

University of Strathclyde
**Department of Naval Architecture, Ocean &
Marine Engineering**



**Natural Language Processing for automatic
classification of human and causal factors from
maritime safety reports**

**By
Hollie Black**

A thesis presented in fulfilment of the requirements for the degree of

Doctor of Philosophy

Glasgow, UK

2026

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my family and friends, especially my mother, Annmarie Black, and my father, William Black, for their continuous support, encouragement, and understanding throughout this PhD. Their unwavering belief in me, patience, and motivation have been my greatest source of strength during both the challenges and successes of this journey.

I would also like to sincerely thank my supervisors, Dr Rafet Kurt and Professor Osman Turan, for their invaluable guidance, expertise, and consistent support throughout this research. Their insightful feedback, academic rigour, and encouragement have been fundamental to the development and completion of this thesis.

My sincere thanks are extended to Louis de Wolff for his support throughout this work and for his assistance in providing access to the near-miss ferry incidents dataset that formed a key case study within this research. I am also grateful to G+ for providing the offshore wind dataset used as part of the analysis. The availability of these real-world datasets was essential to the practical relevance and depth of this research.

I would like to thank my peers at the university for their continued support, valuable discussions, and their assistance with validation during the analysis. Their feedback and collaboration were greatly appreciated and contribute meaningfully to the quality of this work.

Finally, I am grateful to all those who, in different ways, contributed to the completion of the PhD, whether through academic support, professional collaboration, or personal encouragement.

List of Figures

Figure 1. Maritime Accident Investigation flow of investigation processes and oversight.	19
Figure 2. Comparison of accident triangle models (Heinrich's Triangle, Bird, and ConocoPhillips).....	20
Figure 3. Relationship between HOF, human factors, and ergonomics.....	23
Figure 4. Image of Reason's Swiss Cheese Model.....	26
Figure 5. Thesis chapter flow diagram.....	31
Figure 6. Research design and relationship between the research question, objectives and methodological approaches.	35
Figure 7. Literature search, screening and categorisation process for the two review strands.	38
Figure 8. HFACS Framework (Shappell and Wiegmann, 2000).	40
Figure 9. Distribution of HFACS variants/adapted frameworks in maritime papers.	41
Figure 10. SHIELD Taxonomy framework Iceberg visualisation (Stroeve et al., 2023).....	47
Figure 11. Human element relationship graph (Nurduhan and Kuleyin, 2024)	65
Figure 12. Topic co-occurrence network derived from STM applied to Hong Kong bus incident reports (Xu et al., 2025).....	69
Figure 13. BERTopic-derived hazard themes identified from wildfire incident reports using the HEAT framework (Andrade and Walsh, 2023).....	70
Figure 14. Performance comparison of different prompting strategies for LLM-based HFACS classification across multiple subtasks (Liu et al., 2025b).	83
Figure 15. User interface of the incident investigation support system developed by Mathew et al. (2021).....	90
Figure 16. User interface of the intelligent HAZOP analysis system developed by Wang and Gu (2022) for automated hazard identification and consequence analysis.	91
Figure 17. Overview of reviewed methodological approaches.	94
Figure 18. Visual diagram of framework.....	100
Figure 19. BERTopic Flow Diagram.....	105
Figure 20. High information approach methodology workflow diagram.	113
Figure 21. Final SHIELD-NLP Taxonomy.....	130
Figure 22. Bar chart of incident types from SHIELD analysed data.	132
Figure 23. Distribution of SHIELD-NLP justification text lengths measures in tokens.	138
Figure 24. Bar chart of per-category performance results.	140
Figure 25. Bar chart showing the number of codes in each performance band with F1 score.	142
Figure 26. SHIELD-NLP PDF upload interface.....	148
Figure 27. Automatically extracted causal sentences from an investigation report.	150
Figure 28. User validated and stored classification records.	151
Figure 29. ARM network of contributory factors (full dataset).	154
Figure 30. ARM network of contributory factors for collision incidents.	156
Figure 31. ARM network of contributory factors for grounding incidents.	162
Figure 32. ARM network of contributory factors in fire incidents.....	167
Figure 33. ARM network of contributory factors in occupational incidents.	171
Figure 34. Flow diagram of the methodology used in the ferry near-miss case study.....	184
Figure 35. Seasonal patterns of incidents by incident topic and month.	189
Figure 36. Heatmap of yearly incidents by topic from 2019 to 2023.	190
Figure 37. Network of incident type and prevention association rule mining results.	199
Figure 38. Flow diagram of the methodology used in the offshore wind case study.....	214
Figure 39. Monthly distribution of offshore wind incidents by year.	217

Figure 40. Geographic distribution of offshore wind incidents.	218
Figure 41. Heatmap of body part injured, with number of injuries to each body part given.	228
Figure 42. ARM network of offshore wind incidents.	234
Figure 43. Flow chart of preliminary study methodology.....	282
Figure 44. Word cloud of high potential and non-high potential incidents.	283
Figure 45. Word cloud of dropped object and non-dropped object incidents.....	283
Figure 46. Word cloud of ERME and non-ERME incidents.	284
Figure 47. Prototype tool interface for offshore wind classification.	286
Figure 48. Methodology flow diagram of SHIELD preliminary study.....	288
Figure 49. SHIELD-NLP Initial Acts Layer from preliminary study.	290
Figure 50. Word clouds, showing the most frequent words in each of the following layers: acts, preconditions, operational leadership and organisational.....	291
Figure 51. SHIELD Prototype interface preliminary study.....	294

List of Tables

Table 1. Overall comparison of accident causation analysis methods, ...	55
Table 2. Comparison of topic modelling approaches in accident investigation research.	71
Table 3. Frequency distribution of SHIELD-NLP codes within the Acts layer.	134
Table 4. Frequency distribution of SHIELD-NLP codes within the Preconditions layer.	135
Table 5. Frequency distribution of SHIELD-NLP codes within the Operational Leadership layer.	136
Table 6. Frequency distribution of SHIELD-NLP codes within the Organisational layer.	137
Table 7. Macro and Micro summary statistics of Stage 1 model performance.	138
Table 8. Stage 1 classification performance for high-level SHIELD-NLP causal categories.	141
Table 9. Macro and Micro summary statistics of Stage 1 model performance.	141
Table 10. Stage 2 classification performance for fine-grained SHIELD-NLP causal codes.	144
Table 11. Model performance comparison across different feature representation methods.	146
Table 12. Full data association rule mining results.	154
Table 13. Collision incidents association rule mining results.	156
Table 14. Grounding incidents association rule mining results.	162
Table 15. Fire incidents association rule mining results.	167
Table 16. Occupational incidents association rule mining results.	171
Table 17. Clustered topic groups of incident summary topic modelling.	185
Table 18. Clustered topic groups of near miss prevention	190
Table 19. ICC scores per evaluation dimension.	200
Table 20. Mean expert scores across topic validation dimensions.	201
Table 21. ICC scores per evaluation dimension.	202
Table 22. Mean expert scores across topic validation dimensions.	203
Table 23. Summary of the structured offshore wind incident dataset (2017-2021)	214
Table 24. Distribution of recorded incident consequences	216
Table 25. Observed and Expected number of incidents by site type.	219
Table 26. Incident type clusters with associated topics and their percentage of total classified reports.	223
Table 27. Injury type clusters with associated topics and their percentage of total classified reports.	228
Table 28. Causal factors clusters with associated topics and their percentage of total classified reports.....	230
Table 29. ARM results with topic, linked attribute and metrics.	233
Table 30. ICC scores per evaluation dimension.	236
Table 31. Mean expert scores across topic validation dimensions.....	237
Table 32. Model performance metrics for predicting high-potential incidents.	285
Table 33. Model performance metrics for predicting dropped object incidents.	285
Table 34. Model performance metrics for predicting ERME incidents.....	286
Table 35. SVM Model performance metrics for all Acts.	292
Table 36. Naive Bayes Model performance metrics for all ACTS.	292
Table 37. K-nearest neighbour model performance metrics for all Acts.....	292
Table 38. Logistic regression model performance metrics for all Acts.	293
Table 39. Random forest model performance metrics for all Acts.	293
Table 40. Decision tree model performance metrics for all Acts.	293
Table 41. Mapping of Original SHIELD Codes to SHIELD-NLP Codes	295
Table 42. Codes removed from analysis and reasons	296

Abstract

Maritime and offshore operations remain vulnerable to accidents and near misses in which human and organisational factors play a central role. Although narrative incident reports contain rich qualitative evidence about these contributory mechanisms, current analytical practices rely heavily on manual review, making large-scale learning slow, inconsistent, and resource intensive. Recent advances in natural language processing (NLP) offer new opportunities to exploit this data, yet applications in the maritime and offshore wind sectors remain fragmented, rarely taxonomy-aligned, and seldom extend to near-miss and Safety-II perspectives.

This thesis develops and evaluates a unified NLP framework for extracting human and causal factors from maritime and offshore incident narrative summaries under both low-information (short, unstructured summaries) and high-information (structured, taxonomy-labelled) conditions. The high-information branch introduces SHIELD-NLP, a machine-learning-adapted version of the SHIELD taxonomy, combined with SeaSafeBERT, a domain-adapted transformer fine-tuned on maritime accident reports, a hybrid semantic-lexical feature representation, and a hierarchical, probability-calibrated logic regression classifier. Association rule mining is then used to make interaction patterns between human, organisational, and environmental factors across accident types. The low-information branch applies BERTopic to short narratives, supported by GPT-based topic labelling and expert validation, to derive themes, temporal patterns, and links between incident types and preventive actions.

Three case studies demonstrate the framework's applicability. In the taxonomy-coded maritime accident reports, the pipeline achieves strong performance for several high-value SHIELD factors with F1-scores reaching 1 for visual/auditory detection and personnel/training factors, 0.92 for inadequate risk assessment, and 0.85 for inadequate leadership/supervision and failures in information transmission. This enables reliable automatic or semi-automatic coding of failures in detection, communication, planning, leadership, and risk assessment. Association rules reveal shared

vulnerabilities across accident types, including links between adverse operational conditions and decision-making failures; for example, incompatible workspaces were associated with failures in information transmission (confidence =0.61, lift=1.76), while environmental and technological difficulties were consistently associated with incorrect decisions or plans (confidence=0.60-0.62). In the ferry near-miss database (4,360 reports), BERTopic revealed 76 incident themes and 87 prevention themes, extending the operators' taxonomy with 34% of incident-topic clusters represented patterns not captured by existing taxonomy, uncovering seasonal patterns in passenger/traffic and technical risks, and systematically links identified topics to documented prevention actions. This provided empirical evidence of adaptive capacities and resilient practices through which ferry operations are maintained under varying operational conditions, including crew vigilance, teamwork and ship-shore coordination. Analysis of 5,403 offshore incidents from 2017-2021 identified 85 distinct topics using a combination of topic modelling, and association rule mining on pre-structured multi-label outcomes (e.g. high-potential events, dropped objects, ERME). The findings highlight recurring risks in lifting operations, vessel transfers, working at heights, and PPE non-compliance, as well as characteristics of injury patterns and environmental contributors. ARM further identified workplace safety violations as an important risk-related themes, with 94% of occurrences associated with hazard outcomes and 73% associated with high-potential incidents. The theme was also associated with working at heights 72% of the time, highlighting procedural compliance during elevated work as a priority for targeted intervention.

A prototype decision-support interface illustrates how these components can be operationalised into a human-in-the-loop safety intelligence tool, providing calibrated factor predictions, causal pattern visualisations, and near-miss learning support. The principal contribution is an integrated and explainable NLP framework that accommodates both taxonomy-rich and information-sparse safety data, thereby supporting conventional accident analysis alongside Safety-II-oriented learning from near misses and successful preventive actions. By transforming large volumes of previously underutilised narrative safety data into structured and interpretable safety intelligence, the framework

reduces reliance on exclusively manual analysis and enables safety practitioners to identify recurring human and organisational vulnerabilities, detect emerging or previously uncategorised operational risks, prioritise targeted interventions, and systematically learn from effective preventive actions at scale. This provides a practical pathway for embedding NLP within organisational safety-management processes while retaining expert oversight of automated outputs.

Publications Arising from this Thesis

The following peer-reviewed publications have arisen from the research undertaken as part of this thesis. These publications present components of the methodologies and applications developed throughout the thesis and form part of the overall contribution of this research.

1. Black, H., de Wolff, L., Kurt, Y. B., Farag, Y., Waskito, D., Kurt, R. E., & Turan, O. (2026). *Uncovering patterns in ferry near misses: A topic modelling approach with GPT-assisted labelling*. Ocean Engineering, 363(Part 4), Article 126862.
2. Black, H., Kurt, R. E., & Turan, O. (2026). *A domain-adapted NLP pipeline for human factors classification of maritime accidents*. Ocean Engineering, 357(Part 3), Article 125625.

Contents

1. Introduction	13
1.1 Chapter Overview	13
1.2 Maritime and Offshore Safety Context	13
1.2.1 Maritime Safety Context	13
1.2.2 Offshore Safety Context	16
1.3 Accident Investigation and Challenges to Safety Learning	18
1.4 Near-Misses and Proactive Safety Learning	21
1.5 Human and Organisational Factors in Maritime Safety	23
1.6 The Safety Data Analysis Challenge and Research Motivation	27
1.7 Research Layout	28
2. Research Question, Aims and Objectives	33
2.1 Chapter Overview	33
2.2 Research Question	33
2.3 Aims & Objectives	34
2.4 Research Approach	35
3. Critical Literature Review of Accident Learning and AI-Based Safety Analysis	37
3.1 Chapter Overview	37
3.2 Literature Review Scope and Method	37
3.3 Accident Analysis Methods	40
3.3.1 Human Factors Taxonomies	40
3.3.2 System and Accident Causation Approaches	48
3.3.3 Comparison of Accident Analysis Approaches	54
3.4 AI-Based Analysis of Safety Narratives	59
3.4.1 Text Representation and Feature Extraction	59
3.4.2 Unsupervised Analysis of Safety Narratives	64
3.4.3 Supervised Classification of Safety Narratives	73
3.4.4 Analysis and Application of Structured Safety Information	86
3.5 Summary and Identified Research Gaps	92
3.5.1 Accident and human-factors analysis	93
3.5.2 AI-based Safety Analysis	94
3.5.3 Identified Research Gaps and Novelty	97
4 Methodological Framework	100
4.1 Chapter Overview	100

4.2 Framework Overview	100
4.3 Low Information Approach (Unstructured Short Incident Summaries)	102
4.3.1 Topic Modelling (BERTopic)	103
4.3.2 Topic Labelling	106
4.3.3 Validation	107
4.3.4 Trend and Causal Analysis	109
4.4 High-Information Approach (Structured Data from SHIELD)	111
4.4.1 Taxonomy Adaptation	112
4.4.2 Model Pipeline	113
4.5 Chapter Summary	122
5. Automated Human Factors Classification in Maritime Accident Reports using the SHIELD-NLP Framework	123
5.1 Chapter Overview	123
5.2 Data Acquisition and Development of the Database	123
5.3 The Safety Human Incident and Error Learning Database (SHIELD) Taxonomy	125
5.4 Refining and structuring of SHIELD-NLP	129
5.5 Descriptive Statistics of SHIELD-NLP Dataset	132
5.5.1 General Statistics	132
5.5.2 Human Factors Statistics	134
5.5.3 Text-based Analysis	138
5.6 Performance Evaluation of the Two-Stage SHIELD-NLP Classification Framework	139
5.6.1 Stage 1: High-Level Causal Category Performance	139
5.6.2 Stage 2: Fine-grained SHIELD-NLP Code Performance	142
5.6.3 Model Performance Comparisons	147
5.7 Deployment of the SHIELD-NLP Decision Support Tool	147
5.7.1 Document Ingestion and Text Extraction	148
5.7.2 Sentence Segmentation and Casual Sentence Selection	149
5.7.3 Automated SHIELD-NLP Code Prediction	150
5.7.4 Interactive Human-in-the-Loop Evaluation Interface	150
5.7.5 Persistence and Database Storage of Expert-Validated Outputs	151
5.7.6 Operational Significance of the Deployment	152
5.8 Association Rule Mining for Human Factors Interaction Analysis	152
5.8.1 Overall Database	152
5.8.2 Accident Type Analysis	156
5.9 Discussion of Model Performance, ARM Results and Practical Implications	177
5.9.1 Classification Performance	177

5.9.2 Accident Patterns Identified Through ARM	178
5.9.3 Methodological Implications	180
5.10 Chapter Summary	181
6. Uncovering patterns in Ferry Near misses using topic modelling and GPT-assisted labelling	183
6.1 Chapter Overview	183
6.2 Ferry Near Miss Data and Modelling Approach	183
6.3 Results of the Application of BERTopic to Ferry Near-misses	186
6.3.1 Near Miss Incident Summary Classification and Topic Naming	186
6.3.2 Time-series Analysis	189
6.3.3 Classification Preventions	191
6.3.4 Linking Incident Types to Prevention Actions	194
6.3.5 Expert Validation	201
6.4 Discussion of Near-Miss Analysis Findings	204
6.5 Practical Implications of the Findings and Method	210
6.6 Chapter Summary	211
7. Offshore Wind Farm Occupational Incident Analysis using topic modelling and GPT-assisted labelling	213
7.1 Chapter Overview	213
7.2 Offshore Wind Data and Modelling Approach	213
7.3 Statistical Overview of the Offshore Wind Dataset	215
7.3.1 Temporal Analysis of Incidents	217
7.3.2 Geographical Distribution	218
7.3.3 High-Potential Incidents Across Site Types	219
7.3.4 Dropped object vs High-Potential	220
7.3.5 Limitations of this Data	221
7.4 Offshore Wind Incident Topic Modelling Results	222
7.4.1 Incident Types	223
7.4.2 Injury Types and Body Parts Affected	227
7.4.3 Human, Environmental and Causal factors	230
7.5 Association Rule Mining Results	233
7.6 Expert Validation	236
7.7 Discussion of Topic Modelling Findings and ARM	239
7.8 Chapter Summary	243
8. Discussion	245
8.1 Chapter Overview	245
8.2 Research Overview	245

8.3 Research Novelty and Contributions	247
8.4 Achievement of Research Objectives	251
8.5 Limitations	253
8.6 Recommendations for Future Research	257
9. Conclusions	262
9.1 Chapter Overview	262
9.2 Conclusions	262
References	265
Appendix A: Preliminary Study 1: Development of NLP Models for Offshore Wind Incident Classification	282
Appendix A.1 Purpose, dataset and pipeline.....	282
Appendix A.2 Experimental results.....	286
Appendix A.3 Prototype Offshore Wind Classification Tool	287
Appendix A.3 Lessons Learned	288
Appendix B: Preliminary Study 2: Development of SHIELD-NLP	289
Appendix B.1 Purpose and preliminary classification pipeline.....	289
Appendix B.2 Results of SHIELD-NLP preliminary study	290
Appendix B.3 Prototype Interface	295
Appendix B.4 Lessons Learned	295
Appendix C: SHIELD-NLP Development Tables	296

1. Introduction

1.1 Chapter Overview

This chapter establishes the context and motivation for the research by examining maritime and offshore safety, the continuing challenges associated with accident investigation and safety learning, and the role of human and organisational factors in safety outcomes. It considers the value and limitations of learning from accidents, incidents and near misses, including the opportunities presented by more proactive and Safety-II-oriented approaches. The chapter then establishes the safety-data analysis challenge arising from the growing volume of heterogeneous and predominantly unstructured safety information and introduces the potential role of artificial intelligence in supporting more scalable and systematic analysis. Finally, the chapter outlines the structure of the thesis.

1.2 Maritime and Offshore Safety Context

Maritime and offshore industries operate in complex, safety-critical environments where failures can result in significant human, environmental and economic consequences. Despite advances in technology, regulation and safety management, accidents and occupational incidents continue to occur across these sectors. Understanding the development and current state of maritime and offshore safety provides the context for examining how accident, incident and near-miss data can be more effectively used to support safety learning.

1.2.1 Maritime Safety Context

Maritime transport is fundamental to the global economy, carrying more than 80% of world trade (UNCTAD, 2025). The sector has made substantial safety improvements through developments in vessel design, regulation, and safety management, many of which followed major maritime accidents. These have historically catalysed changes in safety regulation and practice. Importantly, the lessons from these events show an evolution in maritime safety thinking, from an early emphasis on technical and design deficiencies to greater recognition of human, organisational, and systemic factors.

The Titanic disaster in 1912, which resulted in the loss of more than 1,500 passengers and crew, exposed deficiencies in iceberg detection, decision-making and evaluation arrangements, alongside organisational overconfidence (Ashraf, 2014). The regulatory response focused largely on improving technical and operational safety requirements and contributed to establishing the International Convention for the Safety of Life at Sea (SOLAS) in 1914. SOLAS was created to specify minimum standards for vessel safety with respect to construction and operation. The convention has continued to be updated, with the last amendment in 1974 (IMO, 1974).

Environmental disasters similarly prompted regulatory change. The grounding of MV Torrey Canyon in 1967 resulted in a major crude-oil spill and contributed to international efforts that culminated in the International Convention for the Prevention of Pollution from Ships (MARPOL) established in 1973 (Wells, 2017). The Exxon Valdez grounding in 1989 similarly exposed limitations in oil-spill preparedness and response in the United States, contributing to the introduction of the Oil Pollution Act 1990 (Haugen and Kristiansen, 2022, Guard), 1990).

Later accidents increasingly highlighted that technical regulation alone could not prevent maritime casualties. The 1987 Herald of Free Enterprise disaster, in which 193 died after the vessel capsized with its bow doors open, initially appeared to result from frontline crew error; however, the subsequent investigation exposed wider deficiencies in supervision, communication and organisational culture (Haugen and Kristiansen, 2022). The accident became an important example of the need to consider management and organisational influences on safety and contributed to the development of the International Safety Management (ISM) Code, adopted in 1993 to establish international standards for the safety management and operation of ships. The 1994 Estonia disaster, resulting in 852 fatalities, further demonstrated the interaction between technical vulnerability and wider approaches to recognise and manage recurring risk, particularly given previous bow-visor incidents (Soomer et al., 2001, Hänninen, 2003).

More recently, the grounding of the Costa Concordia in 2012, resulting in 32 fatalities, demonstrated that catastrophic accidents can continue to occur despite advanced vessel technologies and established regulatory frameworks. The accident involved multiple human and organisational factors, including decision-making, deviation from planned routes, ineffective bridge resource management, and evacuation procedures (Alexander, 2012). Collectively, these accidents illustrate the progressive development of maritime safety from predominantly reactive responses to individual technical failures toward a broader understanding of accidents as outcomes of interactions among technical, human, and organisational factors.

Despite developments in maritime safety regulation and safety management, accidents and incidents remain frequent. European Maritime Safety Agency (EMSA) data indicate an average of approximately 2,675 maritime casualties and incidents annually, although the number of very serious casualties has generally declined since 2014, reaching 51 in 2024, with 37 fatalities recorded in the same year (EMSA, 2025), EMSA (2024). This reduction in severe outcomes contrasts with the continuing burden of non-fatal events, with reported injuries increasing from 624 in 2022 to 809 in 2023 and then to 803 in 2024 (EMSA, 2025).

National-level data reflects similar patterns, with the Marine Accident Investigation Branch (MAIB, 2024) recording 751 marine incidents, 153 less-serious accidents, 39 serious accidents and 7 very serious accidents in 2023. While changes in reporting practices may contribute to observed trends, these figures demonstrate that maritime safety remains an ongoing operational challenge despite reductions in the most severe outcomes.

Importantly, human factors continue to represent a major component of maritime accident causation. Between 2015 and 2024, 64.5% of incidents were attributed to human error, compared with 21% attributed to system or equipment failures, while 80.1% of identified contributing factors for incidents in this period were related to the human element EMSA (2025).

In terms of safety recommendations and actions, EMSA reports that most effort focused on ship-related procedures, followed by human factors and ship structure and equipment. Within the human factors category, most changes have focused on improving training, skills, and experience. This reflects an industry-wide acknowledgement of the need to enhance human performance and reduce errors in critical maritime operations. Correspondingly, safety recommendations on human factors have particularly emphasised improvements in training, skills, and experience. These findings demonstrate that technological and regulatory advances alone have not eliminated the underlying conditions contributing to maritime incidents and reinforce the importance of understanding how human, organisational and operational factors interact within maritime systems.

1.2.2 Offshore Safety Context

Offshore operations present significant safety challenges because of their remote locations, harsh environmental conditions, complex technologies, and dependence on safety-critical activities such as vessel transfers, lifting operations, maintenance, and working at height. The offshore oil and gas sector offers a clear example of how major accidents have shaped safety regulation and organisational learning. The Piper Alpha disaster, which resulted in 167 fatalities, demonstrated how catastrophic outcomes can emerge from interacting technical, operational and organisational deficiencies, including weaknesses in plant design, permit-to-work arrangements, maintenance, training, communication, and emergency management. The accident subsequently contributed to substantial changes in offshore regulation, asset design and approaches to safety management (Binder et al., 2010). More recently, the Deepwater Horizon disaster in 2010 demonstrated the continuing importance of human and organisational factors, with investigators identifying deficiencies in safety-critical communication, risk management, contractor oversight, and safety culture alongside technical failures.

The offshore oil and gas industry has since developed into a comparatively mature safety-critical sector, supported by decades of operational experience, regulatory development and incident reporting. In

2024, International Association of Oil and Gas Producers (IOGP) members recorded 32 fatalities across 21 incidents; however, despite a 26% increase in reported working hours, the fatal accident rate decreased to 0.77 fatalities per 100 million hours worked, while the Total Recordable Injury Rate decreased to 0.81 injuries per million hours worked ((IOGP), 2025). Longer-term data similarly demonstrate a substantial reduction in fatal accident rates since the late 1980s, reflecting developments in regulation, engineering controls, risk management, human factors and safety culture. Nevertheless, offshore workers continue to experience higher injury rates than their onshore counterparts, demonstrating the persistent hazards associated with offshore operations.

Offshore wind presents many comparable hazards but operates within a considerably newer and rapidly expanding industry. Workers are exposed to risks associated with marine logistics, vessel transfers, heavy lifting, working at height, adverse weather and remote operations throughout construction, operation, maintenance and decommissioning. Unlike offshore oil and gas, however, the sector has not accumulated the same depth of operational experience or regulatory development, while safety reporting remains largely industry-led and can vary between operations and regions. G+, the Global Offshore Wind Health and Safety Organisation, provides an important mechanism for industry safety reporting and knowledge sharing, although the absence of a unified reporting framework can limit consistency and completeness across the sector. These differences highlight the need for systematic approaches to learning from the growing volume of offshore wind safety data.

The need is becoming increasingly important as offshore wind activity expands. G+ data shows that reported working hours have increased from 20.4 million in 2019 to 65.6 million in 2024, while the injury rate increased from 19.5 to 27.1 injuries per million hours worked between 2023 and 2024, with fatalities also recorded during this period (G+, 2023a). In 2023, a worker was fatally injured during the onshore assembly of a wind turbine when they became trapped between the basket of a mobile elevated work platform (MEWP) and the turbine blade bearing on which they were working. In 2024, another fatal incident occurred during the disassembly and scrapping of a rotating bucket from a

monopile-up ending tool on a transport barge moored alongside a quayside. During this operation, part of the bucket fell in an uncontrolled manner, causing the supporting scaffold to collapse. This resulted in one worker suffering a fatal fall from height, while a second worker was hospitalised with injuries.

The injury rate has also been reported as approximately four times that of offshore oil and gas (David et al., 2024), emphasising the importance of continued investigation of offshore wind safety as the sector develops. The contrast between a mature offshore oil and gas industry and the rapidly developing offshore wind sector therefore provides an important context for this research, particularly for investigating whether established and emerging safety data can be analysed systematically to identify recurring operational and human-factor vulnerabilities.

1.3 Accident Investigation and Challenges to Safety Learning

Accident investigation is a fundamental mechanism through which maritime and offshore industries learn from adverse events and develop measures to prevent their recurrence. Investigations aim to establish the circumstances and contributing factors associated with an event and translate these findings into recommendations for improving safety (HE, 2020). Internationally, the IMO Casualty Investigation Code (Resolution MSC.255(84)) establishes mandatory standards for investigating marine casualties and incidents and requires flag states to investigate very serious marine casualties involving outcomes such as loss of life, loss of a vessel or severe environmental damage. Investigations are subsequently undertaken by national authorities, such as the Marine Accident Investigation Branch (MAIB) in the UK, with findings contributing to wider international safety learning through organisations including the IMO and European Maritime Safety Agency (EMSA).

However, the maritime safety-reporting landscape remains fragmented. Differences in national regulations, investigative resources and reporting practices create variation in how incidents are investigated, classified and published, while less severe incidents and near misses are not subject to the same level of mandatory investigation. Safety information is also distributed across national

authorities, operators, insurers, P&I clubs, industry associations and confidential reporting programmes such as CHIRP. Consequently, maritime safety data are often dispersed across different systems, inconsistently structured and predominantly recorded as unstructured narrative text. Human and organisational factors may also be represented inconsistently within formal investigations despite their recognised importance to accident causation

Maritime safety data is also generated and distributed through a wider reporting ecosystem involving multiple stakeholders beyond formal investigation bodies, shown in Figure 1. National investigation agencies such as the Maritime Accident Investigation Branch (MAIB) collect and publish investigation findings, which may then be reported to organisations such as the IMO through GISIS and, within Europe, to EMSA for regional analysis. In addition, industry organisations, insurers, P&I Clubs, operators, confidential reporting programmes such as CHIRP Maritime, and membership organisations including INTERTANKO and INTERCARGO also generate and collect large volumes of safety-related information. However, these data sources are often fragmented, inconsistently structured, and heavily reliant on unstructured narrative reports, creating significant challenges for systematic safety analysis (Schröder-Hinrichs et al., 2011). Compared with sectors such as aviation and nuclear energy, where more structured approaches to human-factor reporting and analysis have been established, this limits opportunities for consistent cross-incident analysis and proactive safety learning (IAEA, 2013).

Maritime Safety Reporting and Investigation Ecosystem

Multiple actors generate, collect and share safety information through different channels.
Data flows are often fragmented, voluntary and based on unstructured narrative reports.

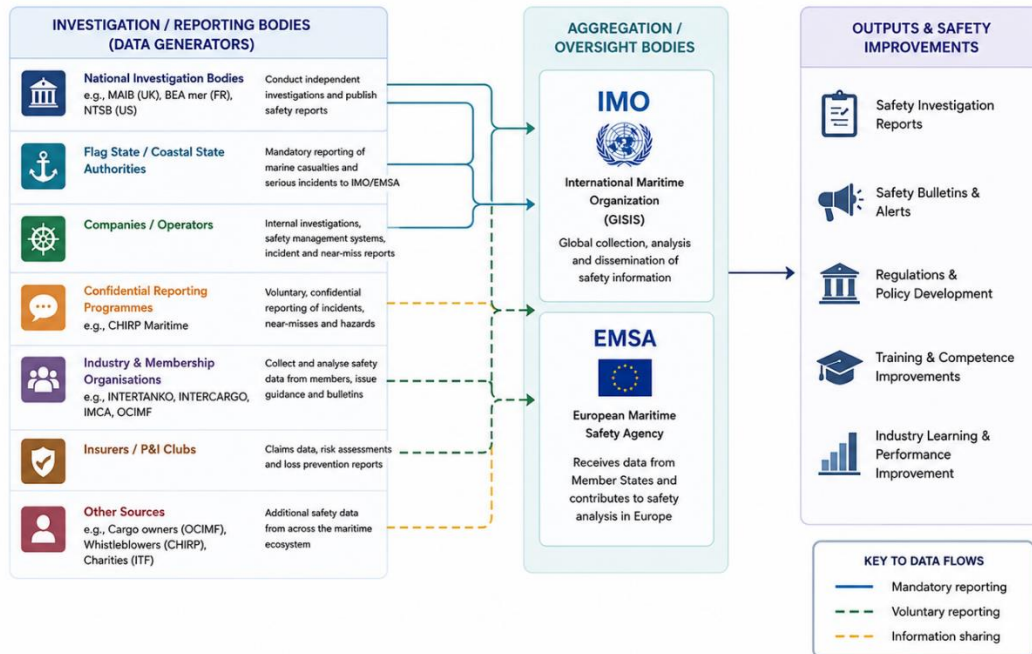


Figure 1. Maritime Accident Investigation flow of investigation processes and oversight.

A further challenge is the disparity between the number of safety events reported and the number receiving detailed investigation. For example, MAIB receives more than 1200 accident notifications annually but typically investigates around 25 to 30 cases each year (fewer than 3% of reports), reflecting the practical need to prioritise events by severity and potential safety value. Under-reporting can further restructure the available evidence, particularly for minor incidents and near misses, where concerns surrounding blame or professional consequences may discourage reporting. Confidential systems such as CHIRP seek to address this by enabling safety concerns and incidents to be reported without attribution. Nevertheless, incomplete reporting, fragmented data sources, and the resource-intensive nature of manual investigation mean potentially valuable safety information may remain underutilised.

These limitations indicate that the challenge is not only whether safety events are reported and investigated, but also whether the resulting information can be analysed consistently and at sufficient

scale to support learning. Detailed accident reports provide rich evidence concerning causal and human factors but are relatively limited in number, while shorter incident and near-miss narratives occur more frequently but are less consistently investigated and contain highly variable information. This created a need to consider how learning can be extended beyond major accidents to the larger body of lower-severity safety events generated during everyday operations.

1.4 Near-Misses and Proactive Safety Learning

While formal accident investigations provide important insights into serious casualties, near misses represent an additional and often underutilised source of safety learning. Near misses are events that had the potential to result in injury, environmental harm or operational damage but did not, either because of intervention, adaptation or favourable circumstances (Hollnagel, 2014). Their potential value has historically been illustrated through accident-pyramid models, shown in Figure 2.

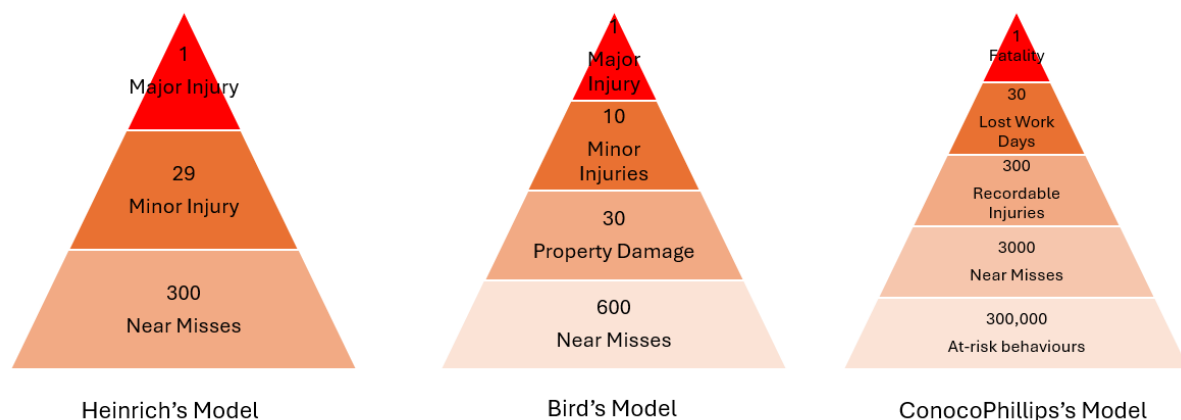


Figure 2. Comparison of accident triangle models (Heinrich's Triangle, Bird, and ConocoPhillips)

Heinrich (1941) proposed that approximately 300 near misses occur for every major injury, while Bird (1969) later proposed a revised ratio of 1 serious injury to 600 near misses. Later industry research has suggested an even greater frequency of lower-severity events relative to fatalities (Marine, 2003). Although these ratios should not be interpreted as universal causal relationships, they illustrate the much greater volume of lower-severity safety events available for learning than major accidents.

Analysis of these events can therefore complement traditional accident investigation by helping to identify recurring hazards, operational deviations and conditions that may contribute to more serious outcomes.

Despite their importance, near-misses remain significantly underreported in the maritime sector (Hassel et al., 2011). Fear of blame or professional consequences, administrative burden, and the perception that events without harmful consequences are of limited importance can discourage reporting. Furthermore, although the IMO Casualty Investigation Code establishes requirements for the investigation of very serious marine casualties, lower-severity events do not receive the same level of mandatory investigation (IMO, 2010).

Near-miss analysis also provides a potential bridge between traditional reactive safety management and more proactive approaches to safety learning. Safety-I predominantly seeks to understand adverse outcomes and their causes, whereas Safety-II places greater emphasis on understanding how work normally succeeds under variable operating conditions and how people and systems adapt to maintain safety performance (Hollnagel, 2014). Near misses remain occurrence-based and therefore should not be considered equivalent to a complete Safety-II approach; however, they can contain information about successful interventions, recoveries and adaptations that prevented an event from escalating. Their analysis can therefore extend learning beyond identifying what went wrong to also consider what prevented a potentially harmful situation from becoming an accident.

This distinction is particularly relevant because conventional maritime-safety-reporting systems rarely capture information about everyday successful operations, whereas many organisations already collect near-miss reports. Near-miss data therefore offer a practical opportunity to support more proactive, resilience-informed safety learning using information that is already available. However, their value depends on organisations being able to identify meaningful patterns across large numbers of heterogeneous reports and to understand the human, organisational and operational factors

associated with both adverse events and successful prevention. This provides an important basis for considering these factors in the following section.

1.5 Human and Organisational Factors in Maritime Safety

Human and organisational factors (HOF) are central to understanding maritime safety because accidents rarely result from a single technical failure or individual error. HOF broadly encompass the interactions between people, tasks, technologies, working environments and organisational systems that influence safe performance. The relationship is illustrated in where Human and Organisational Factors (HOF) form the broadest category, encompassing the design of systems for effective human interaction, personal aspects such as communication and fatigue, and organisational elements such as structure, supervision, and management practices (Kristiansen, 2005). Human factors have consequently been described as the scientific discipline concerned with underlying interactions between humans and other elements of a system (HE, 2020). This systems perspective moves accident analysis beyond attributing outcomes solely to frontline operator error and towards understanding the conditions that shape human behaviour and performance.

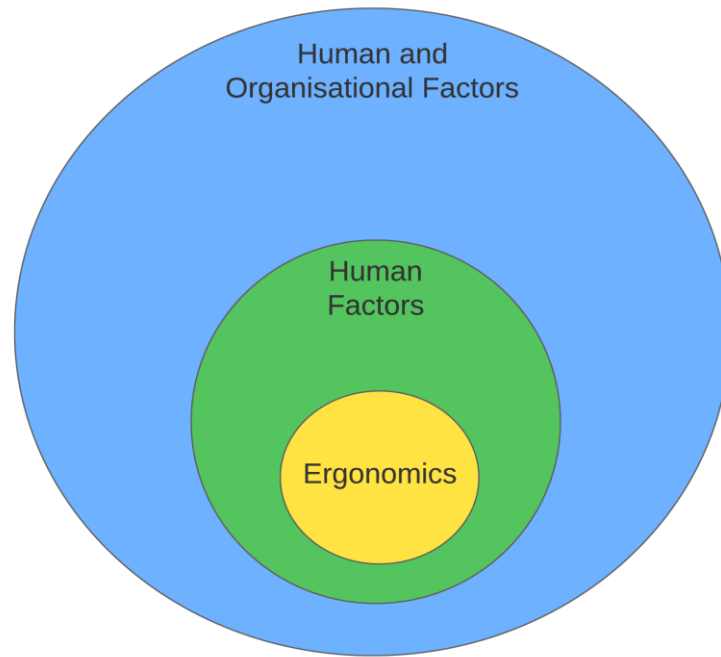


Figure 3. Relationship between HOF, human factors, and ergonomics

Within maritime operations, these influences can occur at multiple interacting levels. **Individual factors** relate to the physical, cognitive, and psychological characteristics of seafarers that can influence performance such as:

- **Fatigue:** Fatigue can be physical and psychological, but both lead to the same effect on human performance with declining alertness, concentration and motivation. Studies have found significant associations between fatigue and increased probability of incidents, such as vessel groundings (Akhtar and Utne, 2014). More recent advances through data-driven approaches have looked to identify fatigue-related risk, linking higher fatigue risk to night-time operations, prolonged voyages, small crew sizes, and adverse environmental conditions (Shiqi and Zaili, 2024).
- **Situational awareness:** refers to an individual's ability to perceive and comprehend elements in their environment accurately, and to anticipate how those elements will evolve over time. Accident analysis by Gommosani et al. (2021) found that 58% of accidents related to

situational awareness deficiencies were linked to failures to perceive key cues (Level 1 SA), and 42% to misinterpreting them (Level 2 SA). These were also linked with poor bridge team coordination, communication and lookout practices.

- Decision making: the process of selecting a course of action from available alternatives, often under conditions of uncertainty and time pressure. Poor decision-making has been linked to collisions, groundings, and other critical incidents, with contributing factors like cognitive overload, inadequate risk assessment, and reliance on habitual responses.

Environmental factors relate to the physical and operational conditions in which seafarers work, many of which can impair performance and increase the likelihood of error:

- Physical work environment: conditions such as temperature, noise, vibration, and illumination can significantly influence comfort, alertness, and task performance. Prolonged exposure to poor environmental conditions has been shown to reduce efficiency and increase human error (Ivergård et al., 1978).
- Vessel motion: ship motion can induce seasickness and motion sickness, affecting individuals differently but often causing nausea, drowsiness, reduced concentration, and vomiting, all of which impair operational effectiveness (Stevens and Parsons, 2002).
- Vision demands: maintaining effective lookout and situational awareness becomes particularly challenging in darkness or reduced-visibility conditions caused by weather.

Organisational and systemic factors influence how work is structured, managed, and supported at both ship and company levels, shaping the broader context in which human performance occurs:

- Communication: effective communication, within the crew, between ship and shore, and with external factors such as pilots or vessel traffic services, is essential for safe operations. Communication breakdowns have been repeatedly identified as major contributors to maritime accidents (Hetherington et al., 2006).

- Bridge resource management: BRM emphasises the coordinated use of all available resources (human, technical and procedural) to ensure safe navigation. Deficiencies in BRM are frequently associated with poor teamwork, unclear roles, and ineffective monitoring.
- Human-machine interface (HMI): Ergonomic HMI design is essential in high-stakes maritime environments. IMO MSC/Circ.982 outlines principles for interface design that promote clear visibility, intuitive control layouts, and alarm prioritisation to reduce operator error (Organisation, 2000).
- Safety culture and safety climate: these refer to the shared beliefs, values, and perceptions within an organisation regarding safety. Strong safety culture supports proactive risk management and open communication, while weak culture can normalise unsafe practices and elevate the risk of incidents.

This systems perspective is reflected in Reason (1990) Swiss Cheese Model (shown in Figure 4, which conceptualises safety defences as multiple layers containing weaknesses through which hazards may propagate when defences align. The model distinguishes between immediate unsafe acts and deeper preconditions, supervisory and organisational influences, illustrating how accidents can emerge through combinations of failures at different levels rather than from a single cause. However, its broad conceptual nature provides limited guidance for systematically identifying and classifying individual contributing factors (HE, 2020). This limitation contributed to the development of more structured human-factor taxonomies, including the Human Factors Analysis and Classification System (HFACS) and subsequent maritime adaptations. The following literature-review chapter reviews and compares these approaches in detail.

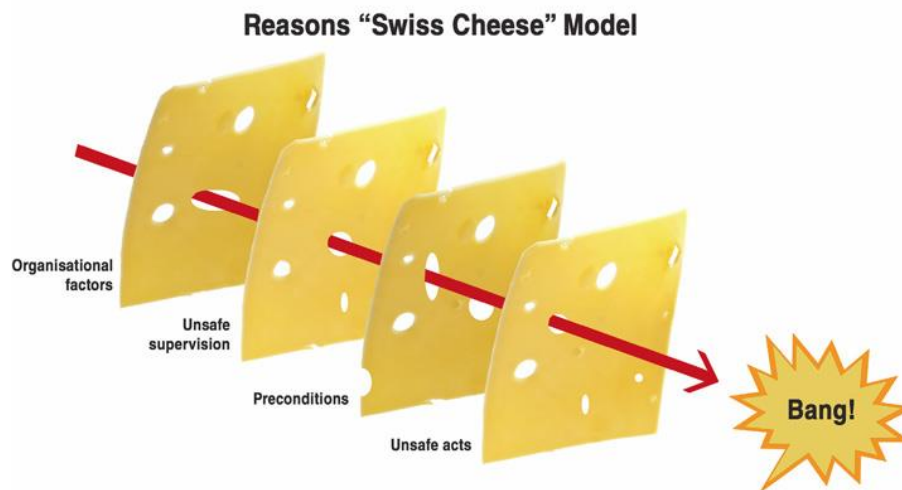


Figure 4. Image of Reason's Swiss Cheese Model

Despite growing recognition of HOF within maritime safety, their identification within accident and near-miss reports remains largely dependent on expert interpretation and manual review. Although structured frameworks and taxonomies can support consistent causal analysis, applying them to large collections of narrative safety reports is resource-intensive and may vary between analysts. This challenge grows as the volume of accident, incident, and near-miss information increases across maritime and offshore industries. Consequently, a key challenge is not simply defining human and organisational factors but determining how to identify them consistently and efficiently across large volumes of heterogeneous safety information.

1.6 The Safety Data Analysis Challenge and Research Motivation

The preceding sections demonstrate that maritime and offshore safety learning relies on information generated at different levels, ranging from detailed accident investigations to shorter incident and near-miss narratives. These sources provide complementary information: full investigation reports contain detailed evidence concerning accident causation and human and organisational factors, while shorter operational reports provide a larger source of information for identifying recurring incidents,

emerging trends and potentially successful preventive actions. The analytical challenge is therefore to extract useful and consistent safety knowledge from these different forms of narrative data.

Traditional analysis relies substantially on manual review by investigators, safety practitioners and human-factors specialists. While expert analysis remains essential, manually reviewing large collections of safety narratives is time-consuming and limits the scale at which recurring patterns can be identified. The problem is particularly significant where structured human-factor taxonomies must be assigned to lengthy accident reports or where thousands of shorter narratives must be compared to identify common themes and relationships. Furthermore, differences in terminology, report length, and level of detail make simple keyword-based analysis insufficient to capture the contextual and semantic meaning in safety narratives.

Recent developments in Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and Machine Learning (ML), provide an opportunity to complement existing expert-led analysis. NLP methods can process unstructured text at scale, while supervised learning can classify reports against predefined safety categories, and unsupervised approaches can identify previously unspecified themes and patterns within larger collections of narratives. These capabilities are particularly relevant to maritime and offshore safety because they can connect two forms of safety learning: structured analysis of known human and causal factors in detailed accident reports, and exploratory identification of themes, trends, associations, and preventive actions in shorter incident and near-miss narratives.

This provides the central motivation for the research. Rather than replacing expert judgement or accident investigation, the study investigates whether AI-based methods can augment safety analysis by increasing scalability, consistency, and the ability to extract actionable knowledge from existing narrative data. The research therefore examines how to develop and evaluate complementary supervised and unsupervised NLP approaches across maritime accidents, ferry near misses, and offshore incidents. In doing so, it addresses the broader question of how increasingly available safety

data can be transformed into structured and interpretable evidence that supports both retrospective accident analysis and more proactive safety learning.

1.7 Research Layout

This thesis is organised into nine chapters, each contributing to a cohesive narrative that explains the reasoning behind the research and culminates in the presentation and analysis of its results. The structure, shown in Figure 5, is designed to guide the reader through the background, methodology, and outcomes, providing a clear understanding of the study's foundation, implementation, and implications.

Chapter 2 presents the central research question that drives the study and outlines the aims and objectives designed to address the identified gaps. It then defines the research approach used to achieve each objective, establishing a clear link between the research question, objectives and the methodological and analytical stages undertaken throughout the thesis.

Chapter 3 critically reviews the literature underpinning the research. It first examines existing accident analysis methodologies and their suitability for maritime accident analysis, establishing the basis for the selection and subsequent adaptation of SHIELD. It then evaluates existing AI, Natural Language Processing (NLP), and Machine Learning (ML) approaches for analysing safety narratives, including supervised classification, language models, topic modelling and related data-driven techniques. Together, these reviews identify the conceptual and methodological gaps addressed by the research and inform the development of the analytical framework.

Chapter 4 introduces the methodological framework developed in this research. Drawing on the findings of Chapter 3, this framework integrates the selected human factors taxonomy with the most appropriate AI-based analytical methods into a unified approach for analysing heterogeneous safety data. It outlines the dual analytical pathway designed to address the differing characteristics and challenges of maritime safety data:

1. Short narrative pathway, which is tailored for brief and unstructured texts as near-misses. This pathway combines BERTopic and LLM-assisted topic labelling to extract coherent themes from sparse textual data, alongside temporal analysis and association rule mining (ARM) to identify how patterns evolve over time and how different themes relate to one another, and
2. Full accident report pathway, which is designed for longer and more detailed investigation reports requiring structured classification. This pathway uses SHIELD-NLP supported by SeaSafeBERT (a maritime-specific BERT model) to capture domain-specific language, combined with SMOTE to address class imbalance and Elastic Net logistic regression and TF-IDF representations to enable robust and interpretable human-factors classification.

This chapter explains how these components are systematically integrated to form a unified and scalable framework, capable of extracting consistent and meaningful insights from heterogeneous safety data sources. The two analytical pathways developed within this framework are then evaluated through a series of case studies presented in Chapter 5-7.

Chapter 5 focuses in detail on the development and implementation of SHIELD-NLP. This chapter addresses the challenge that manually reviewing accident reports and classifying human and organisational factors using taxonomies is both time-consuming and resource-intensive, limiting its application to large accident datasets. By automating this process, the framework aims to provide consistent, scalable, and interpretable classification of accident reports. It outlines the creation of SeaSafeBERT, its training processes, and its integration into the taxonomy-based human-factor classification pipeline. A taxonomy-based approach is adopted to ensure consistent, structured, and interpretable classification of causal and contributory factors, addressing the limitations of purely data-driven methods which often lack transparency and domain alignment in cases where we have large volumes of information per accident. SHIELD-NLP is developed within this research as a machine-implementable adaptation of the SHIELD taxonomy, enabling its application to large-scale unstructured accident report data. The chapter evaluates the performance of the proposed modelling

strategies and explains how SHIELD-NLP advances the automation of causal and contributory factor classification in maritime accident data. This chapter evaluates the fully accident report pathway of the framework before its complementary short narrative pathway is demonstrated using real-world datasets in Chapters 6-7.

Chapter 6 presents the first application of the short narrative pathway on a ferry near-miss dataset, analysing incident narratives and prevention narratives. This case study was selected because ferry operators generate large volumes of near-miss reports, yet resource constraints often limit the detailed analysis of this information. By automating the extraction of themes, trends, and prevention strategies, the framework demonstrates how valuable safety insights can be obtained from routinely collected data to support proactive safety management. This demonstrates the practicality of the framework and its effectiveness on near-miss datasets, uncovering links between incident types and associated preventions. This first case study demonstrates the applicability of the short narrative pathway within the maritime domain, providing a basis for assessing its transferability to offshore wind in Chapter 9.

Chapter 7 extends the framework's application to offshore wind incident data. This case study was selected to demonstrate that the proposed methodology is transferable beyond the maritime sector and can support safety analysis in an emerging safety-critical industry where incident reporting datasets are rapidly growing but remain underutilised. The chapter applied the short narrative pathway to extract the trends in the accident narratives. This showcases the framework's adaptability to a rapidly evolving safety-critical industry, one of increasing global importance. Together the maritime and offshore wind case studies demonstrate the framework's applicability across different safety-critical domains and data types.

Chapter 8 then discusses the theoretical, methodological, and practical implications of the findings across maritime and offshore wind safety. It reflects on how the proposed framework contributes to accident investigation, human factors analysis, safety-II learning, and data-driven safety management. The chapter also addresses limitations and proposed avenues for future work. The findings from both

analytical pathways and case studies are synthesised to evaluate the overall contribution of the proposed framework.

Chapter 9 concludes the thesis by summarising key contributions, highlighting the novelty and significance of the integrated AI-based analytical framework, and outlining clear recommendations for researchers, practitioners, and regulatory stakeholders.

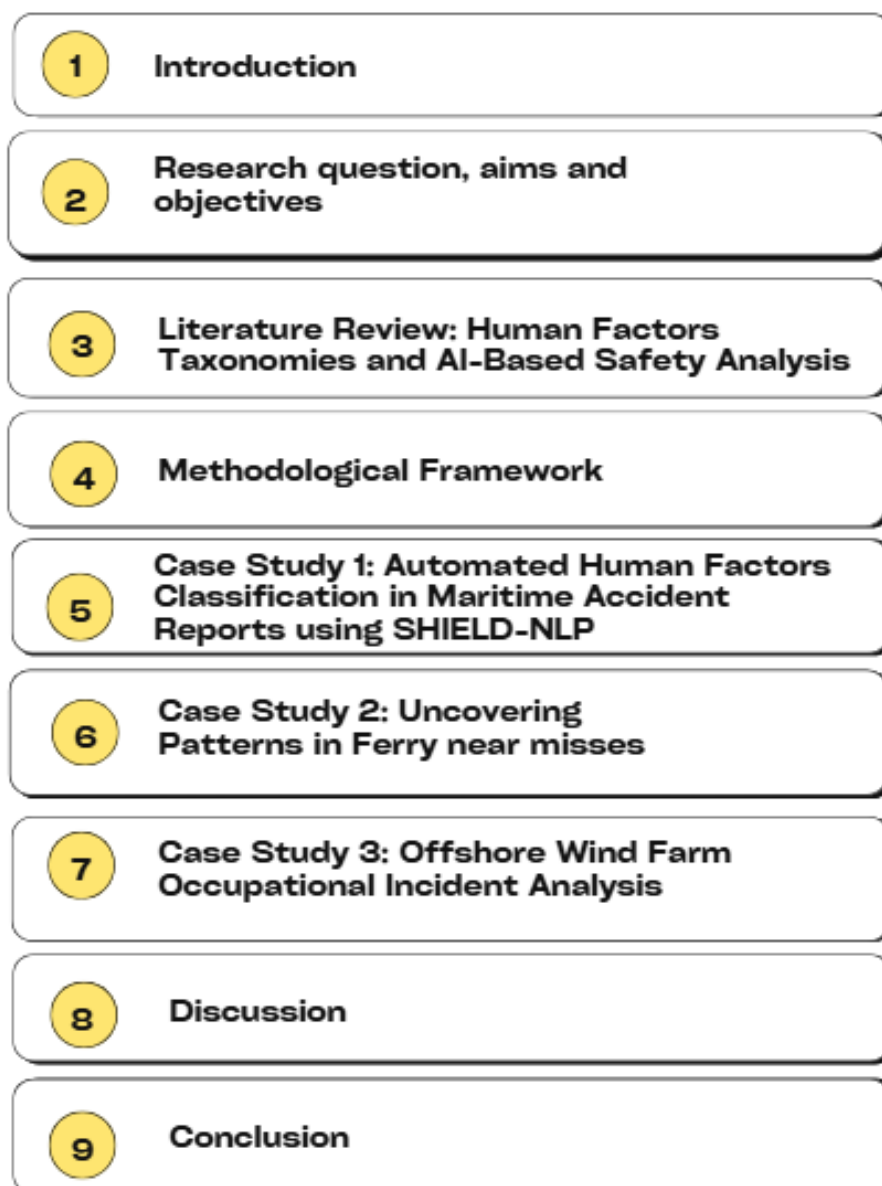


Figure 5. Thesis chapter flow diagram.

2. Research Question, Aims and Objectives

2.1 Chapter Overview

This chapter outlines the central research question guiding the study, followed by a clear articulation of its overarching aim, supporting objectives and research approach. These elements collectively shape the methodological choices and empirical analyses presented in subsequent chapters.

2.2 Research Question

The main research question is:

“How can advances in Artificial Intelligence be used to enhance safety analysis and organisational learning from unstructured safety data in maritime and offshore operations?”

This overarching question reflects the dual focus of the research: (1) automated classification of human and causal factors in full major accident reports, and (2) automated extraction of themes, trends, and associations from short incident narratives and near misses in both the maritime and offshore wind sectors.

To address the central question, the study considers the following sub-questions:

1. What methodological requirements and AI-based analytical approaches are most appropriate for extracting meaningful safety knowledge from unstructured safety reports?
2. To what extent can human and causal factors be automatically and reliably classified from detailed maritime accident investigation reports using a domain-informed taxonomy and supervised NLP methods?
3. To what extent can AI-based unsupervised methods extract meaningful themes, temporal patterns, associations and preventive actions from short incident and near-miss narratives where predefined classifications are limited or absent?

4. To what extent can a unified AI-driven analytical framework provide scalable, interpretable and transferable support for safety learning across different reporting formats and safety-critical operational domains?

2.3 Aims & Objectives

The overarching aim of the research is to enhance safety learning and support more proactive, evidence-based safety management in maritime and offshore operations by advancing the use of AI to transform unstructured safety narratives into actionable safety knowledge.

This aim is pursued through the development and evaluation of an integrated AI-based analytical framework capable of supporting analysis across heterogeneous forms of safety reporting, from detailed accident investigations to short incident and near-miss narratives.

Objectives:

1. Identify the theoretical, methodological and practical gaps that limit the systematic use of unstructured safety narratives for human-factors analysis and organisational safety learning.
2. Develop an integrated AI-based analytical framework capable of extracting structured and interpretable safety knowledge from heterogeneous forms of unstructured safety reporting
3. Develop and evaluate a domain-informed approach for automating human and causal-factor classification from detailed maritime accident investigation reports.
4. Demonstrate and evaluate the framework's capability to identify incident themes, temporal patterns, factor associations and preventive actions from short incident and near-miss narratives, including information-sparse reporting contexts.
5. Evaluate the transferability and practical contribution of the framework across maritime and offshore safety contexts and establish how resulting insights can support proactive, evidence-based organisational safety learning.

2.4 Research Approach

To address the research question and achieve the research objectives, a staged research approach was adopted, progressing from the identification of research gaps and methodological requirements to framework development, empirical evaluation, and assessment of transferability and practical contribution. The resulting framework incorporates two complementary analytical pathways: a supervised pathway for detailed, taxonomy-labelled accident reports and an unsupervised pathway for shorter incident and near-miss narratives where predefined classifications are limited or unavailable.

As illustrated in **Error! Reference source not found.**, the supervised pathway is developed and evaluated using maritime accident investigation reports in Chapter 5, while the unsupervised pathway is evaluated using ferry near-miss and prevention narratives in Chapter 6. The latter pathway is subsequently applied to offshore-wind incident data in Chapter 7 to assess its transferability to a different safety-critical domain. Findings from these applications are synthesised in Chapter 8 to evaluate the overall contribution of the framework to scalable, interpretable and proactive safety learning.

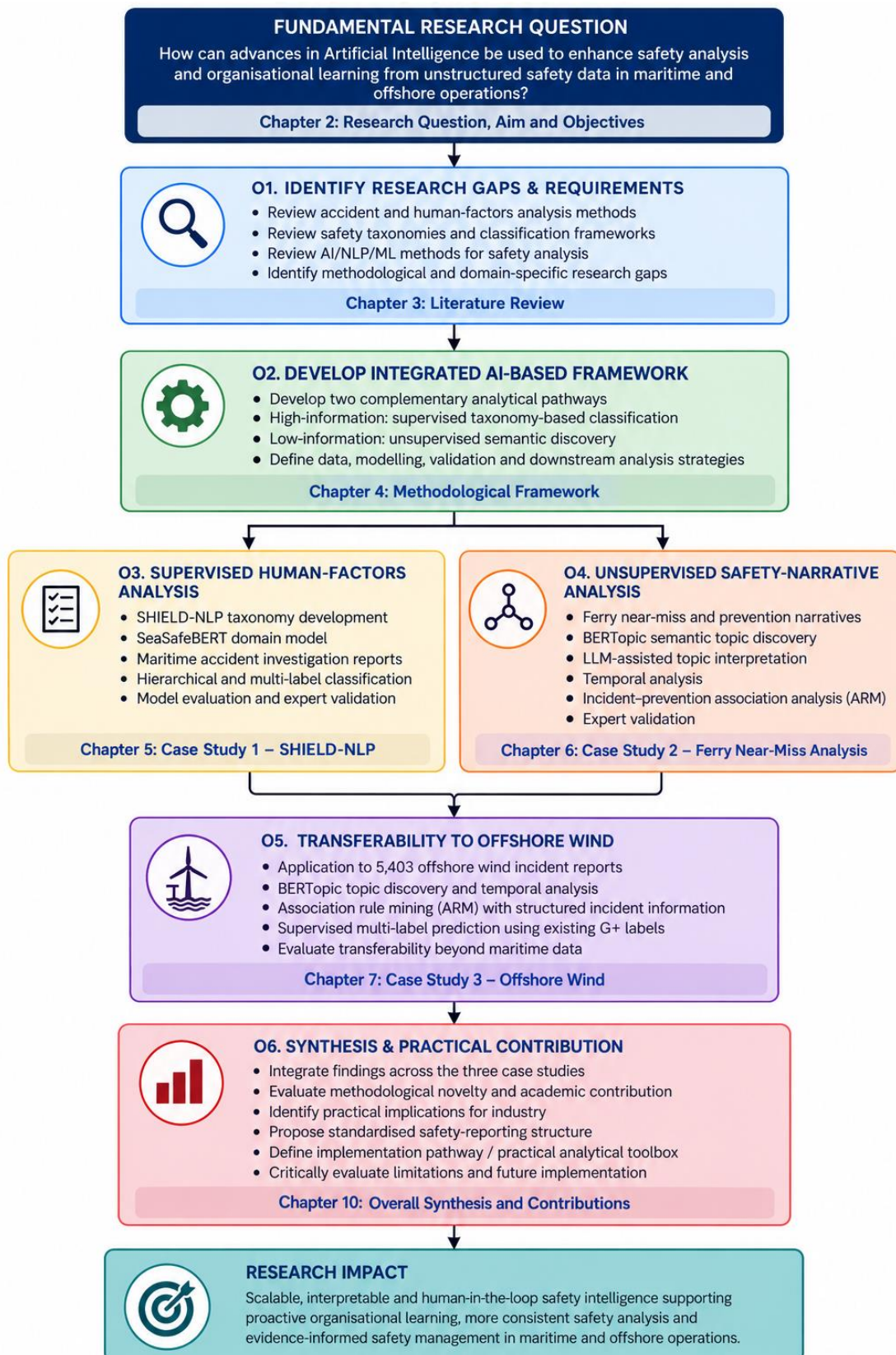


Figure 6. Research design and relationship between the research question, objectives and methodological approaches.

3. Critical Literature Review of Accident Learning and AI-Based Safety Analysis

3.1 Chapter Overview

This chapter critically reviews principal methods used to analyse human and organisational and causal factors within safety narratives, bringing together two complementary areas of literature: structured accident-analysis frameworks and AI-based text-analysis methods. The first considers human factors taxonomies and accident causation models used to identify and organise contributory factors, while the second examines Natural Language Processing (NLP), Machine Learning (ML) and related data-driven techniques used to extract themes, classifications and relationships from unstructured safety reports. The review first evaluates existing accident-analysis approaches to determine their suitability for systematically identifying and structuring human and organisational factors within maritime safety data. This establishes the basis for selecting an appropriate analytical framework for the research. Building on this, AI, Natural Language Processing (NLP) methods are critically reviewed to determine how such analysis can be automated and extended to larger volumes of unstructured safety narratives. Particular attention is given to the capabilities and limitations of existing methods and their suitability for different forms of safety data.

The chapter concludes by critically synthesising these two areas of literature to identify the methodological limitations and research gaps that inform the development of the analytical framework presented in Chapter 4.

3.2 Literature Review Scope and Method

The literature review was structured around two complementary search strands reflecting the two methodological requirements of the research. The first examined structured approaches for analysing human and organisational factors in maritime accidents, including human factors taxonomies, accident causation models and related analytical frameworks. Its purpose was to evaluate how contributing

factors are currently identified and structured and to determine an appropriate basis for systematic human-factors analysis within this research. The second strand examined computational approaches for analysing unstructured safety narratives, including Natural Language Processing (NLP), Machine Learning (ML), topic modelling, transformer-based language models and related analytical techniques. This strand evaluated the extent to which existing computational methods can support automated, scalable extraction, classification, and analysis of safety information.

A structured literature search was undertaken across ScienceDirect, Scopus and Web of Science. Using multiple databases was intended to broaden coverage across maritime safety, human factors, and computational research and reduce the risk of omitting relevant studies by relying on a single database. The two search strands used separate keyword combinations reflecting their different scopes. For the accident-analysis strand, combinations included:

- “taxonomies” AND "human factors" AND "incident" AND “maritime”
- “human” AND “accident” AND “maritime”
- “human” AND “accident” AND “ship”

The computational strand used combinations relating safety and accident terminology to methods including:

- "Natural Language Processing" AND "accident investigation"
- "Natural Language Processing" AND "incident investigation"
- "Natural Language Processing" AND "incident report"

After database searching, we combined records from each strand and removed duplicates before title and abstract screening. Studies progressing beyond this stage underwent full-text assessment against strand-specific eligibility criteria. For the accident-analysis strand, studies were required to apply a human factors taxonomy, accident causation model, or other structured accident analysis approach to identify or analyse contributory factors within the maritime domain. For the computational strand,

studies were required to apply or evaluate computational text-analysis methods to accident, incident, near-miss or comparable safety narratives. Across both strands, included studies had to be original, peer-reviewed research published in English. The screening and selection process is summarised in Figure 7.

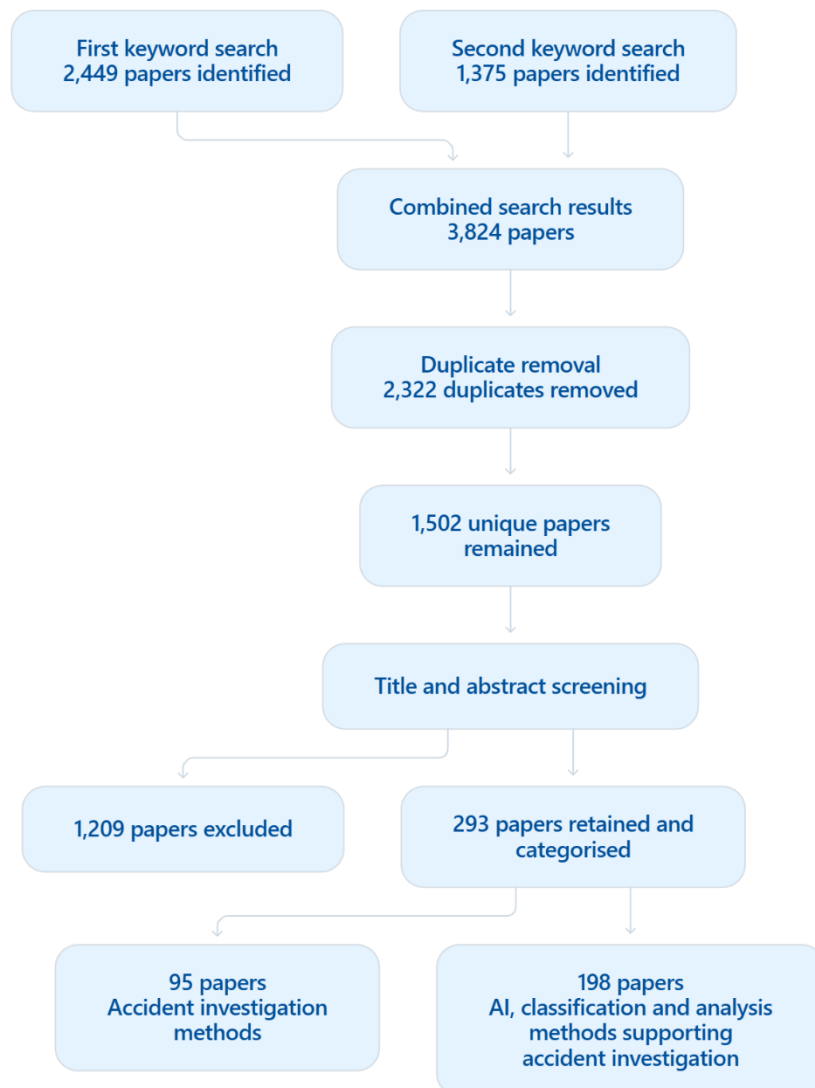


Figure 7. Literature search, screening and categorisation process for the two review strands.

The resulting literature was initially integrated to avoid duplicate papers, then classified and separated according to the two main goals of the review. Accident-analysis studies were grouped according to taxonomy-based, causation/system-based and other structured approaches and compared in terms of their analytical purpose, strengths, limitations, and suitability for systematic application. Computational studies were synthesised by principal analytical capability, including supervised

classification, unsupervised topic discovery, contextual language representation, and subsequent pattern or relationship analysis.

3.3 Accident Analysis Methods

The first review strand identified 95 studies concerning accident-investigation methods. The findings are organised into human-factors taxonomies, accident-causation and systems approaches, and other specialised methods. Rather than describing each study individually, the following sections compare how these approaches identify, structure and analyse contributory factors, with particular attention to their methodological strengths, limitations and suitability for systematic analysis.

3.3.1 Human Factors Taxonomies

Human factors taxonomies provide structured classification schemes for organising the human and organisational factors identified during accident investigations. Unlike accident causation models, which seek to explain how accidents develop, taxonomies provide more defined categories that enable consistent coding, comparison of incidents, and identification of recurring trends. This section critically reviews the principal human factors taxonomies applied within maritime safety research, focusing on their structure, applicability, limitations, and suitability for computational implementation.

3.3.1.1 HFACS

The Human Factors Analysis and Classification System (HFACS), developed by Shappell and Wiegmann (2000) from Reason's Swiss Cheese Model, is one of the most established frameworks for classifying human factors in accident investigation. It structures contributory factors across four levels (shown in Figure 8): acts, preconditions, supervision, and organisational influences (based on DOD HFACS Version 7.0). Within the acts layer, there are three subcategories: performance-based errors, judgement and decision-making errors and violations. Each of these is further divided into subcategories example within performance-based errors we have unintended operation of equipment, checklist not followed correctly, procedure not followed correctly, over-controlled/under-controlled aircraft/vehicle, breakdown in visual scan, rushed or delayed necessary action. This type of subgrouping is present in

each of the four layers (HE, 2020). Its hierarchical structure provides a systematic way to move beyond immediate operator error to the supervisory and organisational conditions that contribute to accidents.

The Human Factors Analysis and Classification System

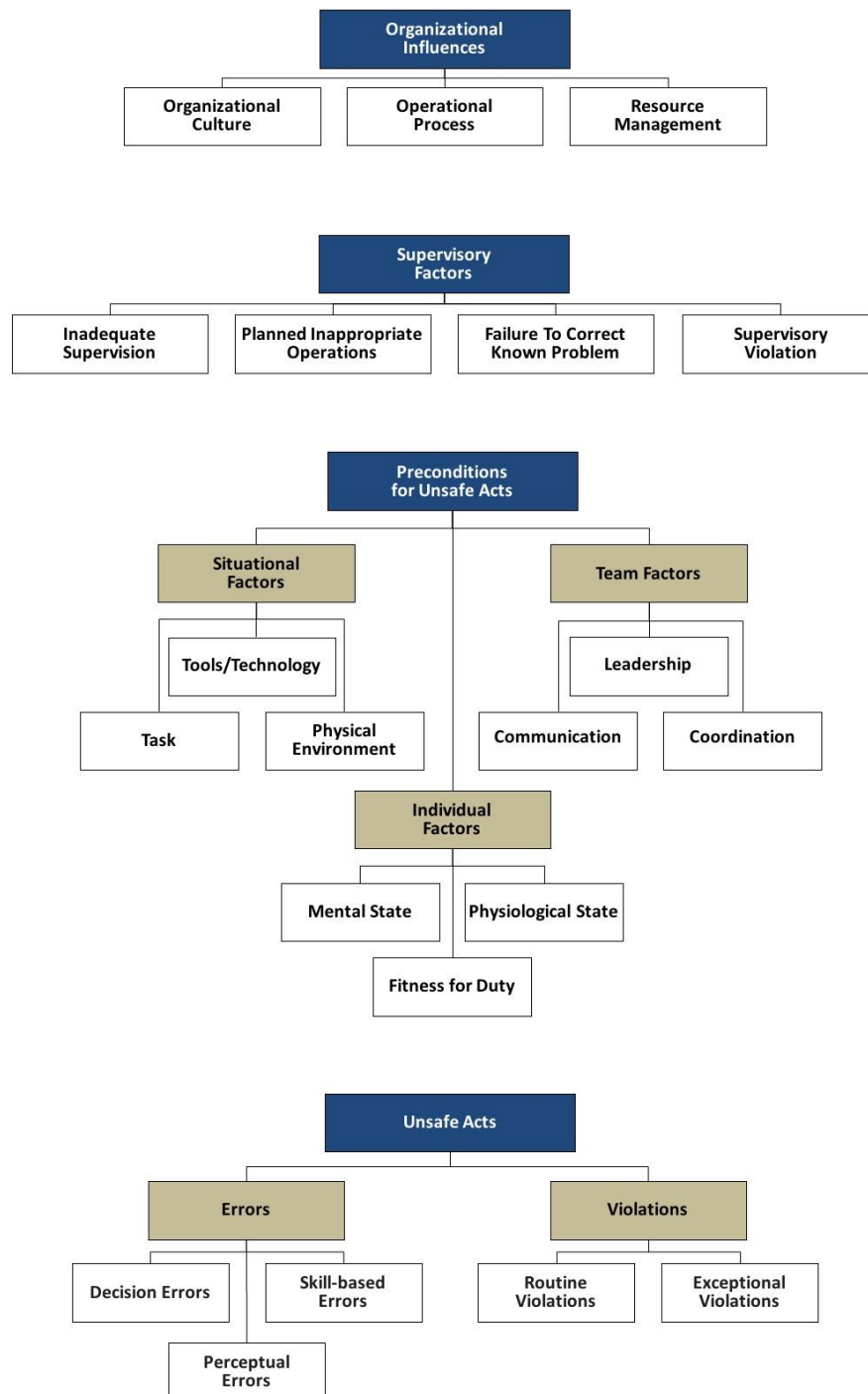


Figure 8. HFACS Framework (Shappell and Wiegmann, 2000).

HFACS has been applied in maritime to investigate various accident types, ranging from collisions and groundings to machinery failures. A common application of HFACS is evaluating the contributing factors behind maritime accidents and conducting statistical analyses to identify trends. For instance, Maternová and Materna (2023) analysed MAIB (Maritime Accident Investigation Branch) and TSB (Transport Safety Board) Canada reports and found that maritime incidents often involve multiple causal factors. Their analysis identified the most frequent error types and led to recommendations for preventing the recurrence of these errors through targeted interventions.

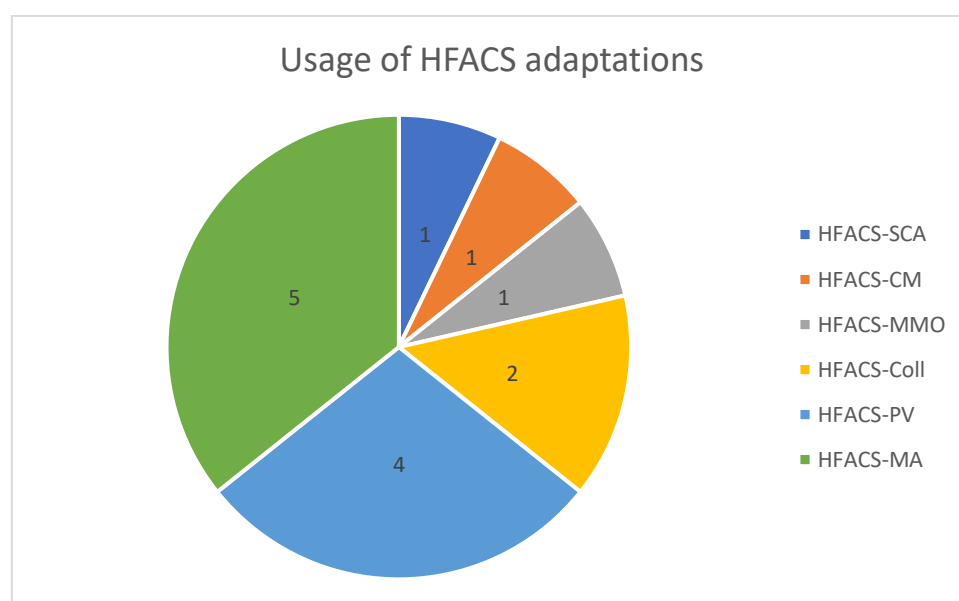


Figure 9. Distribution of HFACS variants/adapted frameworks in maritime papers.

Its adaptability has also led to several variations tailored for maritime incidents, including HFACS-MA (Maritime Accidents), HFACS-Ground (for grounding accidents), HFACS-Coll (for collision accidents) (Chauvin et al., 2013), HFACS-PV (for passenger vessels) (Uğurlu et al., 2018), HFACS-MSS (machinery spaces on ships) (Schröder-Hinrichs et al., 2011), HFACS-MMO (marine maintenance operations) (Kandemir and Celik, 2021) and others, Figure 9 shows the distribution of how these have been used. These applications generally retain the hierarchical principles of the original taxonomy while modifying or extending categories to better represent maritime-specific operational conditions. For example, HFACS-MA integrates a new level for external factors, including legislative gaps and administrative

oversight (Chen et al., 2013), while HFACS-PV incorporates operational conditions to better represent factors specific to passenger-vessel operations (Yildiz et al., 2021, Loughney et al., 2022). More recently, HFACS-SCA further extended the taxonomy for ship collision by incorporating external factors, including legislation, administration and design, and expanding preconditions to represent interactions between personnel, vessels and the operating environment (Mingyang et al., 2025). Liu et al. (2022) also adapted HFACS for collision accidents and used chi-squared and odds-ratio analysis on 168 vessels to identify relationships between factors across its hierarchical levels. An alternative adaptation was suggested by Batalden and Sydnese (2014), who adapted HFACS by separating unsafe supervision into shipboard and shore-based management, before cross-coding 478 causal factors from 94 MAIB cases against the functional sections of the ISM Code. This linked human and organisational failures to specific safety-management requirements, although the ISM coding remained based on qualitative analyst judgement.

These adaptations demonstrate the flexibility of the HFACS hierarchy but also indicate limitations in transferring the original aviation-derived taxonomy directly to maritime accident analysis. In particular, the repeated addition of operational and contextual factors specific to particular maritime accident types suggests that the original framework may not adequately represent the wider contextual and system-level conditions contributing to maritime accidents. A further limitation concerns the source information available for classification; while HFACS provides a useful, structured framework, the quality and depth of analysis remain highly dependent on the level of detail available within accident reports, particularly for organisational and supervisory factors (Mazaheri et al., 2015, Schröder-Hinrichs et al., 2011). A review of the different adaptations of HFACS was explored by Kaptan et al. (2021a). They explain that most modifications to the taxonomy involve only minor adjustments, aimed at making the framework more relevant to specific types of accidents. The largest change in the taxonomy was the addition of a fifth level which they state as “operational factors” which includes inappropriate equipment and ergonomic design faults which they no longer consider an organisational issue.

Beyond adaptations to the taxonomy itself, a prominent development in maritime applications has been the integration of HFACS with quantitative and probabilistic methods. Bayesian and network-based approaches have been particularly common, allowing dependencies between HFACS factors to be modelled while accommodating uncertainty and sparse accident data (Adumene et al. (2022), Wang et al. (2024a), Bayazit and Kaptan (2024), (Akhtar and Utne, 2014, Kaptan et al., 2021b, He et al., 2022)). Human Reliability Assessment (HRA) methods, including HEART and SLIM, have similarly been combined with HFACS to quantify human-error probabilities and adapt error-producing conditions to maritime operations (Akyuz et al., 2016)(Qiao et al., 2024)(Kandemir and Celik, 2021)(Aydin et al., 2022). Chen (2020) combined HFACS-MA with Why-Because Analysis and Grey Relational Analysis to identify common human and organisational factors across multiple accidents while retaining the causal detail of individual cases. These studies collectively suggest that HFACS is rarely used as a standalone framework in modern maritime research. Instead, researchers increasingly integrate HFACS with quantitative risk and reliability methods to improve predictive capability and support probabilistic assessment of human error.

A related trend is the use of HFACS to examine relationships between contributory factors rather than treating individual categories independently. Cognitive mapping (Akyuz and Celik, 2014), complex networks (Chen et al., 2019)(Zhang et al., 2024)(Ma et al., 2024b), system-dynamics modelling (Zhang et al., 2020), Bayesian Networks (Ma et al., 2024a)(Ma et al., 2024c)(Jiang et al., 2025b, Uğurlu et al., 2020, Li et al., 2022b), Analytical Network Process (ANP) (Bicen and Celik, 2022)(Akyuz, 2017), decision-making trial and evaluation laboratory (DEMATEL) (Ma et al., 2022), Association Rule Mining (ARM) (Lan et al., 2022, Xu et al., 2023) and clustering (Paolo et al., 2021) have all been used to identify dependencies and recurring combinations across HFACS levels. For example, association-based approaches allow relationships between human factors and contextual accident characteristics, such as vessel type, time and location, to be examined (Lan et al., 2022). Collectively, this literature reflects a shift from descriptive classification toward understanding interactions among human, organisational, and operational factors.

HFACS has also been combined with Fault tree analysis (FTA) and fuzzy approaches to quantify causal relationships while accounting for uncertainty in expert judgement (Zhang et al., 2019b)(Saralioğlu et al., 2020)(Qiao et al., 2020)(Celik and Cebi, 2009).

More recently, HFACS research has begun to incorporate AI-based and automated approaches. Xingrong et al. (2025) integrated spatiotemporal analysis, causal inference and large language model (LLM)-assisted classification, using automated classification to support the coding of larger volumes of maritime accident reports. Similarly, Laihao et al. (2025) combined an adapted HFACS framework with association rule mining (ARM), Tree-Augmented Naïve Bayes (TAN) and Bayesian Networks to model the individual and joint influence of human and operational factors. These studies indicate an emerging transition from manually coded, descriptive HFACS applications to computational frameworks that support classification and subsequent pattern analysis. However, evidence for fully automated taxonomy-based analysis remains comparatively limited, and the reliability of automated classification against expert-coded human-factor labels requires further investigation.

Previous reviews reinforce both the prominence and limitations of HFACS within maritime research. HFACS has consistently been identified as one of the most widely applied maritime human-factors frameworks, attributed to its relatively simple hierarchical structure, limited training requirements and foundation in Reason's accident causation model (Shi et al., 2021)(Wu et al., 2022)(Hulme et al., 2019). However, these reviews also identify limitations in the availability and quality of maritime safety data and comparatively limited consideration of higher-level influences (Shi et al., 2021)(Hulme et al., 2019). More recent bibliometric evidence indicates a broader shift towards big-data and machine-learning approaches within maritime accident research (Cao et al., 2023), suggesting increasing interests in addressing some of these scalability limitations associated with conventional manual analysis.

Overall, HFACS is considered a well-established framework for categorising human error. The process is shown to be easily understandable and very adaptable. Its structured hierarchical design makes the

framework particularly suitable for AI-driven approaches such as machine learning and natural language processing. The clearly defined categories and subcategories provide a foundation for converting unstructured accident narratives into structured and analysable safety data. However, within the maritime domain incident reports often lack sufficient detail limiting the depth of potential HFACS analyses, particularly at organisational and supervisory levels. For use in the maritime industry, modified versions may therefore be more appropriate, as the original aviation-focused framework may miss domain-specific operational and environmental factors. Lastly, although the framework aims to reduce subjectivity, incorrect classifications are one of the biggest issues with the method, as the same incident may be coded differently by multiple analysts (Galieriková, 2019).

3.3.1.2 Organisation Specific Taxonomies

Organisation-specific taxonomies, such as those used by MAIB and the IMO, provide classifications closely aligned with operational accident investigation and therefore represent an important source of structured maritime safety information. However, their categories are developed for specific organisational purposes and may differ in terminology, granularity and treatment of human factors, limiting consistency and comparison across reporting systems. A study by Alan and Sogut (2020) focused on grounding accidents using MAIB's database. They evaluated the breakdown of human factors to understand the most common issues leading to grounding incidents and using this information to detail prevention methods. The study explains that other human factors taxonomies, such as HFACS, may evaluate human errors more effectively. Similarly, Navas de Maya et al. (2021) identified duplication and ambiguity among 94 MAIB human factor labels and used card sorting to reorganise them into 11 clearer categories, illustrating the challenges associated with maintaining consistent organisational classifications. Collectively, these studies demonstrate that organisation-specific taxonomies provide valuable operationally grounded safety data, but may require consolidation or alignment with more systematic human-factors frameworks where consistent large-scale analysis is required.

3.3.1.3 Ship Accident Root Cause Evaluation (SHARE)

The SHARE framework provides a maritime-specific approach for identifying the root causes of maritime incidents, incorporating both human and job-related factors. In its initial validation, SHARE accounted for 94.1% of the root causes identified within the evaluated accident reports, indicating strong coverage of the causal factors within the sample. The technique is not only for taxonomy usage but also follows into calculating risk using the Fuzzy Analytic Hierarchy Process (FAHP) (Kececi and Arslan, 2017). While SHARE provides a structured approach to identifying root causes, its categorisation of incidents into broad groups like human or job-related factors may lack the necessary granularity for more detailed analysis. By grouping such complex factors into these overarching categories, important nuances or specific contributory elements may be overlooked, limiting the depth of understanding required to fully address and prevent future incidents. Therefore, more detailed sub-categories may be needed to capture the intricacies of maritime accidents effectively.

3.3.1.4 Safety Human Incident and Error Learning Database (SHIELD)

Comparatively, SHIELD was created more recently during the SAFEMODE project, the taxonomy structure is shown in Figure 10. It was created based on NASA-HFACS, a version of HFACS specifically made for analysis of incidents in space operations by NASA) and HERA (Human Error in Air Traffic Management). The goal was to create a domain-generic human factors taxonomy that would be systematic and to report the findings in a way that trends and patterns could be analysed to help prevent incidents by making changes to policies, regulations, and design. The taxonomy has shown great potential, helping companies discover more than double the causal factors in safety investigations when compared to their usual HFACS-based approach (Stroeve et al., 2023). This is because SHIELD offers a deeper understanding of risk factors through the multi-layered approach. The method also requires minimal training due to the well-defined, clear categorisations. However, challenges of time-consuming analysis mean that there is a relatively small number of analysed incidents, limiting the statistical significance of the findings.



Figure 10. SHIELD Taxonomy framework Iceberg visualisation (Stroeve et al., 2023).

3.3.2 System and Accident Causation Approaches

Unlike human-factors taxonomies, which classify contributory factors into predefined categories, accident causation and systems approaches explain how accidents develop through interactions among human, technical, organisational, and environmental conditions. These approaches range from relatively conceptual models of defence failure to systems-based methods that represent non-linear interactions, functional variability and control failures within complex socio-technical systems. This section critically evaluates the principal approaches used in maritime safety, focusing on their ability

to represent accident causation, their practical limitations, and their suitability for systematic analysis of accident data.

3.3.2.1 CREAM

Human reliability approaches provide another means of examining the contextual conditions influencing human performance. The Cognitive Reliability and Error Analysis Method (CREAM), developed by (Hollnagel, 1998), is a second-generation human reliability analysis method that considers human performance through interactions between human, technological and organisational conditions. Central to CREAM are Common Performance Conditions (CPCs), which represent contextual factors affecting performance and the dependencies between them. CREAM can support both qualitative and quantitative analysis, including retrospective examination of accident causes and prospective assessment of human reliability.

Maritime applications have demonstrated the need to adapt these conditions to the operational context. Yoshimura et al. (2014) adapted CPCs for marine collision analysis, including navigation conditions, safety management and communication, while Mitomo et al. (2012) used simulator experiments to examine communication with other vessels as an additional maritime-specific condition. Subsequent work developed CREAM-based support for retrospective marine accident investigation, while also identifying the specialist human-factors knowledge required to construct probable cause-effect relationships (Yoshimura et al., 2015). Quantitative extensions have similarly integrated modified CREAM approaches with probabilistic methods for maritime human-error assessment (Ung, 2018, Yang et al., 2013). Collectively, these studies demonstrate CREAM's ability to incorporate contextual influences on human performance, but also highlight its dependence on domain adaptation, detailed information and expert judgement, potentially constraining consistent application across large accident datasets.

3.3.2.1 Swiss Cheese Model

The Swiss Cheese Model is one of the most well-known human factors theories developed by Reason (1990). The theory visualises the defence layers an organisation has as slices of cheese, with each hole representing weaknesses or failures in the layers. The key idea behind this is that accidents will occur when the holes in these layers align, which allows hazards to pass through and cause an incident. The model has strongly influenced subsequent human factors taxonomies, most notably HFACS, and has also been used as a framework for accident analysis (Antão and Guedes Soares, 2008). However, the model provides limited guidance for systematically identifying or classifying causal factors, making its application dependent on analyst interpretation. Its value in this context therefore lies primarily in providing a theoretical basis for multi-level accident causation rather than a directly implementable analytical framework.

3.3.2.2 Functional Resonance Analysis Method (FRAM)

FRAM is a systems modelling approach that represents how work is performed within complex socio-technical systems. Rather than modelling accidents as linear sequences of failures, FRAM describes system performance in terms of interacting functions and the variability of everyday work. Although FRAM is not itself an accident investigation or risk assessment method, the models it produces have been applied to support accident analysis, risk assessment, and resilience engineering. Maritime applications have demonstrated its flexibility across areas including oil-spill analysis, dynamic positioning, general ship operations and emergency operations (Nasur et al., 2025)(Salihoglu and Bal Beşikçi, 2021)(Peng et al., 2023)(Lee and Chung, 2018)(Akyıldız and Mentés, 2023).

Lee et al. (2020) extended FRAM by incorporating human collaborative relationships, allowing formal and informal interactions within ships, between vessels and with shore-based personnel to be represented. This demonstrated FRAM's value in capturing temporary collaborative arrangements that conventional accident-analysis methods may treat as isolated individual actions. Additionally, Wicaksana et al. (2025) applied FRAM to a maritime near-miss to examine how operators successfully

adapted to disruption and prevented a collision. This illustrates FRAM's value in identifying recovery and resilience mechanisms, rather than focusing exclusively on failures that produced adverse outcomes. Other extensions of FRAM include converting FRAM functional signatures into graphs and using structural-similarity analysis to detect unusual operational patterns that were not identified through conventional statistical measures (Danial et al., 2021).

FRAM's principal strength is its ability to represent non-linear interactions and variability that conventional cause-effect models may overlook. However, this flexibility also introduces methodological challenges. FRAM construction remains dependent on expert interpretation, Salihoglu and Bal Beşikçi (2021) identified subjectivity as a limitation and recommended integrating complementary approaches to improve reliability. Its detailed functional modelling can also be time-consuming and difficult to standardise across large numbers of accidents. Recent approaches have begun combining FRAM with text mining and association analysis to support more data-driven identification of system interactions (Yuhong et al., 2025), indicating potential for computational support. Nevertheless, FRAM is primarily suited to explaining interactions within specific socio-technical systems rather than providing consistent predefined categories for large-scale classification of accident narratives.

3.3.2.3 AcciMap

AcciMap is a system-based accident causation model that graphically maps interactions among actors, organisations, technologies, and regulatory influences that contribute to an accident (Svedung and Ramsmussen, 2002). Its principal strength is the ability to visualise how decisions and actions at different organisational levels interact, extending accident analysis beyond immediate operator actions. Maritime applications have demonstrated its usefulness for representing system-level relationships (Cerqueira et al., 2019) and it has also been combined with quantitative methods to support prioritisation and risk analysis (Lee et al., 2017)(Akyuz, 2015). Morel and Chauvin (2006) combined an abstraction hierarchy with Generic AcciMap to analyse fishing-vessel collisions,

demonstrating how commercial and operational pressures can move everyday work beyond permeable safety boundaries. This illustrates AcciMap's value in connecting frontline actions with the wider regulatory, organisational and economic conditions shaping them.

However, the construction and interpretation of AcciMap remains substantially dependent on analyst judgement. Filho et al. (2019), comparing AcciMap and STAMP applications, reported relatively low consistency in identified causal factors and suggested that such systems approaches may function more effectively as tools for generating explanations than as highly reproducible diagnostic frameworks. Gucma et al. (2021) also noted that the model should be updated dynamically when operating conditions diverge from the vessel's original design assumptions, requiring risks to be reassessed as the operational context changes. Although quantitative extensions may support subsequent analysis (Jiaxin et al., 2025) the initial identification and structuring of causal relationships remains largely expert-driven. AcciMap therefore provides strong system-level representation but presents challenges for consistent and scalable application across large accident datasets.

3.3.2.4 System Theoretic Accident Model (STAMP) -CAST

STAMP is a systems-theoretic accident causation model developed by Leveson (2004) for analysing accidents with complex socio-technical systems. Rather than viewing accidents as linear chains of failures, STAMP considers causes arising from inadequate system control and feedback. It aims to identify the system's control structure, including systems design and operations controls. Causal Analysis Based on STAMP Theory (CAST) uses the theoretical base of STAMP but goes beyond the analysis, providing a holistic view of the event. For this reason, CAST is usually preferred to the basic STAMP model. Maritime applications demonstrate its value for analysing complex events where conventional linear cause-effect explanations may be insufficient (Ceylan et al., 2021)(Ceylan et al., 2022)(Kim et al., 2016).

STAMP-CAST is well suited for analysing complex socio-technical maritime systems and high-level organisational interactions. The reviewed studies demonstrate its strength in identifying system

control failures and broader operational influences beyond simple cause-effect relationships. However, the methodology can be resource-intensive and difficult to apply consistently in routine investigations. Its systems-oriented structure may nevertheless support future AI-driven modelling and network analysis approaches.

Additionally, Uddin and Awal (2020) applied System-Theoretic Process Analysis (STPA) to identify unsafe control actions within inland passenger-ship operations and derive corresponding safety requirements. This extends the system-theoretic approach from retrospective accident analysis through CAST to proactive hazard identification before an accident occurs.

3.3.2.5 Other Approaches

Several more focused approaches have also been applied within maritime accident analysis. The Bow-tie model is an accident analysis and risk management method that combines elements of fault tree analysis and event tree analysis to illustrate the pathways leading to an accident and the barriers available to prevent or mitigate its consequences. The method visually represents the relationship between hazards, initiating events, preventive controls, the central accident event, and mitigation measures, making it particularly useful for identifying weaknesses in safety barriers and supporting risk management. However, its relatively linear representation may oversimplify the interaction characteristics of complex socio-technical accidents and still depends on expert identification of hazards and barriers. Recent combinations with qualitative coding approaches have attempted to make this process more systematic (Dingding et al., 2025).

TRACER and SHEL focus more specifically on human performance and interaction. TRACER is a cognitive human error analysis framework developed by Shorrock and Kirwan (2002) for air traffic control. It provides a structured approach to analysing cognitive errors, human-machine interaction, and performance-shaping factors in both predictive and retrospective assessments. The framework has been employed as a systematic method for analysing human errors in grounding and collision accidents (Graziano et al., 2016)(Sotiralis et al., 2016). However, its detailed cognitive categorisation

requires substantial analyst expertise, while limited guidance for reconstructing task errors can affect consistency. Some studies have extended TRACER, integrating the Information, Decision and Action (IDA) model with clustering and association-rule mining, allowing task errors to be linked systematically to underlying cognitive failures and performance-shaping factors (Ma et al., 2024d).

SHEL similarly is a human factors interaction framework designed to analyse the interfaces between people, equipment, procedures, and the operating environment. Rather than classifying the contributory factors or modelling accident causation directly, SHEL focuses on identifying mismatches between system components that may contribute to human error. The method was initially created for use in aviation studies by Hawkins (1993) and was made to help identify potential gaps that could lead to human error by focusing on how humans interact with their environment and technology. Chen et al. (2020a) then tried to further expand on Reason's model by integrating the Swiss Cheese theory with SHEL (Software, Hardware, Environment Liveware). The integration of SHEL to Reason's model provides a more detailed understanding of how failures at different system levels interact to cause accidents. For instance, missing certifications are a software-to-liveware issue, while poor communication between crew members is a liveware-to-liveware issue. However, despite its value in identifying interactions between people, equipment, procedures, and the environment, the SHEL model has been criticised for providing a largely static representation of human performance and for placing less emphasis on wider organisational and systemic influences that contribute to accidents.

Alternatively, other qualitative HRA-based approaches were also proposed. Bowo et al. (2019) proposed a hybrid HEART-4M methodology that classifies error-producing conditions across human, machine, environmental and management factors while retaining quantitative human-error estimates. This supports more targeted mitigation, although the weighting of these conditions remains dependent on analyst judgement. Another approach, DEMATEL, has also been used in the research. This is used to model directional relationships and was applied to a study to evaluate accident factors

during pilotage (Demirci et al., 2023). The method supports intervention prioritisation but again remains dependent on expert pairwise judgement.

3.3.3 Comparison of Accident Analysis Approaches

This review, spanning four decades, reveals a notable trend toward increased publication frequency in recent years. This increase indicates a growing interest in human factors and safety analysis within maritime research, particularly as the industry faces evolving challenges. HFACS's dominance in the reviewed literature reflects its effectiveness and widespread acceptance as a primary human factors taxonomy for classifying contributory factors in maritime incidents. HFACS is easily adaptable, as shown by the adaptations for specific maritime incidents explored in this study. However, the method has limitations such as the degree of subjectivity, inconsistent coding can pose significant issues with reliability of the method. Although highly adaptable, many studies proposed modifications to the original aviation-based taxonomy to better capture maritime operational and environmental factors, suggesting that a universally applicable "off-the-shelf" version is not yet available for all maritime applications.

Unlike HFACS, which provides predefined classification categories, FRAM models functional variability and system interactions. The method's key strength is its ability to model complex interactions within maritime operations. Some studies highlight the method's ability to provide deeper insights than conventional accident analysis tools. However, the model's flexibility is also a weakness, as it is highly subjective and lacks the structure seen in HFACS. Analysts using the framework need to be skilled in interpreting functional variability, which can lead to inconsistent or unreliable findings if not managed carefully. Combining the method with more structured frameworks is often recommended for a balanced approach.

As an accident causation model, AcciMap offers a broader view of system-level interactions, graphically representing the relationships between various influencing factors. The method helps visualise interdependencies but tends to focus more on system-level suggestions than individual-level

interventions. However, the method has been criticised for its relatively low reliability, as found by Filho et al. (2019). Moreover, the limited validity and tendency to generate ideas rather than concrete diagnostics suggest that it may be more suited for preliminary analysis rather than final decision-making.

Unlike the broader accident causation models reviewed, TRACEr offers a psychological and cognitive perspective on human errors, particularly in how individuals interact with their environment and technology. Despite its effectiveness in analysing cognitive errors, TRACEr struggles to reconstruct complex event sequences and offers little guidance on structuring task errors.

STAMP-CAST is another methodology well suited to analysing complex socio-technical systems. Its focus on system design and operations makes it an excellent tool for understanding how various factors interact in complex maritime systems. CAST, in particular, adds value by moving beyond analysis to provide a holistic view of the incident and offer detailed system-level recommendations. However, the strength in addressing complex accidents can also be its limitation in more straightforward scenarios, where a simpler model may suffice.

All of these methods face challenges related to inherent subjectivity and reliance on the analyst's expertise and judgment. This can lead to inconsistent findings and undermine the overall reliability of the analysis. Additionally, organisational factors are often the most under-reported contributors in accident investigations, as many analyses fail to delve deeply enough into the organisation's safety culture or history, while organisations themselves may be reluctant to expose these factors for fear of blame. Multiple studies, including those by Schröder-Hinrichs et al. (2011), have highlighted that although organisational issues are likely to be underlying factors in many incidents, they are rarely reported. Consequently, taxonomies and frameworks designed to identify organisational problems may be less effective in capturing and addressing these crucial elements, limiting their ability to provide a full picture of incident causation.

Table 1. Overall comparison of accident causation analysis methods,

Methodology	Focus	Strengths	Weaknesses
HFACS	Human error categorisation	Structure, multi-level analysis, easy to understand, adaptable	Over-simplification, limited non-human focus
AcciMap	System interactions	Visual representation, broad factor inclusion	Reliability issues, complex interpretation
FRAM	Functional interactions	Holistic, flexible	Subjective, complex application
TRACer	Cognitive processes	Predictive and retrospective, cognitive focus	Analyst dependency
STAMP/CAST	Socio-technical systems	Holistic	Implementation complexity, potential oversights
SHARE	Root cause analysis	High validity rate, comprehensive approach	Broad categorisation, lack of granularity
SHIELD	Incident learning	Domain-generic, in- depth causal analysis	Time-consuming, subjective

Furthermore, a significant drawback of these methodologies is their time-consuming nature. Conducting thorough, in-depth analyses, whether using SHIELD's multi-layered structure or STAMP-CAST's system theory approach, requires substantial time and resources. This makes it difficult to implement the methods regularly in day-to-day investigations (Kristiansen, 2005). So, despite this

ability to offer a detailed understanding of accident causation, the extensive time and effort involved often prevent these analyses from being conducted on a routine basis, limiting their practical application.

Among the reviewed human factors taxonomies, SHIELD offers several characteristics that make it particularly suitable for AI-assisted classification. Unlike many of the models and frameworks, SHIELD provides a domain-generic and hierarchical structure capable of representing both human and organisational factors across a wide range of maritime operations. This structured yet flexible design provides a strong foundation for machine implementation, allowing the taxonomy to be adapted for automated, consistent and interpretable classification of human and organisational factors from large collections of unstructured accident reports.

A clear trend emerging from the reviewed literature is the increasing integration of traditional accident analysis methods with artificial intelligence and data-driven techniques. Recent studies have combined HFACS, FRAM, AcciMap and Bayesian Networks with natural language processing, machine learning, knowledge graphs and probabilistic modelling to improve scalability, reduce subjectivity and enable the analysis of increasingly large collections of accident reports. Despite these advances, most approaches still rely on manual identification or coding of human and organisational factors before automated analysis can be performed, highlighting the need for frameworks capable of directly classifying contributory factors from unstructured safety reports.

3.4 AI-Based Analysis of Safety Narratives

This review strand examined 198 studies applying AI and computational methods to the classification and analysis of accident, incident and other safety narratives. The literature is evaluated according to the main stages of the analytical process: text representation and feature extraction, supervised classification, unsupervised pattern discovery, contextual language modelling, and relationship analysis. Attention is given to the extent to which these methods support scalable safety analysis while preserving domain-specific meaning, interpretability and classification reliability.

3.4.1 Text Representation and Feature Extraction

3.4.1.1 Text Pre-processing

Text pre-processing is a critical step in all natural language processing workflows, regardless of whether the downstream model is supervised, unsupervised, or deep learning based. It ensures that text is converted into a form that is more suitable for modelling, improving both efficiency and model performance. Common operations include stop word removal, stemming/lemmatisation (Liang et al., 2020, Borjali et al., 2021), lowercasing, whitespace and punctuation removal, removal of numbers, keyword filtering, tokenisation, sentence splitting, Part-of-speech (POS) tagging, automated preprocessors, rule-based abbreviation toolkits, term-frequency filtering, spelling corrections and domain-specific dictionaries. One recurring theme is the value of domain-specific preprocessing. Liang et al. (2020), for example, developed a custom list for their safety domain-specific corpus using the Rainbow algorithm, while Tanguy et al. (2016) used a rule-based toolkit containing manually defined rules for abbreviations and terminology. Similarly, Zou et al. (2017) manually developed a domain-specific lexicon and expanded it using Word2Vec, whereas Na et al. (2021) proposed a more data-driven lexicon-development approach based on information entropy. These studies suggest that generic preprocessing pipelines may not adequately capture specialist terminology within safety narratives and that incorporating domain knowledge can improve the relevance of extracted information.

A second theme concerns the trade-off between reducing noise and removing useful information. Term-frequency filtering can reduce sparsity, as demonstrated by Evans et al. (2020), while keyword filtering has been shown to reduce class-related bias in some applications (Ong et al., 2010). However, aggressive removal of words, numbers, or low-frequency terms may also eliminate information that is operationally important within specialist domains, such as vessel identifiers, technical abbreviations or rare hazard terminology. Preprocessing choices therefore need to be aligned with the characteristics of the reporting systems rather than applied uniformly.

Finally, several studies demonstrate increasing use of more automated and linguistically informed preprocessing, including POS tagging and parsing (Grouin et al., 2013), spelling correction (Mathew et al., 2021) and N-gram tokenisation (Evans et al., 2020). While these approaches can improve consistency in noisy free-text data, they also introduce additional complexity and may require domain-specific adaptation. Overall, the literature indicates a shift from generic text cleaning towards preprocessing strategies that preserve domain meaning and support the specific requirements of the downstream classification task.

3.4.1.2 Feature and Information Extraction

Feature extraction provides a means of transforming narrative text into variables or representations that can be used for subsequent analysis. Earlier approaches frequently relied on manually or statistically identified textual attributes. Tixier et al. (2016) for example, combined a domain-specific dictionary with detection rules to extract precursor conditions and accident attributes from construction narratives, achieving precision and recall of 95% and 97%, respectively. Statistical approaches can reduce reliance on predefined dictionaries; Luo et al. (2023) applied mutual information and information entropy to segment and extract high-information word strings during preprocessing. These statistical measures helped identify word combinations with strong co-occurrence relevance and uncertainty reduction, enhancing the quality of feature input for accident severity prediction.

Named Entity Recognition (NER) extends feature extraction by identifying predefined entities directly within narrative text and can therefore support conversion of unstructured reports into structured safety information. Applications have included extraction of causal and consequence information (Kamil et al., 2023b) and crash-related entities (Idakwo et al., 2025). Importantly, both studies demonstrate the value of domain adaptation; customised NER models outperform generic or simpler baseline approaches, indicating that general-purpose entity definitions may not adequately represent specialised safety terminology. External knowledge sources can also enrich entity extraction; Single et al. (2020) for example, incorporated PubChem information to support hazardous-attribute identification and ontology development. Maritime applications similarly demonstrate the value of combining domain knowledge with statistical extraction. Mackenzie et al. (2021) combined rule-based NER heuristics with statistical classifiers for maritime security narratives, reporting F1 scores of 82-97% across selected incident and response categories.

3.4.1.3 Text Representation and Embeddings

Text representation determines how narrative information is converted into numerical features for subsequent classification or pattern analysis. Approaches identified within the reviewed literature range from sparse lexical representations, such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), to static semantic embeddings including Word2Vec and Doc2Vec and, more recently, contextual representations derived from transformer models such as BERT and RoBERTa. TF-IDF remains widely used within safety-related NLP because of its computational simplicity and ability to identify terms that discriminate between documents or categories (Huang et al., 2022, Hughes et al., 2019).

However, lexical representations have limited ability to capture semantic context. Yokota et al. (2018), for example, found substantial overlap between the most important BoW features associated with two incident categories, limiting their discriminatory capability and motivating the use of representations capable of retaining greater contextual information. Word2Vec and Doc2Vec provide one means of

addressing this limitation by representing semantic similarities between words or documents and have been applied to accident narratives for this purpose (Madeira et al., 2021). Nevertheless, greater representational complexity does not necessarily produce superior classification performance. Macedo et al. (2023) found that TF-IDF combined with SVM outperformed BoW and Doc2Vec alternatives, demonstrating that relatively simple lexical representations can remain competitive where distinctive terminology is strongly associated with the target categories.

More recent research has increasingly employed contextual transformer representations. BERT and related models generate representations according to surrounding linguistic context and have been used either as feature generators for downstream classifiers or as components of end-to-end classification architectures (Macêdo et al., 2022)(Chen et al., 2022)(Ramos et al., 2024). Transformer-based representations have also been combined with clustering and other analytical approaches, including maritime applications examining human-element interactions and near-miss narratives (Nurduhan and Kuleyin, 2024)(Chen et al., 2025b). Sentence-level representations such as SBERT further enable semantic similarity between complete narratives or passages to be examined (Elizabeth et al., 2025). Collectively, these applications demonstrate the flexibility of contextual embeddings as a common representation that can support both supervised classification and unsupervised analysis.

Comparative evidence also indicates potential advantages from contextual and domain-adapted representations. Tawfik and Spruit (2020) found contextual embeddings such as ELMo and BioBERT generally outperformed static alternatives across biomedical NLP tasks, while the performance of domain-specific BioBERT demonstrates the potential value of adapting representations to specialist terminology. However, improved semantic capability comes with additional computational requirements. Parikh et al. (2024), for example, obtained similar results using BERT and the smaller DistilBERT model despite the greater computational cost of BERT. Furthermore, transformer representations do not necessarily remove the need for appropriate downstream model selection:

Chen et al. (2024) obtained their strongest patient-safety classification performance by combining RoBERTa representations with SVM and incorporated LIME to improve prediction explainability.

Overall, lexical and contextual representations offer complementary strengths. TF-IDF provides an efficient and interpretable representation of distinctive terminology, while transformer-based representations capture semantic context and can be adapted to specialist domain language. As greater model complexity does not consistently improve performance, evaluating both representations provides a stronger basis for safety-narrative classification than assuming one approach is universally superior.

3.4.1.4 Rule-based Learning

Rule-based learning and keyword extraction represent some of the most traditional and interpretable approaches to natural language processing. These methods rely on predefined rules, keyword lists, or linguistic patterns to extract and classify information from unstructured text. While they are often less flexible than machine learning or deep learning models, rule-based methods offer high transparency, lower computational requirements, and can be useful in low-data or safety-critical environments where interpretability is essential. Keyword and term-matching approaches have successfully been used to screen large collections of safety-related narratives and identify potentially relevant records (Forster et al., 2005)(Shen et al., 2022). However, their performance depends strongly on the completeness and specificity of predefined vocabulary, creating a risk that relevant reports expressed using different or more contextual language are not identified.

The limitation becomes particularly important when identifying causal and human factors rather than explicit accident characteristics. Hughes et al. (2018), for example, found greater classification errors for categories such as human error than for more directly identifiable conditions, reflecting the greater complexity of language associated with accident causation. This difficulty is also evident in maritime accident analysis. Tirunagari (2015) compared supervised and linguistic-pattern approaches for identifying causal-factor sentences in accident reports, with F1 scores of approximately 58–65%. The

relatively modest performance illustrates the difficulty of identifying causal information through explicit linguistic cues alone. Similarly, linguistic characteristics and keywords can reveal useful safety signals within narratives (Leurs et al., 2022), but their presence alone does not necessarily establish the contextual meaning or causal role of a factor. Rule-based approaches are therefore better suited to identifying explicit and consistently expressed information than abstract, implicit or multi-causal human and organisational factors.

To overcome these limitations, research has increasingly combined rules and keyword extraction with statistical or machine-learning approaches. Keyword extraction has been integrated with topic modelling to reduce dependence on labelled training data and improve topic identification (Ahadh et al., 2021)(Sarkar et al., 2023), while rule-based extraction has been combined with supervised classifiers to retain interpretability while allowing more adaptable classification (Yang and Demichela, 2023)(Sahoo et al., 2024b). Rule-based pre-screening can also improve downstream classification by enriching the relevant samples presented to ML models (Byon et al., 2023). These hybrid approaches demonstrate that rules can provide useful domain constraints or preliminary extraction without necessarily forming the complete analytical model.

Nevertheless, hybridisation does not fully resolve the difficulty of representing complex safety narratives. Chang and de Jesus Martinez (2024), applying combined rule-based NLP and ML to oil and gas incident reports, has repeated difficulties obtaining accurate predictions and attributed these partly to narrative variability, accident complexity, and limitations in available training data. Collectively, the literature suggests that rule-based methods offer valuable transparency and can perform effectively where relevant information is explicit and terminology is sufficiently stable but become less reliable as classification requires greater semantic and contextual interpretation. This limitation is particularly relevant to human-factors taxonomies, where contributory factors may be implicit in the narrative rather than represented by distinctive keywords, supporting the use of supervised and context-sensitive language models.

3.4.2 Unsupervised Analysis of Safety Narratives

Unsupervised learning refers to AI methods that find patterns in unlabelled data, making it ideal for domains like incident analysis where structured labels are scarce. It is used to identify clusters, themes and causal relationships in safety narratives, often serving as a preliminary step before classification or structured modelling.

3.4.2.1 Clustering

Clustering has been widely applied to identify latent structures within unlabelled accident and incident narratives by grouping records according to lexical or semantic similarity. Approaches identified in the reviewed literature include hierarchical clustering, k-means, and density-based methods such as DBSCAN, often combined with TF-IDF or embedding-based text representations. Hierarchical clustering has demonstrated utility for identifying recurring incident patterns and structuring causal or operational themes (Liu et al., 2019)(Abe et al., 2022)(Wang et al., 2024b). Within wind-energy safety, Gurdal Ertek et al. (2017) similarly applied hierarchical clustering to identify relationships within wind-turbine incident narratives and examine their association with temporal, geographical and component characteristics. More advanced applications have combined clustering with qualitative or systems-based analysis; Ma and Chen (2024), for example, clustered sentence embeddings before applying grounded theory and constructing an AcciMap, illustrating how unsupervised grouping can provide a preliminary structure for subsequent causal analysis. The studies demonstrate that hierarchical clustering is a flexible and interpretable approach for analysing accident narratives and identifying patterns within safety data. Its ability to group semantically similar incidents without requiring predefined categories makes it particularly suitable for exploratory analysis in domains where causal relationships are complex and not fully understood.

K-means offers greater simplicity and computational efficiency and has produced interpretable groupings when combined with semantically rich representations. For example, Nurduhan and Kuleyin (2024) combined BERT embeddings with k-means to identify maritime human-element clusters, as

illustrated in Figure 11, while Ono and Nakanishi (2023) applied a similar approach to aviation narratives. However, performance across safety datasets has been inconsistent. Studies have reported missed contextual information, overlapping clusters and weak separation of incident narratives (Miyamoto et al., 2022)(Rose et al., 2022)(Liu et al., 2021)(Khan et al., 2024). Comparisons with alternative approaches have also found k-means to be outperformed by co-occurrence networks, DBSCAN and STM in particular applications. These findings indicate that clustering performance depends strongly on both the representation of the underlying text and the semantic complexity of the narratives, rather than the clustering algorithm alone.

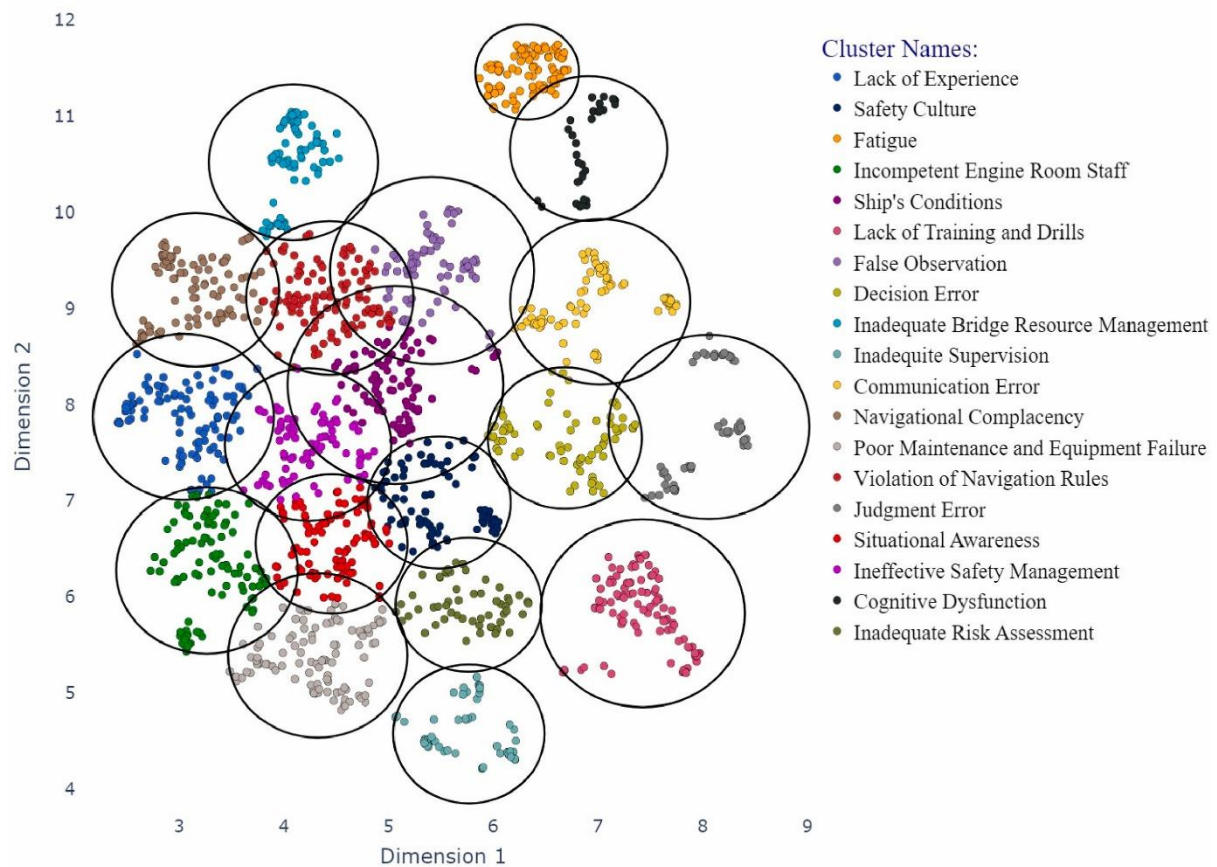


Figure 11. Human element relationship graph (Nurduhan and Kuleyin, 2024)

Alternative and hybrid clustering approaches attempt to address some of these limitations. DBSCAN can identify irregular clusters and treat atypical observations as noise without requiring a predefined number of clusters, making it potentially useful for heterogeneous safety datasets containing rare events. Its performance remains data-dependent: one study found agglomerative clustering slightly

more effective for document assignment, whereas Valcamonico et al. (2025) showed that DBSCAN combined with BERT embeddings could produce domain-specific groupings while retaining atypical cases. Hybrid approaches have similarly combined clustering with rule-based segmentation, dimensionality reduction and co-occurrence networks to improve interpretability (Nakata, 2017)(Nakata and Sohrab, 2018)(Liu et al., 2021). Clustering has additionally been used to evaluate predefined taxonomies, revealing categories that may be overly broad or obscure latent distinctions within safety data (Fujita et al., 2012)(Akiyama and Fujita, 2013).

Overall, clustering provides a useful exploratory mechanism where predefined labels are unavailable, but the resulting groups do not inherently constitute meaningful safety categories. Their quality depends on preprocessing, text representation, algorithm selection and the characteristics of the underlying narratives, while interpretation generally remains dependent on domain expertise. These limitations motivate approaches that retain unsupervised discovery while providing more explicit representations of the thematic content contained within groups of safety narratives.

3.4.2.2 Topic Modelling

Topic modelling is a widely used unsupervised natural language processing technique that identifies latent themes in large text corpora. It allows researchers to discover patterns and structures in unlabelled data, often revealing risk factors, causal mechanisms, or event sequences that are not immediately visible through manual review. The most commonly used topic modelling methods found in the reviewed literature are Latent Dirichlet Allocation (LDA), Structural Topic Modelling (STM), and the more recent transformer-based approach BERTopic. Several hybrid or extended approaches were also observed.

LDA remains widely applied because of its relative simplicity, interpretability and ability to identify recurring themes across large collections of safety narratives. Applications have used LDA to identify accident themes, risk factors and temporal patterns and to generate structured inputs for subsequent association, network and predictive analyses (Sahoo et al., 2024a)(Suh, 2021)(Yan et al., 2023)(Pereira

et al., 2013)(Mertes et al., 2024)(Wali et al., 2021). Its potential efficiency is illustrated by Li et al. (2021), who reported that automated topic extraction required approximately 10% of the time of manual analysis while recovering 95% of manually identified themes. LDA-derived information has also been incorporated into subsequent analytical pipelines, including deep-learning models (Xu et al., 2022)(Tingjiang et al., 2023), neuro-fuzzy risk prediction (Sarkar et al., 2023) and causal-network analysis (Cao et al., 2024b). These applications demonstrate that topic modelling can provide not only exploratory themes but also structured information for subsequent safety analysis.

Applications within wind-energy safety remain comparatively limited. Gurdal Ertek and Kailas (2021) applied LDA to wind-turbine incident narratives and identified ten recurring incident themes, demonstrating the feasibility of unsupervised thematic analysis within this domain. However, overlapping terminology reduced topic distinctiveness, reflecting a broader limitation of bag-of-words approaches, where related safety themes share similar vocabulary.

Comparative studies further support the continued relevance of LDA while also showing that no single topic-modelling approach is universally preferable. Williams and Betak (2018), comparing LDA and Latent Semantic Analysis (LSA) on railway accident reports, identified both overlapping and method-specific themes and concluded that the approaches could provide complementary insights. Xiong et al. (2024a) selected LDA over LSA and probabilistic LSA because of its interpretability and lower susceptibility to overfitting, while Costantino and Sabetta (2025) compared LDA with NMF, STM and an LLM-based approach for robotic incident analysis, demonstrating that LDA remained useful when combined with expert validation and semantic refinement. This suggests that newer or more semantically sophisticated methods do not necessarily eliminate the value of established probabilistic topic models; suitability depends on the analytical objective, data characteristics, and interpretability requirements.

However, LDA represents documents through word distributions rather than contextual semantic relationships, limiting its ability to distinguish meanings that depend on surrounding language.

Davidson et al. (2021) for example, found that identified themes could not distinguish between different contextual explanations for similar terminology, while Macedo et al. (2022) reported poor downstream classification performance from LDA-derived topics in noisy data. Attempts to incorporate seed or anchor words have improved topic relevance (Ahadh et al., 2021)(Bao et al., 2023), but these approaches introduce additional domain knowledge into what is otherwise an unsupervised process.

STM extends LDA by incorporating metadata (e.g. incident type, time, or severity), allowing for nuanced exploration of how topics vary across contextual dimensions. Applications to aviation and transport safety narratives have reported improved topical separation and interpretability relative to LDA and k-means, particularly where contextual metadata are analytically important (Kuhn, 2018)(Rose et al., 2022)(Xing et al., 2024)(Li et al., 2024a). This makes STM particularly valuable where the objective extends beyond identifying themes to understanding how their prevalence varies across operational or accident characteristics. The literature also demonstrates how STM-derived topics can provide structured inputs for subsequent safety analysis. Muguro et al. (2022) combined STM with keyword-scoring measures to map discovered topics onto predefined traffic-safety categories, while Kwayu et al. (2021) combined STM-derived information with random forest classification for crash prediction. Xu et al. (2025) extended this approach by constructing a co-occurrence network from topics identified in bus incident narratives (Figure 12), enabling examination of relationships between safety themes. The network illustrates how extracted themes can be transformed into structured relational information rather than treated as isolated topic outputs.

However, STM remains fundamentally based on bag-of-words representations and therefore has limited ability to represent semantic and syntactic relationships within the underlying narratives. Its main advantage over LDA is integrating metadata and contextual covariates, rather than substantially richer linguistic representation. Consequently, where safety narratives are short, terminology is specialised, or meaning depends strongly on linguistic context, embedding-based topic-modelling approaches may offer an alternative.

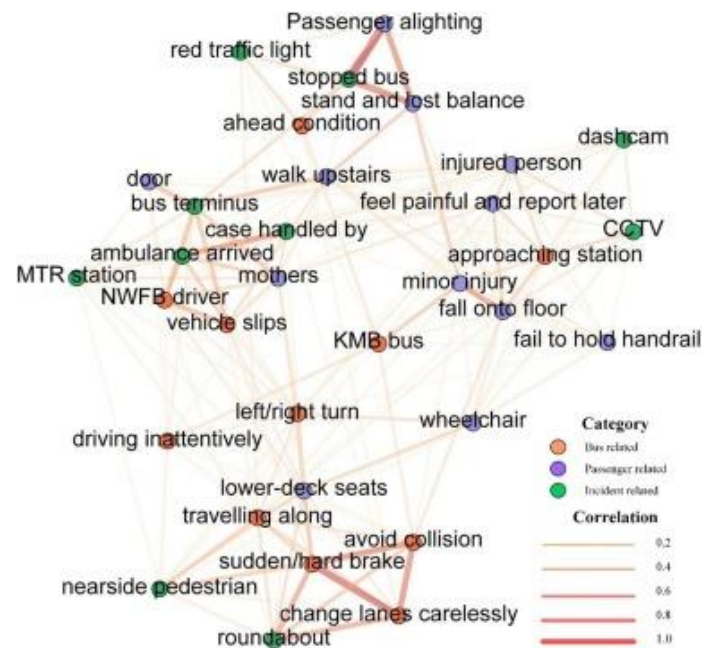


Figure 12. Topic co-occurrence network derived from STM applied to Hong Kong bus incident reports (Xu et al., 2025)

BERTopic is a more recent embedding-based approach that combines transformer sentence embeddings with dimensionality reduction and density-based clustering to identify semantically related topics. Unlike bag-of-words topic models, its initial grouping of documents is therefore informed by contextual semantic representations, making it potentially advantageous for short or specialised narratives where similar concepts may be expressed using different terminology. Safety applications have demonstrated its use for extracting hazard themes, examining relationships between topics and supporting semantic retrieval from incident collections (Andrade and Walsh, 2023)(Hong et al., 2024)(Irshad and Walsh, 2024). For example, Figure 13 illustrates the distinct hazard themes identified by Andrade and Walsh (2023), demonstrating how BERTopic can transform incident narratives into interpretable thematic groupings for subsequent safety analysis. Its application to maritime piracy reports further demonstrates emerging use within the maritime domain (Fahreza and Hirata, 2024).



Figure 13. BERTopic-derived hazard themes identified from wildfire incident reports using the HEAT framework (Andrade and Walsh, 2023).

These advantages do not eliminate the interpretability problem inherent in unsupervised analysis. BERTopic identifies semantically similar groups but does not independently establish their safety significance or causal meaning, meaning expert interpretation and validation remain important. The reviewed applications similarly show that the resulting topics are often supplemented by manual validation, semantic search, or subsequent analytical methods.

3.4.2.3 Unsupervised Model Comparison

The review of the literature suggests that the choice of topic modelling approach should be guided by the characteristics of the available safety data and the analytical objectives. As summarised in Table 2, traditional probabilistic approaches such as LDA remain attractive due to their interpretability, low computational requirements, and suitability for large corpora. STM extends these capabilities by incorporating contextual metadata, making it particularly valuable when analysing how themes vary across operational conditions or incident characteristics. More recently, transformer-based approaches such as BERTopic have shown superior performance for short, unstructured incident narratives because they can capture semantic relationships beyond bag-of-words assumptions. For maritime and offshore safety research, where reports are often short, domain-specific, and context-

rich, embedding-based approaches may offer significant advantages for uncovering nuanced safety themes and latent risk patterns.

Table 2. Comparison of topic modelling approaches in accident investigation research.

Feature/Capability	LDA	STM	BERTopic
Underlying approach	Probabilistic bag-of-words topic model	Extension of LDA incorporating metadata	Transformer embeddings and clustering
Semantic Understanding	Limited	Limited-Moderate	High
Interpretability	High	High	Moderate
Performance on short narratives	Often weak	Moderate	Strong
Handling of domain-specific terminology	Limited without tuning	Moderate	Strong with embeddings
Computational complexity	Low	Moderate	High
Main strengths	Simplicity, transparency, easy implementation	Context-aware topic discovery	Rich semantic understanding and clustering
Main limitations	Poor contextual understanding, bag-of-words assumptions	Still constrained by BoW assumptions	Computationally intensive, only one cluster grouping per document

Suitability for maritime/offshore safety reports	Good for large exploratory datasets	Strong for contextual safety analysis	Strong for near-miss and narrative-rich datasets
---	-------------------------------------	---------------------------------------	--

3.4.3 Supervised Classification of Safety Narratives

Supervised learning has been widely applied across numerous studies to classify various aspects of incident data. These techniques involve training ML models on labelled datasets to predict specific attributes such as the type of incident, its cause, associated risk or severity, or characteristics such as the injured body part. Unlike unsupervised approaches, supervised models learn relationships between textual representations and previously assigned labels, allowing established classification structures to be reproduced across larger collections of reports. This makes supervised learning particularly relevant where safety organisations already employ taxonomies or classification systems. However, its effectiveness depends on the availability, quality and distribution of labelled examples, as well as the representation of the underlying narrative text.

3.4.3.1 Machine Learning Models

Traditional machine-learning approaches have been extensively evaluated for safety-narrative classification, with Support Vector Machines (SVM), logistic regression, Naïve Bayes, decision-tree methods and ensemble models among the most commonly applied. Across the reviewed literature, SVM emerges particularly consistently as a strong classifier for high-dimensional textual data. Comparative evaluations have found SVM to outperform Naïve Bayes, KNN, decision trees, random forests and ensemble approaches in some classification tasks (Mujtaba et al., 2018), while similar strong performance has been reported relative to bagging, rule-based and regression-based classifiers (Le et al., 2018). SVM has also achieved strong classification performance at substantially different dataset scales, including an F1 score of 0.79 when applied to more than 600,000 narratives (Byon et al., 2023). Importantly for human-factors analysis, Morais et al. (2022) demonstrated its application to

identifying human-error factors across maritime and industrial accident reports, although performance varied between domains.

Evidence from safety-related classification more broadly reinforces this pattern. SVM achieved the highest AUC among evaluated approaches for incident-type and severity classification in a study by Evans et al. (2020), while other applications reported F1 scores of 0.826 for incident-type classification and 0.716 for severity classification (Wang et al., 2020)(Pence et al., 2020). Strong performance has similarly been reported for consequence classification, including an F1 score of 0.808 using a linear SVM (Mathew et al., 2021)(Syed et al., 2020). Earlier applications also demonstrated effective SVM classification with relatively small datasets (Ong et al., 2010) and for incident-cause identification (Liang et al., 2020), while Toyabe (2012) demonstrated its use alongside Bayesian approaches. Collectively, these studies indicate that SVM performance is not restricted to a single safety domain or dataset scale.

However, SVM is not consistently superior. Yoon et al. (2025) found strong SVM performance but reported slightly better results using Random Forest, while comparisons with deep-learning approaches have produced mixed findings. Baker et al. (2020b) found that SVM frequently outperformed CNN and Hierarchical Attention Network models for injury-precursor classification, except for body-part classification. In contrast, Cheng et al. (2020) reported an F1 score of 0.69 for an optimised GRU compared with 0.58 for SVM, although SVM remained the strongest conventional ML approach evaluated. SVM has also been used for multi-output accident classification (Shrestha et al., 2020) and outperformed labelled-LDA for incident-description classification (Guan et al., 2017), with the resulting classifications subsequently supporting similarity-based retrieval. These comparisons show that increased architectural complexity does not consistently produce better classification and that performance depends on the target category, text representation, and characteristics of the available training data.

Naïve Bayes, while simple, has proven useful in specific contexts. Ku and Leroy (2014) found it outperformed logistic regression in crime report classification and was even more accurate than human reviewers in some cases. Shome et al. (2024) found that although random forest was the best performer in detecting attacks on healthcare professionals in pandemic-related news articles, Naïve Bayes still delivered reliable results that were very close to the performance of random forest.

Tree-based ensemble methods have also performed strongly. Random Forest has outperformed alternative classifiers in several applications, particularly where class balancing or richer attribute sets were available (Grouin et al., 2013)(Ebrahimi et al., 2023)(Luo et al., 2023). Its ability to accommodate heterogeneous features and provide estimates of feature importance provides an additional advantage for safety analysis. Macedo et al. (2022) demonstrated that its performance could be improved through topic-derived features and model tuning even with relatively small and imbalanced data, while Yoon et al. (2025) and Shome et al. (2024) reported the strongest classification performance from Random Forest among their evaluated conventional classifiers. Decision trees provide a simpler alternative and may retain advantages for particular minority categories; Zhang et al. (2019a) for example, found a decision tree most effective for a very-low-support minority class.

Logistic regression remains a widely used baseline model in NLP classification tasks due to its simplicity, interpretability, and efficiency, especially when datasets are not excessively large or complex. Comparative studies generally indicate that more complex or regularised approaches can improve performance: Ridge regression outperformed conventional logistic regression in a study by Tabaie et al. (2024), while alternative ML and deep-learning models achieved stronger results in several other applications (Parikh et al., 2024)(Wong et al., 2020)(Ku and Leroy, 2014). Nevertheless, logistic regression remains useful for evaluating alternative text representations because its simplicity lets researchers distinguish improvements tied to the underlying features from those driven by classifier complexity (Tawfik and Spruit, 2020). Its continued effectiveness has also been demonstrated with free-text narratives (Groeneveld et al., 2025) and contextual representations derived from BERT-based

models (Nielsen et al., 2024)(Elizabeth et al., 2025). Its application within wind-energy safety was demonstrated by Sobhan Asian et al. (2017), who compared five conventional classifiers for incident-report analysis and found logistic regression most effective for identifying incidents involving fatalities or injuries. This reinforces the wider finding that increasingly sophisticated text representations can improve classification without necessarily requiring equally complex downstream classifiers.

Boosting methods provide another alternative, with XGBoost demonstrating competitive performance across several safety-classification tasks. Kumi et al. (2024) found XGBoost to outperform Naïve Bayes, SVM and KNN, while Parikh et al. (2024) reported that XGBoost combined with TF-IDF provided a computationally efficient alternative to more complex BERT-based models while retaining competitive performance. XGBoost can also operate on information derived from unsupervised text analysis: Li et al. (2024a) combined STM-derived features with XGBoost and used SHAP values to identify the topics contributing to severity predictions, thereby integrating predictive performance with model explainability. Jiao et al. (2023) similarly identified a favourable balance between predictive performance and computational efficiency. These findings further demonstrate that classifier complexity alone does not determine performance and that conventional ML can remain competitive when combined with appropriate textual or topic-derived representations.

Alternatively, Zhang et al. (2019a) suggests using combinations of multiple machine learning algorithms to best classify causes of construction incidents. They proposed using SVM, linear regression, k-nearest-neighbours, decision tree and naïve bayes. With these combined, they suggest using a sequential quadratic programming algorithm to optimise the weighting of each of the algorithms. The study found that their proposed model had the highest F1 score of 0.68 compared with the algorithms used individually. However, performance continued to vary between labels, leading the authors to recommend manual checking of categories with the weakest classification performance. This highlights an important limitation of aggregate model evaluation: improved overall

performance does not necessarily indicate reliable classification across all categories, particularly where individual labels are sparsely represented.

Reported differences between classifiers are also influenced by model optimisation. Grid search combined with cross-validation has commonly been used to select model configurations (Baker et al., 2020a, Macedo et al., 2022), while more computationally efficient optimisation strategies have increasingly been explored for complex models. Cheng et al. (2020), for example, applied Symbolic Organisms Search to optimise a GRU, which achieved an F1 score of 0.69 and outperformed the conventional classifiers evaluated in the study. Bayesian optimisation has similarly been reported to improve model selection while reducing the computational demands associated with exhaustive search, particularly for complex parameter spaces (Liu et al., 2024)(Lázaro et al., 2025). These findings demonstrate that comparative performance reflects not only model architecture but also the optimisation procedure applied, reinforcing the importance of consistent tuning and evaluation when comparing alternative classifiers.

Collectively, the literature demonstrates that conventional ML provides an effective and computationally efficient basis for safety-narrative classification, without identifying a universally superior algorithm. SVM demonstrates particularly consistent performance across textual classification tasks, while Random Forest and XGBoost provide advantages for particular datasets and feature representations. Logistic regression remains particularly relevant where interpretability, computational efficiency and probabilistic classification are important, and its effectiveness with both lexical and contextual representations demonstrates that increasingly sophisticated text representations do not necessarily require equally complex classifiers. Ensemble methods may further improve aggregate performance, although substantial variation can remain between individual categories. Overall, performance depends on the interaction between classifier, text representation, class distribution, model optimisation and target classification, supporting comparative evaluation rather than selection based on model complexity alone.

3.4.3.2 Challenges of Labelled Safety Data

A fundamental limitation of supervised safety classification is its dependence on sufficiently large and representative labelled datasets. Safety data are frequently highly imbalanced, with common incident or causal categories substantially outnumbering rare but potentially safety-critical events. This can bias classifiers towards majority categories and produce apparently strong aggregate performance while minority categories remain poorly identified. The problem is particularly relevant to taxonomy-based human-factors classification, where individual causal categories may have relatively few labelled examples but remain important to retain within the analytical framework.

Several strategies have therefore been applied to address class imbalance. Resampling provides a relatively straightforward approach, with Madeira et al. (2021) using majority-class downsampling in HFACS classification to improve balance between categories. Class weighting provides an alternative that retains the available observations while increasing the influence of underrepresented categories during model training (Baker et al., 2020a)(Zhang et al., 2021) to adjust model bias towards overrepresented categories, leading to fairer predictions across the class spectrum. Synthetic oversampling has similarly been used to increase minority-class representation, with SMOTE improving the stability of precision and recall for underrepresented categories in some safety-classification applications (Ebrahimi et al., 2023, Luo et al., 2023).

More advanced data-augmentation approaches attempt to increase minority representation while introducing greater variation into the generated training examples. Macedo et al. (2023) reported classification improvements of up to 15% after augmenting underrepresented accident-report patterns, while Liu et al. (2024) used GAN-generated examples to increase minority-class diversity. These approaches can increase the information available to the classifier but also introduce a trade-off between increasing sample size and ensuring that synthetic observations remain representative of genuine safety narratives. Consequently, augmentation cannot fully compensate for extremely limited or heterogeneous examples of particular categories.

Weak supervision addresses a related problem by reducing dependence on manually annotated training data. Rather than generating new observations, approximate or automatically generated labels can be assigned to larger collections of existing data. Parikh et al. (2024), for example, expanded a dataset containing 2,700 manually labelled cases to approximately 70,000 weakly labelled positive examples, while Zhou et al. (2025) similarly applied weak supervision to increase the availability of labelled safety data. Such approaches can substantially reduce expert-annotation requirements and improve representation of rare outcomes but introduce uncertainty regarding label reliability. The challenge therefore shifts from insufficient labelled data towards balancing label quantity against label quality.

This demonstrates that class imbalance cannot be treated solely as a model-selection problem. Resampling, weighting, augmentation and weak supervision can improve representation of minority categories, but each introduces different trade-offs concerning information loss, synthetic-data validity or label uncertainty. This is particularly important for fine-grained human-factors taxonomies, where numerous categories may be represented by relatively few examples and collapsing or removing minority categories can sacrifice analytically meaningful distinctions. Evaluation should therefore consider performance at the individual-category level rather than relying solely on aggregate classification metrics.

3.4.3.3 Deep Learning Classification

Deep-learning approaches have increasingly been applied to safety-narrative classification to capture more complex linguistic patterns than those represented through conventional feature-based ML. The reviewed literature principally applies convolutional and recurrent architectures, frequently in hybrid configurations, with more recent studies combining these approaches with pretrained transformer representations. Comparative evidence, however, indicates that increased architectural complexity does not consistently translate into improved classification performance.

CNN-based classifiers have demonstrated strong performance in several safety-related text-classification tasks. Borjali et al. (2021) reported substantially greater agreement using CNN than the best-performing conventional ML alternatives (97% compared with 69%), while Xu et al. (2022) found CNN achieved higher F1 scores than SVM in four of six evaluated categories. Baker et al. (2020b) similarly reported an F1 score of 95.49% for CNN on one classification task, although SVM remained superior for several other target categories. These findings show that CNN can outperform conventional ML, but its advantage depends on the specific classification target rather than being consistent across categories.

Comparisons with more context-sensitive architectures reveal further limitations. Kim et al. (2024) found CNN substantially less effective than BERT-based models, particularly with limited training data, while Jia et al. (2024) similarly reported stronger performance from BERT than TextCNN. Wang et al. (2020) also identified limitations when applying CNN to multi-label classification. Hybrid architectures have produced mixed results: Jiang and Wang (2024) reported improved performance, particularly for rare categories, by combining CNN with a contextual BiGRU encoder, whereas Kumi et al. (2025) found CNN alone outperformed both LSTM and a combined CNN-LSTM architecture. Conversely, Hou et al. (2025) reported that a hierarchical multi-task CNN outperformed BERT, GPT-2 and RNN baselines. Collectively, these results show that neither hybridisation nor transformer-based complexity guarantees superior performance and that architecture effectiveness depends on data characteristics and classification structure.

Recurrent architectures, particularly BiLSTM, have also been applied where classification depends on sequential and contextual relationships within narratives. Strong performance has been reported when BiLSTM is used for information extraction and categorisation, including F1 scores exceeding 95% in a combined NER and classification framework (Liu and Yang, 2022). However, a prominent trend is integrating BiLSTM with other neural or pretrained language models rather than using it as an isolated classifier. BiLSTM has improved BERT-based incident classification in some applications (Feng et al.,

2021), while CNN-BiLSTM combinations have outperformed conventional ML baselines (Batbaatar et al., 2019). BERT-BiLSTM-CRF architectures have similarly demonstrated improvements over evaluated baselines across several safety-related information-extraction tasks (Liu and Cheng, 2024)(Cao et al., 2024a)(Huang et al., 2025a), while Malik et al. (2023) combined BiLSTM with XLNet for accident-related classification. These applications show a broader shift from standalone recurrent architectures to hybrid models that combine sequential modelling with contextual language representations.

Graph Convolutional Networks are increasingly applied in safety-related NLP tasks to model complex relationships between entities in unstructured data. However, their performance is sensitive to dataset quality and size. For instance, Pan et al. (2022) reported relatively low F1 scores for a GCN-based accident-classification approach, partly attributing this to data inconsistency and insufficient volume. In contrast, Chen et al. (2022) combined GCN with a pretrained BERT model in a Relational GCN framework to classify causal factors in construction incidents. Their model achieved higher performance than traditional models like decision trees, GRU, and logistic regression, demonstrating the value of combining context-aware embeddings with graph-based reasoning.

Lin et al. (2025) went further by employing a spatiotemporal GCN (BSTGCN) for predicting vessel trajectories from video data, enabling robust performance even where AIS data was missing. This adaptation of GCN to spatiotemporal and visual data further expands its relevance beyond text classification, into real-time risk modelling and monitoring.

3.4.3.4 Transformer-Based Classification

Pretrained transformer models represent a further development in safety-narrative analysis, enabling contextual language representations to be adapted to specific classification and information-extraction tasks. BERT and its derivatives have demonstrated strong performance across safety datasets, although the reviewed evidence indicates that performance depends substantially on the relevance of the pretraining corpus and subsequent domain adaptation. Oliaee et al. (2023), for example, fine-tuned BERT on approximately 740,000 reports and achieved approximately 84% accuracy for severity

classification, although ambiguous intermediate-severity labels remained difficult to distinguish. Nguyen et al. (2022) similarly reported that domain-specific BlueBERT outperformed SVM, while Nielsen et al. (2024) improved classification through additional domain-specific pretraining of RoBERTa, indicating that pretrained language representations can reduce reliance on manually engineered textual features, but adaptation to domain-specific terminology and classification structures remains important.

Comparisons between general and specialised transformer models provide further evidence of the importance of pretraining and domain alignment. Kim et al. (2024) found BioBERT and SciBERT more effective than baseline BERT and CNN for causal-sentence classification, while Kizaki et al. (2024) demonstrated substantial performance differences between Tohoku-BERT, UTH-BERT and ELECTRA. Transformer-based models have also outperformed conventional CNN and recurrent architectures in causal classification (Jia et al., 2024), although the wider deep-learning evidence demonstrates that this advantage is not universal across all datasets and classification targets. Collectively, these studies suggest that the suitability of the underlying pretrained representation may be as important as model architecture itself.

Transformer representations have also been incorporated into hybrid architectures. BERT-BiLSTM and BERT-BiLSTM-CRF approaches have been used to classify incidents and extract contributory factors and event information, frequently outperforming the evaluated baseline models (Feng et al., 2021)(Wang and Zhao, 2022)(Xiong et al., 2024b)(Liu and Yang, 2023)(Huang et al., 2025a). Malik et al. (2023) similarly combined XLNet representations with BiLSTM, achieving 94.2% classification accuracy. These studies demonstrate that contextual transformer representations can be combined with sequential architectures where additional sequence-level modelling is beneficial, although the resulting increase in architectural complexity must be justified against simpler fine-tuned alternatives.

Beyond document classification, transformer models have supported information extraction and other narrative-processing tasks. Multi-task BERT architectures have demonstrated strong performance for

NER and simultaneous extraction or classification of safety information (Chen et al., 2024)(Wong et al., 2024). Transformer models have also been applied to safety-narrative summarisation, with Cosma et al. (2024) demonstrating differences in performance between BART, T5 and DistilBART and identifying variation in summarisation quality between groups. These applications demonstrate the broader potential of pretrained transformers to convert unstructured narratives into structured or condensed information, while also highlighting that performance can vary across tasks and subsets of the underlying data.

3.4.3.5 Large Language Models and Emerging Approaches

More recent research has investigated generative large language models (LLMs), which differ from conventional fine-tuned transformer classifiers by enabling safety information to be extracted or classified through natural-language instructions and prompting. Earlier generative-model applications demonstrated the value of domain adaptation, with Jose et al. (2024) reporting improved representations from domain-adapted GPT-2 and approximately 30% improvement in downstream predictive performance when textual information was incorporated. More recent LLM applications increasingly use prompting rather than conventional task-specific classifier training, potentially reducing dependence on large manually labelled datasets.

Prompt-based approaches have demonstrated the ability of general-purpose LLMs to extract structured information directly from incident narratives. Sufi (2024) used prompt engineering to extract predefined incident characteristics, while Nakamura et al. (2024) compared zero- and one-shot GPT-4 classification against human-coded ground truth. One-shot prompting approached human-level performance and exceeded zero-shot performance for more complex incident categories, indicating that even limited task-specific examples can substantially influence LLM classification. This flexibility is potentially valuable in safety domains where labelled examples are limited, although performance remains sensitive to prompt design and task complexity.

Particularly relevant to human-factors analysis, Liu et al. (2025b) integrated HFACS with chain-of-thought prompting in the HFACS-CoT framework for classification of incident narratives using GPT-4o. Structured sequential prompting improved precision, recall, F1 score and logical consistency relative to basic zero-shot prompting, with the reported performance also slightly exceeding the human-expert comparison. As illustrated in Figure 14, performance varied across both prompting strategies and HFACS subtasks, demonstrating that LLM classification is sensitive not only to prompt structure but also to the human factor category being classified. This provides evidence that established safety taxonomies and domain knowledge can be incorporated directly into LLM prompting rather than relying solely on unconstrained language-model reasoning. More broadly, Qi et al. (2025) demonstrated similar sensitivity to interaction and prompt design when applying ChatGPT to STPA, with iterative interaction producing more comprehensive and expert-aligned outputs than one-off prompting.

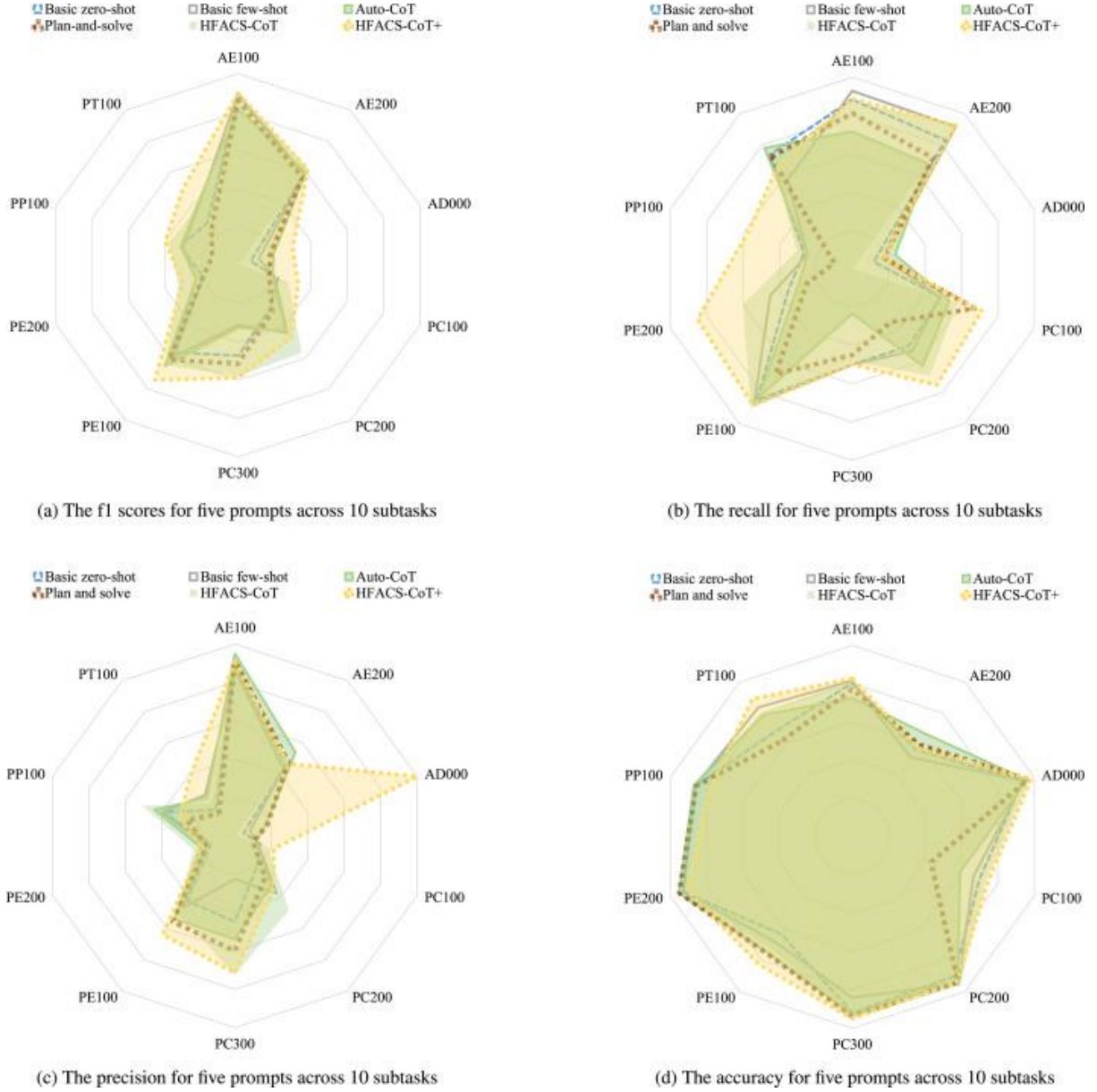


Figure 14. Performance comparison of different prompting strategies for LLM-based HFACS classification across multiple subtasks (Liu et al., 2025b).

Similarly, Qi et al. (2025) investigated the application of ChatGPT to STPA (System-Theoretic Process Analysis) in two real-world case studies. They compared different collaboration modes (one-off, recurring) and prompt styles (STPA-specific vs general). Their results showed that recurring duplex prompting produced the most comprehensive and expert-aligned outputs, and that prompt specificity had mixed effects, domain-agnostic prompts surprisingly uncovered more unsafe control actions in some cases.

However, the flexibility of prompt-based LLMs introduces concerns regarding consistency and reliability. Lázaro et al. (2025) found inconsistent human-factor assignments from GPT-4 Turbo, demonstrating that general-purpose language capabilities do not necessarily produce stable taxonomy classification without sufficient task structure. Sabetta et al. (2025) similarly reported high overall performance for structured extraction but identified errors where narrative phrasing was ambiguous. Prompt formulation can itself influence the nature of generated safety conclusions: Ren et al. (2025) found that targeted prompts applied to Qwen2-7B generated more concise and actionable causal information, whereas general prompts produced greater irrelevant verbosity. These findings indicate that LLM performance cannot be considered independently of prompt design, output constraints and the ambiguity of the underlying narratives.

Open-source LLMs provide an alternative where external API use or data governance presents concerns. Siddeshwar et al. (2024) demonstrated the use of LLaMA-2-7B for structured extraction from aviation narratives without reliance on an external API, although the relatively small evaluation of 50 reports limits the extent to which performance can be generalised. More broadly, the emerging LLM literature demonstrates considerable potential for flexible safety classification and information extraction but remains less mature than conventional supervised classification. Reported performance is sensitive to prompting strategy, domain knowledge and narrative ambiguity, while consistency, reproducibility and validation against expert-labelled data remain important considerations for safety-critical applications.

3.4.4 Analysis and Application of Structured Safety Information

3.4.4.1 Knowledge Graphs

Knowledge graphs and related network approaches extend narrative analysis by representing extracted safety entities and their relationships in structured form. Rather than treating causal factors as independent classifications, graph representations can connect hazards, causes, consequences, accident characteristics and other safety information, enabling recurring relationships and causal

pathways to be explored. Construction approaches range from manually or rule-based encoded relationships to increasingly automated pipelines combining NLP, deep learning and domain ontologies.

Early and comparatively interpretable approaches have used text-mining or manually coded accident factors to construct relational representations. Zhu et al. (2023) combined keyword extraction with graph construction to identify interconnected accident factors, while Li et al. (2024b) structured manually coded 24Model factors to identify shared root causes. Cao et al. (2024b) similarly combined association mining with causal-network construction to quantify coupling strengths and identify dominant accident pathways. These approaches demonstrate the analytical value of relational representations, but their dependence on predefined or manually extracted factors limits automation.

More recent approaches increasingly automate graph construction by extracting entities and relationships directly from narratives. Xue et al. (2025) combined ontology design with BERT-BiLSTM-based relation extraction to generate semantic triplets, while Zhou et al. (2025) used BiLSTM-CRF-based entity extraction and clustering before organising the identified factors within a domain-specific graph. Jia et al. (2024) similarly combined neural text classification with graph-based representation of extracted causal factors. These approaches demonstrate an increasing convergence between NLP and knowledge-graph construction, whereby automated text analysis provides the structured entities and relationships required for subsequent relational analysis.

This development is also evident within transport safety. Knowledge graphs constructed from maritime and railway narratives have represented accidents through entity-relation triplets and causal or risk-propagation chains (Gan et al., 2023)(Liu et al., 2025a)(Huang et al., 2025b). Particularly relevant to maritime safety, Liu and Cheng (2024) developed a Maritime Accident Knowledge Graph containing more than 16,000 entities and 20,000 relationships, linking accident cases, vessels and regulatory information. More recent maritime approaches have combined transformer-based extraction with graph construction to map relationships between human, technical and causal factors (Zhang et al.,

2025)(Chen et al., 2025a). Similar automated graph-construction approaches have also been applied to structured safety assessment and validation (Li and Zhao, 2025).

This shows that knowledge graphs provide a mechanism to move beyond independent classification toward analysing relationships between safety factors. Their principal advantage is preserving relational structure, enabling extracted information to support causal exploration, retrieval, and identification of recurring pathways. However, the reliability of the resulting graph depends on the quality of the underlying entity and relationship extraction, and graph construction often requires predefined ontologies, taxonomies, or domain rules. Knowledge graphs therefore complement rather than replace narrative classification: NLP provides the structured safety information from which relational representations can subsequently be constructed.

3.4.4.2 Bayesian Networks

Bayesian Networks (BNs) provide a complementary relational approach by representing probabilistic dependencies between structured safety variables. Whereas knowledge graphs primarily represent entities and their relationships, BNs additionally quantify conditional dependencies and support predictive, diagnostic and sensitivity analysis under uncertainty. Their application therefore generally occurs after experts have identified relevant accident factors through coding, taxonomies, text mining, or other information-extraction methods.

The reviewed literature demonstrates extensive integration of BNs with methods for converting accident information into structured causal variables. Text mining and semantic clustering have been used to identify factors before probabilistic modelling (Li et al., 2022a)(Ni and Zhang, 2023), while established accident-causation frameworks have provided the underlying causal structure in other applications. These include integration with the 24Model (Yuxin et al., 2024), HFACS (Jiang et al., 2025a), bow-tie analysis (Valcamonico et al., 2024), interpretive structural modelling (Kamil et al., 2023a), fuzzy causal modelling (Kamil et al., 2024)(Jin et al., 2025b) and association-rule mining (Yuan et al., 2025). Collectively, these approaches show that BNs often serve as a subsequent analytical layer

through which previously identified safety factors can be connected and their probabilistic relationships examined.

Recent maritime applications demonstrate similar methodological diversity. BNs have been used to examine occupational-accident severity, oil-spill severity and broader maritime accident-risk relationships (Jiang et al., 2025a)(Wang et al., 2025)(Bin and Shiqi, 2025). Other studies have integrated BNs with complementary causation approaches: Ivana et al. (2025) combined Fault Tree Analysis with probabilistic modelling of human and operational factors, while Juliet Wangui et al. (2025) integrated Grounded Theory, systems thinking, Fault Tree Analysis and fuzzy reasoning to construct a BN under conditions of limited accident data. Bairami-Khankandi et al. (2025) similarly combined CAST, association-rule mining and Bayesian analysis to investigate systemic maritime accident patterns. These studies illustrate the value of BNs for quantifying relationships initially identified through qualitative, systemic or data-mining approaches.

Increasingly, data-driven BN architectures extend this principle further. Jingpeng et al. (2025) combined object-oriented and Tree-Augmented Bayesian Networks with clustering, causal-path searching and Shapley-value analysis across 698 maritime accident cases, seeking to retain interpretable causal structure while improving predictive analysis. Similar developments across other safety domains have combined BNs with association mining, fuzzy inference and automated text processing to identify influential pathways and support sensitivity analysis. These developments demonstrate increasing convergence between probabilistic safety modelling and data-driven analytical techniques.

A further limitation of data-driven BNs is their dependence on sufficient and representative data to estimate probabilistic relationships reliably. As network complexity and the number of variables increase, larger datasets are generally required to populate conditional relationships adequately, which can restrict their application to relatively small accident datasets. Highly imbalanced data present an additional challenge, as rare accident outcomes or causal factors may be represented by

too few observations to estimate stable conditional probabilities. Consequently, although BNs provide an interpretable mechanism for examining dependencies between safety factors, their reliability depends on both the quantity and distribution of the available data, particularly where rare but safety-critical events are of interest.

3.4.4.3 Association Rule Mining (ARM)

Association Rule Mining identifies recurring relationships between variables within structured datasets. Rules are evaluated using support, confidence and lift, representing respectively the frequency of the combination, the conditional occurrence of the consequent, and the strength of association relative to statistical independence. Its principal value in safety analysis is therefore its ability to extend frequency-based analysis by identifying combinations of contributory factors or accident characteristics that repeatedly occur together.

ARM has been applied across safety domains both independently and within broader analytical frameworks. Cao et al. (2024b) combined association mining with causal-network analysis to identify influential accident pathways, while Yuan et al. (2025) integrated ARM with Bayesian Networks to examine relationships between accident factors. Maritime applications include a study by Jin et al. (2025a) who used ARM to identify associations between accident-related factors, and a study by Bairami-Khankandi et al. (2025), who combined ARM with CAST and Bayesian analysis to investigate systemic accident patterns. These applications demonstrate ARM's value as an interpretable exploratory method and as an intermediate step between structured factor identification and more complex relational or probabilistic analysis.

However, association does not establish causation, and strong rules may be influenced by the prevalence of individual factors within the dataset. Results are also sensitive to the thresholds selected for support, confidence and lift: restrictive thresholds can exclude uncommon but safety-relevant relationships, whereas permissive thresholds can generate numerous weak associations. ARM is

therefore most appropriately used to identify potentially meaningful relationships for subsequent interpretation rather than as evidence that one safety factor causes another.

3.4.4.4 Decision-Support and Human-in-the-loop Applications

Several studies have translated their NLP models into user-friendly tools designed for domain experts without technical backgrounds in computing or programming. These tools aim to make advanced NLP functionalities accessible and actionable in real-world settings.

For example, Mathew et al. (2021) developed an interactive tool to support incident investigation and analysis. Users input an incident description, and the NLP model processes the text to generate dropdown suggestions for incident type and location, ranked by probability. For contributing factors, where multiple classifications may apply, the system allows users to select more than one category. As shown in Figure 15 the interface presents classification recommendations alongside associated probability scores, enabling investigators to review and validate the model outputs. This human-in-the-loop design supports consistency in incident reporting while maintaining expert oversight during the investigation process.

(a)

SOLIS
Search
Report An Incident
Dashboard
Incidents
Actions
Statistics

Report an Incident

Incidents:
An incident is any unwanted or unexpected change from a normal system behaviour that causes, or has the potential to cause, an adverse effect to persons or equipment.
This includes reportable circumstances and near-misses.

Instructions:
All of the fields on this initial report are **mandatory** for online submission.
Please note the ID number provided to you upon successful submission of an online report. You may use this number to follow-up on the incident during the investigation phase.
Please note that an email will be sent to notify the associated physician for near-misses and actual incidents.
For complete instructions on reporting incidents, please refer to the [tutorial on DigDocx](#).

Type of Report:

Event Type:

Functional Work Area:

Date Incident was Detected:

Time Period Detected:

Incident Description:

Incident Descriptor:

Reported By:

Coordinator Comments:

Staff Support Required?:

Investigator:

Username:

Password:

Process step where incident occurred

	Probability
Treatment delivery	85%
Treatment planning	10%
Post-treatment completion	3%
Imaging for radiotherapy planning	2%

Figure 15. User interface of the incident investigation support system developed by Mathew et al. (2021).

A practical healthcare application is described by Hinrichsen et al. (2007), who designed a tool for early detection of vaccine-related adverse events. The system monitors patients who seek care within 14 days of vaccination and prompts healthcare providers to assess whether the case could be vaccine related. If so, the provider is guided through a symptom checklist, and the tool suggests reporting the case to VAERS, improving consistency in adverse event reporting.

In the process safety domain, Wang and Gu (2022) created a three-step automated HAZOP analysis tool that supports risk prediction, consequence mapping, and structured summarisation of incidents, streamlining the traditional HAZOP process. As illustrated in Figure 16, the system provides an interface through which users can review identified deviations, potential causes, consequences, and associated risk levels. Such interfaces demonstrate how NLP outputs can be translated into practical decision-support tools, enabling experts to interact with model results without requiring specialist knowledge in machine learning or natural language processing. Similarly, Kurian et al. (2020) used machine learning outputs to power a user-interactive risk matrix tool. The system first suggests relevant hazard

labels, allows user confirmation, and then produces a frequency-consequence matrix along with mitigation recommendations.

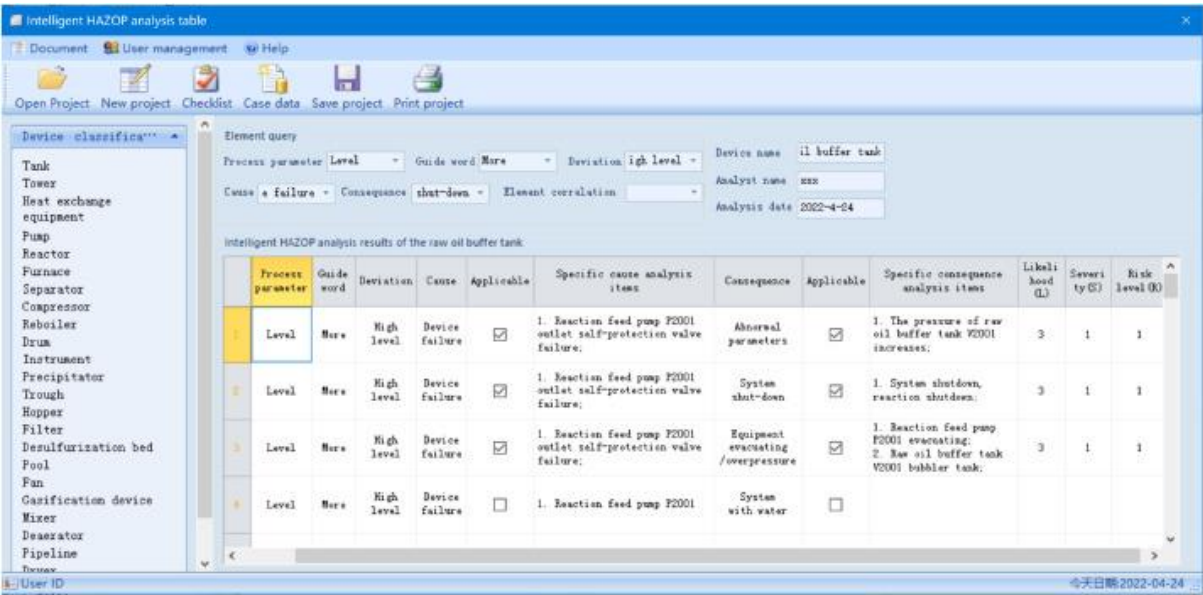


Figure 16. User interface of the intelligent HAZOP analysis system developed by Wang and Gu (2022) for automated hazard identification and consequence analysis.

Perboli et al. (2021) developed a tool that links incident report sentences to SHEL categories, presenting correlation scores to assist human analysts in reviewing and refining classification outputs. This hybrid model preserves automation while ensuring expert oversight.

Lastly, Gangadhari et al. (2024) integrated unsupervised clustering results into a hybrid decision-support framework. Term importance (TF-IDF scores) served as objective weights, which were then balanced against subjective inputs from experts using the Best-Worst Method (BWM), creating a consensus-driven prioritisation of risks.

3.5 Summary and Identified Research Gaps

This review examined two complementary bodies of literature required for the systematic analysis of maritime and offshore safety information: established approaches to accident and human-factors analysis, and computational methods for extracting and analysing safety information from unstructured narratives. The first determines how accident causation and human factors can be

represented meaningfully, while the second determines how such analysis can be automated and scaled. Together, the literature shows that neither established accident-analysis methods nor AI-based approaches alone resolve the challenge of extracting consistent, actionable human-factors information from large collections of heterogeneous safety narratives.

3.5.1 Accident and human-factors analysis

The reviewed accident-analysis methods range from structured human-factors taxonomies to systems-based approaches that examine interactions across human, technical, organisational, and regulatory levels. Several key findings emerge from this literature:

1. Human-factors taxonomies provide structured and repeatable classification but require domain adaptation.

HFACS has been extensively adapted for maritime applications, with repeated additions of operational, environmental and external factors demonstrating the importance of domain-specific representation. Maritime-specific frameworks such as SHARE provide alternatives more closely aligned with maritime operations, while frameworks such as SHIELD adopt a more domain-independent structure intended to support human-factors analysis across different operational contexts.

2. The quality of human-factors analysis remains dependent on the underlying accident evidence.

Detailed taxonomies cannot identify organisational, supervisory, or individual factors when the source report lacks sufficient causal information. Increasing classification detail therefore does not necessarily increase analytical reliability.

3. Organisation-specific taxonomies restrict comparison between reporting systems.

Although classifications used by investigation and regulatory organisations support operational analysis, differences in terminology, structure and analytical depth make aggregation and comparison across datasets difficult.

4. Systems-based methods provide greater causal depth but are difficult to scale.

Approaches including FRAM, AcciMap and STAMP/CAST provide richer representations of interactions across accident systems but generally require substantial expert interpretation and case-specific reconstruction, limiting consistent application across large report collections.

5. Taxonomies and systems approaches therefore provide complementary analytical capabilities.

Taxonomies offer greater standardisation and computational tractability, whereas systems approaches provide greater explanatory depth. For large-scale analysis, structured taxonomies provide a practical basis for automated classification, while systems thinking highlights the importance of examining relationships between identified factors.

6. Accident analysis is increasingly being extended through quantitative and relational methods.

Integration with Bayesian Networks, association analysis and related approaches demonstrates the value of progressing beyond individual factor frequencies towards examining dependencies and recurring combinations. However, these analyses remain dependent on the reliability of the factors identified during the preceding accident-analysis stage.

Collectively, the literature demonstrates a trade-off between analytical depth and scalability. Existing accident-analysis approaches provide theoretically and operationally meaningful representations of accident causation, but detailed application remains heavily dependent on expert interpretation. This creates an opportunity to combine the structured representation provided by human-factors taxonomies with computational methods capable of applying such analysis more consistently across large collections of accident information.

3.5.2 AI-based Safety Analysis

Figure 17 summarises the principal methodological categories identified across the reviewed literature. As illustrated, NLP applications in accident investigation span a broad spectrum of approaches, ranging from traditional rule-based methods and machine learning algorithms to

transformer-based language models and knowledge representation techniques. The review also highlights a general progression towards methods with greater semantic capability and contextual understanding, albeit often at the expense of increased computational complexity and reduced interpretability.

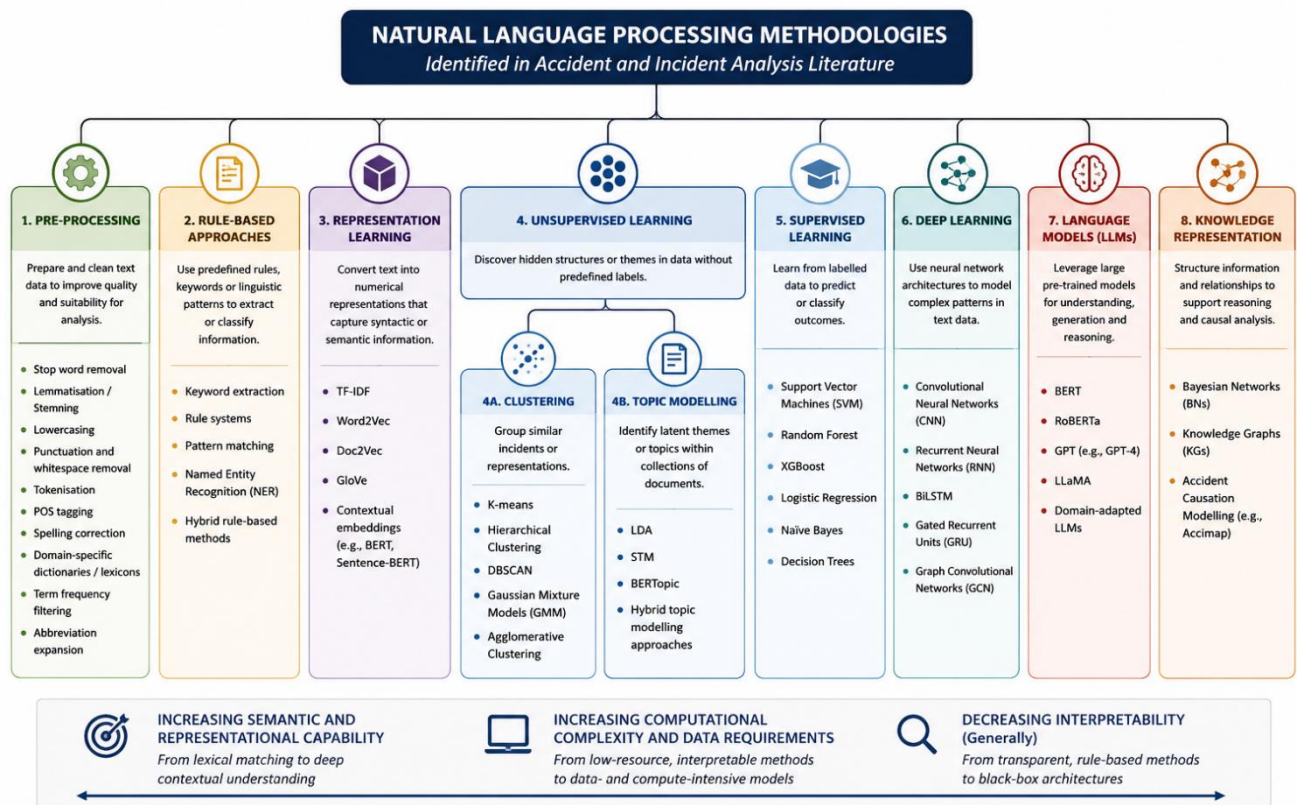


Figure 17. Overview of reviewed methodological approaches.

Some key findings from the literature are as follows:

1. Data structure and text consistency profoundly affect model performance.

Studies consistently show that structured or semi-structured narratives produce far stronger results than free-text fields. This reinforces the need for domain-specific taxonomies, consistent annotation, and probability-calibrated classification methods.

2. Unsupervised learning remains widely used, but interpretability weaknesses persist.

Methods such as LDA or k-means often produce overlapping or ambiguous themes. BERTopic offers notable improvements through transformer embeddings, yet all unsupervised methods still require extensive expert interpretation.

3. BERT-based models and hybrid architectures now outperform traditional ML approaches.

Domain-adapted transformers combined with classifiers (e.g. BERT + SVM, BERT + BiLSTM-CRF) offer the strongest results for both classification and NER. However, many studies lack probability calibration, hierarchical structure, or expert-in-the-loop validation, limiting their practical deployment.

4. LLMs (GPT, LLaMA, etc) show promise but lack robustness without structured prompting or domain constraints.

Recent work shows that unstructured prompting yields inconsistent outputs, and domain-specific constraints significantly improve reliability. However, the work also shows that the LLMs often override direct prompts due to the largeness of the model reverting to its original form.

5. Few studies move beyond classification to model causal pathways or systemic interactions.

While some integrate association rule mining or Bayesian networks, these efforts are not aligned with established human-factor frameworks (e.g. HFACS, AcciMap) or do not connect narrative-level findings to systemic risk insights.

6. Real-world tool adaptation is still limited.

Although some studies produce decision-support tools, they rarely use hierarchical modelling, probability calibration, or causal pathways. Most tools only visualise topics or provide simple classification suggestions – revealing the need for a more holistic, industry-ready framework.

7. Critical domain gaps remain – especially in offshore wind, maritime human factors, and cross-boundary taxonomies.

Research in aviation and healthcare is more mature, whereas offshore wind has virtually no taxonomy-aligned NLP research. Most maritime studies do not integrate structured human-factors taxonomies (e.g. HFACS, SHIELD) or calibrate predictions in a way suitable for safety-critical deployment. This stands in contrast to the needs of high-risk, rapidly expanding sectors.

Collectively, these findings highlight several challenges associated with the automatic population of structured human factors taxonomies such as SHIELD. Although many studies demonstrate high predictive performance, most focus on flat classification tasks and do not support the hierarchical and multi-label structures required by human factors frameworks. Incident narratives frequently contain ambiguous language, implicit causal relationships, and domain-specific terminology, limiting the reliability of automated extraction. Furthermore, existing approaches often lack explainability, probability calibration, and expert-in-the-loop validation, all of which are essential for deployment within safety-critical environments. While transformer-based models and LLMs provide improved semantic understanding, few studies successfully integrate these capabilities with established human factors frameworks or model the complex interactions that exist between organisational, supervisory, environmental, and individual factors. Consequently, there remains a clear need for an integrated methodology capable of combining semantic understanding, hierarchical classification, causal reasoning, and human oversight to support the reliable and explainable population of SHIELD within maritime and offshore safety domains.

3.5.3 Identified Research Gaps and Novelty

Considered together, the accident-analysis and AI literature identifies a fundamental limitation in maritime and offshore safety analysis: the lack of scalable and consistent methods for extracting human-factor insights from large volumes of unstructured safety data. Established accident-analysis frameworks provide meaningful structures for interpreting human and organisational contributions to accidents but remain heavily dependent on manual investigation and expert judgement. Conversely, AI-based approaches can process safety narratives at scale but rarely integrate with comprehensive

human-factors frameworks or adapt to the substantially different characteristics of short incident narratives and detailed accident investigation reports.

The methodological gap therefore does not arise from the absence of individual AI techniques, but from how they are integrated and adapted to specific safety-analysis requirements. The reviewed literature demonstrates that detailed accident reports and short safety narratives present fundamentally different analytical problems. Detailed reports contain greater contextual and causal information and can support supervised classification against an established human-factors taxonomy. Short incident and near-miss narratives often lack predefined labels and sufficient causal evidence for such classification, making them better suited to unsupervised discovery of recurring safety themes.

The first novel component of the proposed framework addresses short, unlabelled safety narratives through an integrated pipeline combining BERTopic-based topic discovery, LLM-supported semantic interpretation, temporal analysis and Association Rule Mining. While these techniques have been applied to safety data individually, their integration enables analysis to move beyond identifying isolated topics to examining how safety themes change over time and which themes repeatedly co-occur. This provides a mechanism for extracting broader patterns from high-volume incident and near-miss data without imposing detailed causal classifications unsupported by the narratives.

The second component addresses detailed accident investigation reports through an AI-enabled, taxonomy-based classification pathway. SHIELD-NLP adapts the SHIELD human-factors taxonomy for computational implementation, enabling accident narratives to be classified within a structured, hierarchical and multi-label maritime human-factors framework. The pathway incorporates SeaSafeBERT as a maritime accident-domain language representation alongside lexical representations and supervised ML techniques, with class-imbalance methods used to address sparsely represented human-factor categories. Importantly, the literature shows that increasingly complex models do not consistently outperform conventional approaches; the framework therefore evaluates alternative

representations and classifiers rather than assuming the most sophisticated architecture will deliver the strongest performance.

The overarching novelty lies in integrating these two complementary analytical pathways within a unified framework for heterogeneous maritime and offshore safety data. Rather than applying a single NLP method irrespective of data characteristics, the framework selects the analytical approach according to the information available within the source narratives: taxonomy-aligned supervised classification where sufficiently detailed information is available to support human-factors classification, and unsupervised semantic discovery where such classification is unsupported by the available evidence. Subsequent temporal and association analysis further enables structured outputs to be examined for recurring patterns and relationships rather than treating classification or topic discovery as the analytical endpoint.

The contribution therefore lies in the targeted integration of established human-factors analysis with AI methods selected according to the characteristics and analytical potential of the available safety data. The following chapter formalises this approach within the proposed methodological framework and defines the two analytical pathways used in this research.

4 Methodological Framework

4.1 Chapter Overview

This chapter presents a multi-level analytical framework designed to systematically extract, interpret, and model human and causal factors in maritime accidents under varying data conditions. Built upon insights from the methodological review, this framework accommodates both minimal-data and rich-data environments, allowing for scalable, intelligent analysis of maritime safety data using a combination of NLP, clustering, temporal analysis, and ML techniques.

The methodology adopted in this thesis evolved through a series of preliminary investigations that explored conventional Natural Language Processing techniques for safety incident classification. Summaries of these proof-of-concept studies are provided in Appendices A and B, while the refined methodology developed from these investigations is presented throughout the remainder of this thesis.

4.2 Framework Overview

The framework was developed to recognise the heterogeneous nature of maritime safety data. Maritime accident investigations and safety reporting systems vary considerably in terms of data completeness, structure, and level of detail. While some datasets contain only limited narrative information, others include extensive structured variables, causal classifications, and contextual metadata. Consequently, a single analytical approach may not be appropriate across all data environments. The proposed framework (shown in Figure 18), therefore, adopts a flexible multi-level design, enabling analytical techniques to be selected according to data availability while ensuring that meaningful insights into human factors and systematic safety issues can still be extracted.

The proposed framework is composed of two primary branches:

1. Low-information approach for when limited or unstructured textual data is available
2. High-information approach for when extensive, structured, and tagged data is accessible.

Each branch applied methods most suited to the data availability and granularity, while maintaining a consistent objective: to extract meaningful insights about human factors and systemic risks.

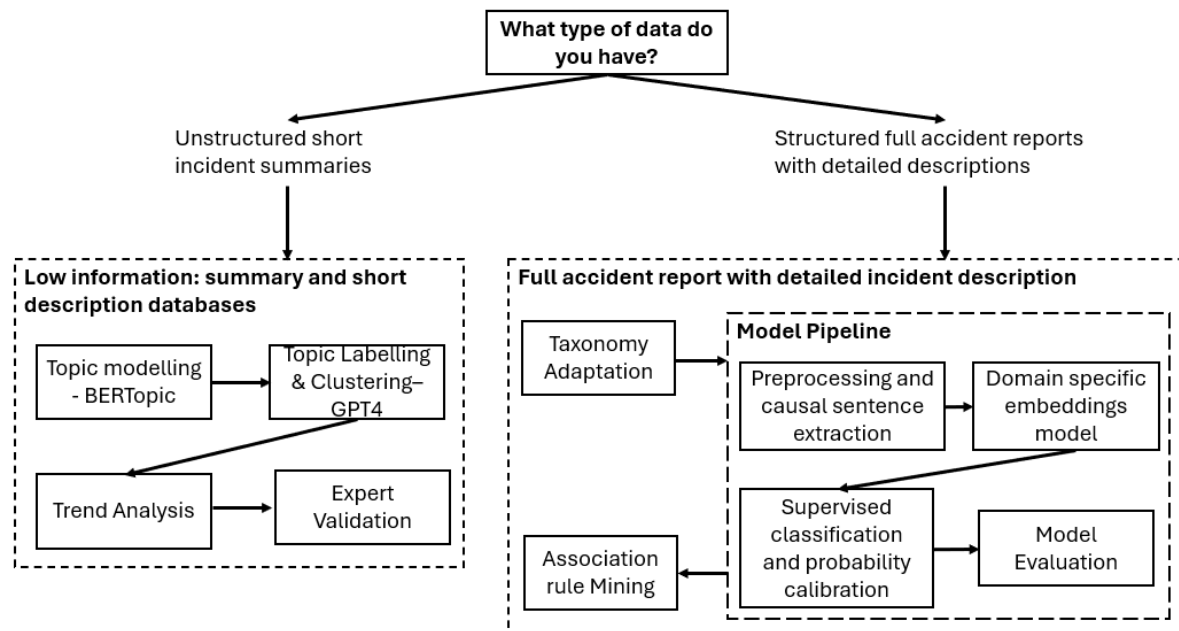


Figure 18. Visual diagram of framework.

The distinction between the two pathways is primarily determined by the level of detail contained within the available safety data. Where only short narrative descriptions are available, such as near-miss reports or brief incident summaries, there is insufficient contextual information to reliably identify and classify contributory human and organisational factors using a structured taxonomy. In these cases, unsupervised analytical techniques are more appropriate for identifying recurring themes, trends, and relationships within the data. Conversely, detailed accident investigation reports provide comprehensive descriptions of the sequence of events, contributing factors, and organisational context, enabling the application of a human factors taxonomy to systematically classify causal and contributory factors. The high-information pathway therefore incorporates the SHIELD taxonomy to provide consistent, interpretable, and scalable human factors classification, whereas the low-information pathway focuses on extracting operational insights without attempting detailed causal classification.

Two distinct pathways are necessary because differences in data richness affect not only the suitability of particular analytical techniques, but also the level of inference the available evidence can reasonably support. Applying a detailed human factors taxonomy to sparse or highly abbreviated reports could require assumptions about causal or contextual factors that are not explicitly documented, potentially introducing uncertainty or misclassification. Conversely, restricting information-rich investigation data to broad exploratory techniques would fail to utilise the detailed causal and contextual information available. The two-pathway structure therefore matches analytical depth to the dataset's evidential depth, allowing the framework to remain applicable across heterogeneous maritime safety data sources without imposing a uniform level of analysis where the underlying data cannot support it.

4.3 Low Information Approach (Unstructured Short Incident Summaries)

The low-information approach was developed for safety datasets comprising short, largely unstructured descriptions with limited supporting metadata. Such data are commonly encountered across maritime and offshore industries, including near-miss reports, incident summaries, hazard observations, and operational safety reports. Although these narratives are often brief and lack detailed causal information, they contain valuable contextual information regarding operational conditions, human actions, equipment issues, and emerging safety concerns. Consequently, analytical techniques capable of extracting meaningful patterns from sparse textual data are required to support proactive safety management and organisational learning.

This research applies the low-information pathway in two case studies: ferry near-miss reports and offshore wind occupational incident reports. Both applications use the same core process of semantic topic discovery followed by analysis of the resulting structured information, but the subsequent analyses adapt to the information available in each dataset. In the ferry case study, separate incident and prevention narratives enable examination of relationships between identified incident and prevention themes. In the offshore wind case study, incident narratives are accompanied by structured incident characteristics and outcome classifications, enabling discovered themes to be examined

alongside wider incident information. The pathway therefore provides a common analytical structure while allowing downstream analysis to reflect each dataset's characteristics.

4.3.1 Topic Modelling (BERTopic)

The central analytical approach adopted for short, unstructured incident summaries, in this research is BERTopic. The method was selected for its capability to extract semantically rich and contextually meaningful topics from unstructured textual data. Unlike traditional probabilistic topic modelling techniques, BERTopic leverages transformer-based embeddings that capture nuanced semantic relationships within text, enabling it to identify coherent themes even when data is noisy, brief, or domain-specific.

The textual inputs differ between the two applications. For the ferry near-miss case study, BERTopic is applied separately to incident descriptions and corresponding prevention narratives, allowing themes describing the reported events and proposed preventive measures to be identified independently. For the offshore wind case study, BERTopic is applied to the available occupational incident narratives to identify recurring operational and safety themes. The underlying BERTopic process remains consistent across both applications.

The BERTopic pipeline comprises four core components, the embeddings model, dimensionality reduction, clustering and topic representation. The embedding model transforms the incident descriptions or summaries into high-dimensional vector representations using Sentence-BERT (SBERT). SBERT builds upon the BERT architecture to produce sentence-level embeddings that retain contextual meaning and semantic similarity between texts. In this research, the pre-trained *all-MiniLM-L6-v2* Sentence-BERT model was utilised. This model generates 384-dimensional sentence embeddings and was selected because it provides an effective balance between semantic performance and computational efficiency for large collections of short incident narratives. This step eliminates the need for heavy text preprocessing (e.g. stemming, lemmatisation, or stop-word removal), as contextual encoding is inherently captured by the transformer model.

Since the resulting embeddings are typically high-dimensional, Uniform Manifold Approximation and Projection (UMAP) is applied to improve computational efficiency and cluster separability. UMAP constructs a high-dimensional graph representing local neighbourhood relationships and optimises a low-dimensional representation by minimising the cross-entropy between the two graphs. This allows it to preserve the local and global structure of data while reducing dimensionality, allowing for clearer topic boundaries and improved visualisation. UMAP was implemented using the default BERTopic parameters (*n_neighbors*=15, *n_components*=5, *min_dist*=0.0, *metric*='cosine'). Within this configuration, *n_neighbors*=15 determines the local neighbourhood used to construct the manifolds, balancing local semantic structure against broader relationships between documents, while *n_components*=5 substantially reduces the embedding dimensionality while retaining sufficient information for subsequent HDBSCAN clustering. A *min_dist* of 0.0 allows neighbouring documents to be represented compactly in the reduced space, facilitating the formation of dense clusters, while cosine distance is appropriate for comparing the orientation of high-dimensional semantic embedding vectors. These parameters were retained as standardised BERTopic configuration rather than being tuned towards a predetermined topic structure. The use of default BERTopic settings as an initial configuration is consistent with previous applied BERTopic research (Tang et al., 2024). As the purpose of the analysis was exploratory topic discovery rather than optimisation of the clustering solution, no dataset-specific hyperparameter tuning was undertaken.

The reduced embeddings are then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is well-suited for text data because it can automatically determine the number of clusters, identify clusters of varying density, and label outliers without requiring a predefined number of topics. This flexibility allows it to capture both dominant and rare incident types in the dataset. In addition to discrete cluster assignments, HDBSCAN produces a soft clustering probability score that reflects the model's confidence that a document belongs to its cluster. Points with low membership probability are labelled as noise, which is advantageous for short or ambiguous reports where incident descriptions may not form coherent themes.

HDBSCAN was implemented using the default BERTopic configuration (*min_cluster_size*=10, *min_samples*=None, *cluster_selection_method*='eom'). The *min_cluster_size* of 10 defines the minimum number of documents required to form a cluster, allowing relatively small but potentially meaningful topics to be retained while limiting the formation of very small clusters. With *min_samples* left unspecified, HDBSCAN derives this value from *min_cluster_size*, retaining its standard density criterion for distinguishing stable clusters from noise. The Excess of Mass ('eom') method selects clusters according to their persistence with the hierarchical density structure, favouring stable clusters rather than requiring a predetermined number of topics. Documents identified as low-density observations were assigned to an outlier class rather than being forced into inappropriate clusters. As with UMAP, these settings were retained as a standardised starting configuration rather than optimised towards a preferred topic solution. This approach is consistent with previous exploratory BERTopic applications that employed the default configuration and allowed the model to determine the topic structure from the data (Tang et al., 2024)

To generate interpretable representations of each cluster, Class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) is used. This approach identified the most representative words and phrases for each cluster by comparing their relative frequency within and across clusters. The weight of each term *t* in a document *d* is computed as:

Equation 1

$$cTFIDF_{t,c} = \frac{f_{t,c}}{|c|} \cdot \log\left(\frac{N}{df_t}\right)$$

Where $f_{t,c}$ is the frequency of term *t* within all documents assigned to cluster *c*, $|c|$ is the total number of terms across documents in cluster *c*, *N* is the total number of documents, and df_t is the number of documents containing term *t*. This formulation amplifies terms that are distinctive to each topic while suppressing globally frequent words. The resulting keywords are used to describe and interpret the semantic meaning of each cluster, providing a foundation for subsequent labelling and thematic analysis.

Documents classified as outliers were excluded from subsequent topic interpretation and thematic analysis. This approach avoided artificially inflating topic coherence by forcing semantically unrelated reports into existing clusters. Given the brevity and operational diversity of many near-miss narratives, retaining an outlier category was considered preferable to maximising document coverage at the expense of interpretability.

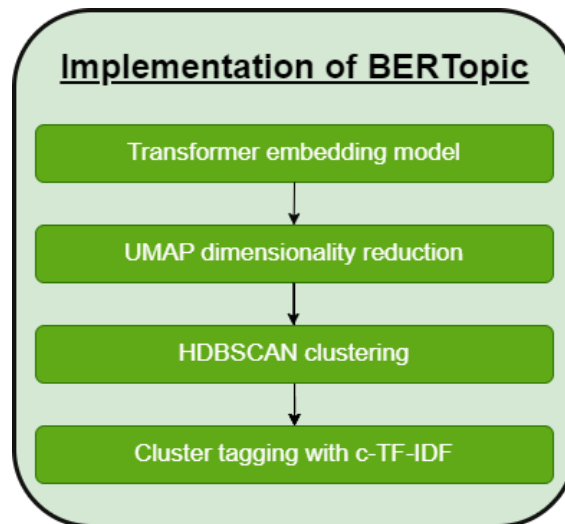


Figure 19. BERTopic Flow Diagram.

4.3.2 Topic Labelling

While BERTopic produced representative keywords for each cluster, these keywords alone may not always yield immediately interpretable topic names. To enhance interpretability and reduce the cognitive burden on human reviewers, a large language model (LLM) such as GPT-4 is integrated into the workflow for automated topic labelling.

The LLM is provided with both the most significant keywords for each cluster, and a representative sample of incident descriptions belonging to that cluster. A one-shot prompt engineering approach is employed, where the model is given an example of an accurately labelled cluster to use as a reference. GPT-4 was accessed programmatically through the OpenAI API. For each cluster, the model was supplied with the top-ranked c-TF-IDF keywords together with representative incident narratives selected based on proximity to the cluster centroid within the embedding space. A standardised one-

shot prompt structure was automatically applied to every cluster, consisting of: (1) a description of the labelling task, (2) an illustrative example demonstrating an appropriate topic label, and (3) cluster-specific keywords and representative texts. To maximise reproducibility and minimise output variability, deterministic model settings were employed with the temperature parameter set to zero. Generated labels were automatically parsed and stored for subsequent expert review and validation. Based on this template, the LLM generates concise, human-readable topic names and groups conceptually similar clusters under broader thematic categories.

This hybrid approach combined the data-driven precision of BERTopic with the semantic interpretability of GPT-based reasoning. It ensures that topics are both linguistically coherent and contextually relevant, minimising subjective variability typically associated with manual topic naming.

Final topic names and thematic groupings are reviewed by domain experts to ensure alignment with maritime safety terminology and operational contexts. This labelling procedure is applied consistently across both case studies, with the LLM interpreting the topics within the terminology and operational context of the respective ferry and offshore wind datasets.

4.3.3 Validation

To ensure the reliability, accuracy, and interpretability of the generated topics, a multi-level validation framework was implemented. The process combines quantitative agreement metrics with qualitative expert evaluation, following established best practices in human-machine validation studies.

The first validation method is mapping LLM-labelled topics to existing organisation taxonomies where the data is available. This ensures conceptual alignment between automatically derived topics and recognised operational classifications.

Additionally expert panel evaluation is important where a panel of subject matter experts (typically with postgraduate qualifications and substantial experience in maritime or safety management) evaluates each topic based on three main criteria:

1. Keyword coherence: Do the extracted keywords form a semantically coherent grouping?
2. Label appropriateness: Does the LLM-generated topic name accurately represent the theme?
3. Label-document fit: Do representative incident descriptions align with the assigned topic?

The number of experts was selected based on recommendations from topic modelling validation and qualitative content analysis studies, which suggest that panels comprising between three and ten domain experts are generally sufficient when participants possess substantial subject-matter expertise. The purpose of validation was not to estimate population-level opinion but rather to evaluate whether generated topics were operationally meaningful within a maritime safety context. To ensure breadth of evaluation, a hybrid validation strategy is used. A subset of core topics is reviewed by all experts for inter-rater reliability, while additional topics are distributed across reviewers to maximise coverage.

Quantitative assessment was then conducted using multiple inter-rater reliability and internal consistency measures. Krippendorff's Alpha was selected for its ability to handle ordinal data and missing ratings, providing a robust measure of agreement across multiple judges. Kendall's W was used to quantify the degree of consensus in ranking topics, while the Intraclass Correlation Coefficient (ICC) assessed absolute agreement among raters in evaluating the same topics. Cronbach's Alpha measured the internal consistency of score scales, ensuring that expert judgements were reliable across criteria. Following Krippendorff (2004) guidelines, values above 0.8 were considered indicative of reliable agreement, while values between 0.667 and 0.8 were considered sufficient only for drawing tentative conclusions. Values below 0.667 were therefore considered below the recommended minimum reliability criterion.

The available validation evidence differs between the two applications according to the information accompanying each dataset. Consequently, validation against existing organisational classifications is undertaken where suitable reference categories are available, while expert assessment provides a

common mechanism for evaluating topic coherence, label appropriateness and alignment between generated topics and the underlying narratives.

4.3.4 Trend and Causal Analysis

Following topic generation and validation, the framework incorporates temporal and causal analytical layers to interpret how patterns in incident data evolve and interact over time. With temporal analysis topic frequencies are aggregated and analysed using time-series methods to reveal dynamic trends. This is split into examination of seasonal trends with monthly variations used to identify demand or weather-related risk patterns. Additionally, yearly analysis is used to track longitudinal shifts related to operational safety focus or external disruptions. The results are visualised through heatmaps or time-frequency plots, allowing for intuitive comparison across topics and time periods.

Incident dates were converted into standard datetime variables and aggregated at both monthly and yearly levels. Topic frequencies were transformed into topic-by-time matrices and row normalisation was applied to account for differences in overall topic prevalence. Specifically, the frequency of each topic within a given time period was divided by the total frequency of that topic across all periods, enabling comparison between topics of differing sizes. Heatmaps were then generated to visualise temporal distributions. To improve visibility of lower-frequency patterns, a power-law (gamma) scaling transformation was applied to the normalised values, enhancing visual contrast while preserving relative trends.

More formally, the normalised frequency for topic (i) during time period (t) is defined as:

Equation 2

$$f'_{i,t} = \frac{f_{i,t}}{\sum_{t=1}^T f_{i,t}}$$

Where $f_{i,t}$ represents the observed frequency of topic i in period t, and T is the total number of time periods. This normalisation enables comparison of temporal patterns between topics irrespective of their absolute frequency.

Association Rule Mining (ARM) is subsequently used to identify recurring co-occurrence relationships between the structured information available within each dataset. The specific antecedents and consequents differ between case studies according to the accompanying safety information. In the ferry near-miss case study, ARM links topics identified from incident narratives with topics identified independently from corresponding prevention narratives, enabling preventive measures associated with particular near-miss themes to be explored. In the offshore wind case study, ARM instead combines identified incident themes with available structured incident information to examine relationships between topics, incident characteristics and safety outcomes. ARM therefore provides a common relational-analysis component within the framework, while the variables included in the analysis are determined by the structure of the individual dataset.

ARM identified statistically significant co-occurrence patterns between categorical variables, expressed as if-then rules. Each association rule is evaluated using three standard metrics: support, confidence, and lift. Support measures how frequently an antecedent and consequent occur together within the dataset and is calculated as:

Equation 3

$$\text{Support } (A \rightarrow B) = \frac{\text{count } (A \cap B)}{N}$$

Where A represents the antecedent, B the consequent, $\text{count } (A \cap B)$ the number of transactions containing both itemsets, and N, the total number of transactions.

Confidence measures the conditional probability of observing the consequent given the antecedent:

Equation 4

$$\text{Confidence } (A \rightarrow B) = \frac{\text{count } (A \cap B)}{\text{count } (A)}$$

Confidence therefore indicates the reliability of the rule, with higher values representing stronger associations between incident and prevention themes.

Finally, lift evaluates the strength of an association relative to random co-occurrence and is calculated as:

Equation 5

$$Lift(A \rightarrow B) = \frac{Support(A \cap B)}{Support(A) \times Support(B)}$$

In this framework, rules were applied to ensure only associations surpassing minimum thresholds for confidence and lift are retained to ensure interpretability and significance. This process enables the discovery of meaningful links between incident types and preventive actions, helping to infer how specific operational failures or conditions may lead to, or prevent, particular outcomes.

Each incident report was converted into a transactional record containing both incident-topic labels and associated prevention-topic labels. Prior to mining, categories occurring fewer than ten times, together with poorly defined labels such as unknown or outlier categories, were removed to improve rule stability and interpretability. Association rules were generated using the Apriori algorithm implemented within the *mlxtend* Python package. Given the exploratory nature of the analysis and the large number of fine-grained topic categories produced by BERTopic, relatively permissive thresholds were employed (minimum support = 0.0025, confidence = 0.07, lift = 1.01). These thresholds allowed infrequent but operationally important incident-prevention relationships to be retained while limiting spurious associations. Rules were constrained to infer prevention measures from incident topics, reflecting the practical objective of identifying preventative actions associated with specific near-miss themes. The retained rules were subsequently visualised as directed bipartite networks, with edge width representing confidence and colour intensity representing lift.

4.4 High-Information Approach (Structured Data from SHIELD)

In cases where incident data is richly annotated, structured, or categorised, for example, datasets containing predefined human factors codes or taxonomies, a high-information analytical approach is applied. This branch of the framework is designed to exploit the structured nature of the data to

develop, train, and validate machine-learning models capable of automatic classification and causal inference.

The goal of this approach is to map unstructured textual content (e.g. causal statements, narrative explanations) to structured taxonomies of human or organisational factors while maintaining interpretability and alignment with domain logic. To achieve this, a multi-stage human-in-the-loop text classification pipeline was developed, incorporating domain adaptation, hybrid language representations, imbalance management, and probabilistic calibration.

4.4.1 Taxonomy Adaptation

Structured taxonomies in safety domains often exhibit label sparsity, where certain categories are under-represented relative to others. Sparse labels can undermine the generalisability of supervised classifiers, as they lack sufficient positive examples to form stable decision boundaries.

To address this, the original taxonomy is consolidated into a balanced label set that maintains conceptual and causal integrity while ensuring adequate data representation across categories.

A minimum support threshold for positive examples is established to determine which codes or categories can be reliably modelled. Labels below this threshold are either merged with semantically adjacent categories or excluded.

In this study, the original SHIELD taxonomy comprising 99 human-factor codes was analysed to determine label frequency distributions. A minimum support threshold of 20 positive instances was established to ensure sufficient representation for supervised learning. Codes falling below this threshold were systematically reviewed. Where conceptually similar low-frequency codes existed, they were merged to preserve the underlying human-factors construct while improving statistical representation. Conceptually distinct but extremely rare codes were removed from the machine-learning taxonomy. This process resulted in a consolidated taxonomy, termed SHIELD-NLP, consisting

of 53 machine-learning-ready labels while retaining the original multi-label and hierarchical structure of SHIELD.

4.4.2 Model Pipeline

This study developed a human-in-the-loop text classification pipeline designed to automatically map free-text causal statements from maritime incident reports to the predefined SHIELD-NLP codes. The unit of analysis in this study is the individual causal sentence extracted from an accident investigation report. Each report may yield multiple causal sentences treated as an independent instance while retaining a reference to its source report. Sentence-level instances inherit the expert-assigned SHIELD-NLP labels associated with the originating report, allowing each instance to be associated with multiple contributing factors in a multi-label representation.

The pipeline comprises five principal stages, shown in Figure 20: (1) document ingestion and causal sentence extraction, (2) hybrid domain-specific language representation, (3) supervision classification with imbalance handling and probability calibration, (4) decision logic incorporating confidence thresholds and domain rules, and (5) model validation. This design supports both automated pre-classification and expert-validated refinement.

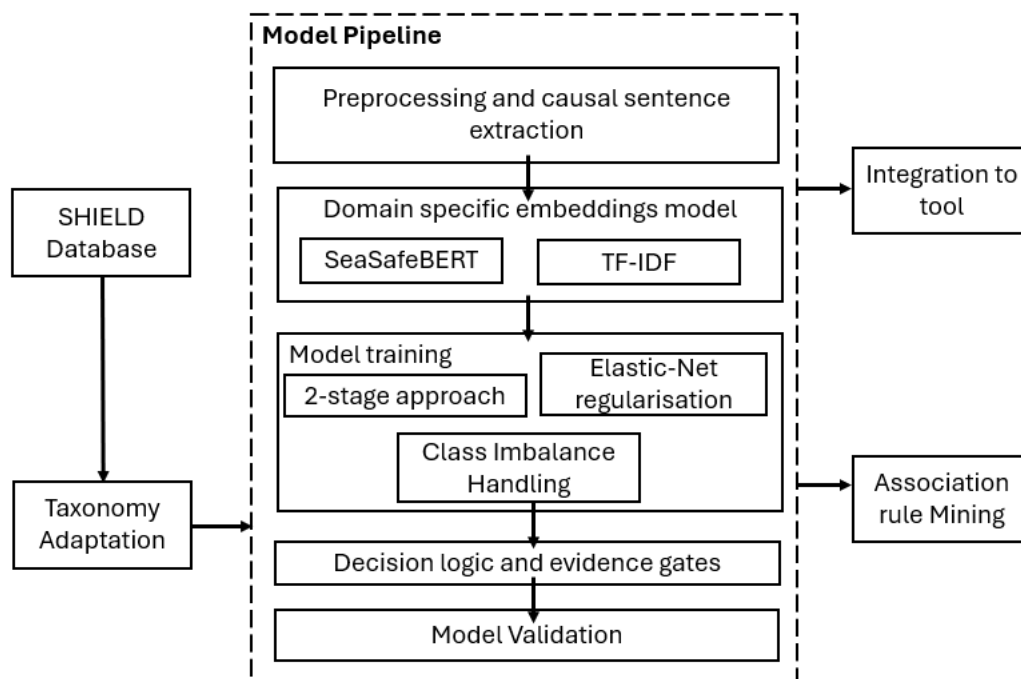


Figure 20. High information approach methodology workflow diagram.

4.4.2.1 Preprocessing and causal sentence extraction

Incident reports often contain a mix of contextual, procedural, and causal information. To improve modelling precision, the framework includes a causal extraction module that isolates sentences most relevant to causal attribution.

This stage involves the following:

- Sentence segmentation: splitting reports into individual sentences.
- Cue scoring: assigning each sentences a score based on the presence of causal markers (e.g. because, led to, resulted in) and penalising non-causal indicators (e.g. section headers, introductions). Length normalisation: filtering out sentences that are too short (<6 tokens) or excessively long (>60 tokens) to reduce false positives.
- Causal selection: Sentences exceeding a causal score threshold of 0.8 were retained for downstream processing. Where insufficient material was extracted, a fallback mechanism retained the top 150 highest-scoring sentences to ensure adequate textual coverage.

This preprocessing step increases the signal-to-noise ratio and improves model accuracy by ensuring that training inputs are focused on the causal reasoning expressed in the reports. Following causal sentence extraction, the dataset was partitioned into training, validation, and testing subsets using a stratified sampling strategy to preserve the distribution of SHIELD-NLP labels across all splits. Initially, an 80:20 split was performed, allocating 80% of the data to model development and reserving 20% as an independent held-out test set for final evaluation. Validation data were then derived exclusively from the training portion by further partitioning the training set into training (80%) and validation (20%) subsets, resulting in an effective allocation of approximately 64% training, 16% validation, and 20% testing data. Stratification was applied at each stage to maintain class prevalence and ensure adequate representation of both parent and subcategory labels. To prevent information leakage, all resampling procedures (SMOTE), probability calibration, and decision-threshold tuning were performed using only the training and validation data. The held-out test set was not used during model fitting, resampling, calibration, or threshold selection and was reserved solely for final performance evaluation.

Splits were performed at the sentence level rather than the report level. While this may allow sentences from the same report to appear in different splits, the use of sentence-level units reflects the intended operational setting in which individual causal statements are classified independently.

4.4.2.2 Hybrid language representation

Accurate classification of human-factor categories requires capturing both semantic context and domain-specific terminology. To achieve this, the framework employs a hybrid embedding strategy that combines both dense semantic embeddings, generated by transformer-based language models to capture contextual meaning and relational structure within sentences, and sparse lexical features, extracted through TF-IDF to preserve sensitivity to critical operational keywords.

To enhance performance on maritime classification tasks, a version of BERT was fine-tuned using the Masked Language Modelling (MLM) objective on a domain-specific corpus of accident reports. This

adapted model is termed SeaSafeBERT. Although alternative approaches such as training a language model from scratch, supervised fine-tuning on labelled datasets, or using general-purpose pre-trained models without adaptation were considered, continued pre-training using MLM was selected due to the limited availability of large, labelled maritime datasets. MLM enables the model to learn domain-specific terminology, contextual relationships, and linguistic patterns directly from unlabelled maritime safety reports while retaining the general language understanding acquired during the original BERT pre-training. Consequently, this approach provides an effective and computationally efficient means of adapting BERT to the maritime safety domain without requiring extensive manually annotated data.

The corpus was constructed from 250 maritime accident investigation reports, comprising reports from the UK Marine Accident Bureau (ATSB) (n=100), Australian Transport Safety Bureau (n=50), Transport Safety Board of Canada (n=50) and the United States National Transportation Safety Board (NTSB) (n=50). These authorities were selected because they provide detailed, narrative-based investigation reports with structured causal analysis and consistent documentation standards. Reports were converted to plain text, and only documents exceeding 500 characters were retained. The text was tokenised using a pre-trained bert-base-uncased tokenizer, which inputs padded or truncated to a fixed length of 128 tokens. The domain-adaptation corpus comprised approximately 3.6 million tokens.

The training was conducted using the HuggingFace Trainer API with the following key hyperparameters: batch size 8, 3 full epochs over the corpus, and a learning rate of $5e-5$. Fine-tuning BERT on this corpus allows the model to internalise the vocabulary, stylistic patterns, and semantic structures characteristics of maritime incident narratives, making SeaSafeBERT better suited than general-purpose language models for downstream tasks. Additionally, all experiments used a fixed random seed (42) to ensure reproducibility. Training was conducted on a CPU-based workstation and took approximately two hours.

Then, to retain sensitivity to domain-critical keywords (e.g. “handover”, “lookout”, “standing orders”), the SeaSafeBERT embeddings were concatenated with a sparse TF-IDF lexical head. This hybrid feature vector balances semantic context with precise operational terminology. The TF-IDF model utilised term-frequency scaling, a minimum document frequency of ≥ 3 , a maximum document frequency cap of 0.95, and unigram and bigram token extraction with maximum 5,000 features.

The final embedding for each sentence was therefore a hybrid vector combining the dense contextual semantics from SeaSafeBERT and the high-precision lexical features from TF-IDF, the final dimensionality of the vector was 5768. This design helps preserve both semantic generalisation and sensitivity to operational vocabulary.

4.4.2.3 Supervised classification and probability calibration

Given the hierarchical structure common to safety taxonomies, the framework employs a two-stage hierarchical classification pipeline which predicts both the parent category and then the subcategory. The parent category is a top-level classifier that predicts broad human-factor domains, and the subcategory is the code level and is a further refined prediction presented within the category level.

Thus, hierarchical setup ensures logical consistency between predictions and the taxonomy structure while improving interpretability and error isolation. These stage two classifiers are applied conditionally based on stage one outputs. A SHIELD-NLP code is emitted only if both its parent category and corresponding stage 2 classifier exceed their respective decision thresholds, enforcing consistency with the SHIELD hierarchy while allowing multiple child codes to be predicted.

To handle data imbalance, a frequent issue in a safety dataset, multiple strategies are combined:

- SMOTE (Synthetic minority oversampling technique) for under-represented classes
- Class weighting and inverse frequency sampling, and
- Regularisation techniques to reduce overfitting and noise.

Finally, probabilistic outputs from classifiers are calibrated to improve reliability. Calibration techniques such as isotonic regression or Platt scaling are applied to align predicted probabilities with observed frequencies, ensuring that the confidence levels of predictions are statistically meaningful.

The classification model used was Logistic Regression, a supervised learning algorithm widely used for categorical prediction tasks. Logistic regression models the probability that an observation belongs to a particular class by fitting a linear combination of input features through the logistic (sigmoid) function, defined as:

Equation 6

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}}$$

Where $P(y = 1|x)$ represents the probability of class membership, x_i are the feature inputs, and β_0 and β_i are the model coefficients estimated through maximum likelihood optimisation.

Unlike linear regression, logistic regression constrains the output to the interval [0,1], allowing probabilistic interpretation. It is particularly advantageous for text classification tasks where interpretability, efficiency, and calibration of predicted probabilities are essential. Each coefficient β_i indicates the direction and strength of association between a predictor and the likelihood of a category being present.

However, in high-dimensional datasets such as those derived from text embeddings or TF-IDF features, multicollinearity and overfitting become major risks. To counter these issues, Elastic-Net regularisation was used, which combined the penalties of both L1 (Lasso) and L2 (Ridge) regularisation. The regularised optimisation objective can be expressed as:

Equation 7

$$\min_{\beta} [L(\beta) + \lambda(\alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2^2)]$$

Where $L(\beta)$ represents the logistic loss function (negative log-likelihood), which measures the model's prediction error. The term λ controls the overall strength of regularisation. The parameter α

determines the balance between L1 and L2 penalties. The L1 term promotes sparsity by driving less informative coefficients towards zero, effectively performing feature selection. The L2 term then shrinks coefficients smoothly, stabilising the model against multicollinearity and small-sample noise. By combining these effects, Elastic-Net regularisation achieves a balance between model simplicity and predictive generalisability, making it particularly suitable for safety datasets where the number of potential text features can be largely relative to the number of labelled examples. This formulation ensures that the classifier remains interpretable, robust to correlated predictors, and resistant to overfitting.

Hyperparameters for the Elastic-Net logistic regression classifier were tuned using grid search cross-validation over $\lambda \in \{0.01, 0.1, 1.0\}$ and $\alpha \in \{0.1, 0.5, 0.9\}$. The configuration yielding the highest macro-F1 score on the validation set was selected.

With this, there is a further probability calibration approach which is implemented to enhance the human-in-the-loop decision mechanism. For this, the predictions are categorised into two confidence states: automatic and suggestive. This logic is driven by dual prediction thresholds:

1. Fixed thresholds: a simple baseline was implemented with a fixed high-confidence threshold of $P \geq 0.5$ (Auto) and a low-confidence threshold of $P < 0.5$ and $P \geq 0.2$ (Suggest).
2. F1-Tuned Thresholds: The primary approach involved tuning the high threshold on the validation set to maximise the F1-score. The corresponding low threshold was then set using the following adaptive rule, where ϵ is a small positive margin and LOW_{min} is a minimum allowable value for the low threshold, and thr_{high} is the F1-optimised high-confidence probability threshold:

$$thr_{low} = \min(0.9 \cdot thr_{high}, \max(LOW_{min}, thr_{high} - \epsilon))$$

To further reduce implausible hits, especially for rare codes, evidence gates (domain rules) were applied after probability scoring. These gates included required keywords, veto keywords and post-probability filters.

To further reduce implausible hits, especially for rare codes, evidence gates (domain rules) were applied after probability scoring. These gates included required keywords, veto keywords and post-probability filters.

4.4.2.4 Model Evaluation

The performance of the classification models was assessed using a combination of standard evaluation metrics to ensure balanced and meaningful interpretation. Accuracy indicates the overall proportion of correct predictions across all categories, offering a general view of performance. However, as safety data are often imbalanced, with some categories much more frequent than others, accuracy alone can be misleading. Therefore, additional metrics were used to provide a more nuanced understanding. Therefore, precision, recall, F1 score and ROC-AUC were prioritised to ensure equal weighting across classes.

Precision measures how many of the model's positive predictions were correct, showing how reliably the system identifies true cases without producing excessive false alarms. In contrast, recall (or sensitivity) reflects how well the model detects all relevant instances of a class. High recall ensures that important but less frequent safety issues are not overlooked. Precision and recall are defined as:

Equation 8

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

Equation 9

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

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

Because precision and recall can trade off against each other, the F1 score, their harmonic mean, was used to provide a balanced indicator of performance and is calculated as:

Equation 10

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

A high F1 score means the model performs well in both identifying relevant cases and avoiding false positives. To evaluate the model's ability to distinguish between classes independently of a single decision threshold, two area-based metrics were applied. The receiver operating characteristic area under the curve (ROC-AUC) assesses how effectively the model separates true positives overall. This provides insight into the underlying separability of the classes, even when threshold metrics such as F1-score may be affected by class imbalance. Together, these metrics offer a comprehensive and reliable picture of model performance in a multi-label, imbalanced safety dataset.

All preprocessing, resampling (SMOTE), class weighting, probability calibration, and threshold optimisation were confined strictly to the training and validation subsets within each split. The held-out test set remained untouched until the final evaluation.

4.4.2.5 Association rule mining for causal mapping

Association Rule Mining (ARM) was applied to identify recurrent patterns of co-occurring contributory factors within the SHIELD-NLP taxonomy. The objective was to uncover statistically meaningful relationships between human, environmental, organisational, and supervisory factors to examine whether these relationships differed across accident types. ARM was conducted using the Apriori algorithm, which identifies frequent items and derives association rules that meet user-defined support, which was set to 5% for this analysis. This means that the rule had to occur in at least 5% of cases to be shown in the network. This was selected to balance the interpretability and statistical reliability. Lower thresholds produced a large number of sparse or unstable rules, while higher thresholds excluded meaningful but less frequent contributory patterns.

Rules were also filtered to be interpreted within the SHIELD causal hierarchy to ensure alignment with the multi-layer pathway (organisational > operational leadership > preconditions > acts). The filtered rules were then visualised as directed networks in which nodes represent SHIELD-NLP factors and edges represent association rules. These networks enabled qualitative exploration of how contributing factors interact within and across taxonomy layers.

ARM was conducted using expert-validated SHIELD-NLP labels, not model predictions, to ensure discovered patterns reflect the incident coding rather than classification artefacts. Formal permutation testing was not conducted in this exploratory phase. Therefore, ARM results should be interpreted as descriptive pattern discovery rather than confirmatory statistical inference.

4.5 Chapter Summary

Overall, this chapter has presented an integrated methodological framework for analysing heterogeneous maritime safety data under varying levels of information availability. Drawing on the findings of the taxonomy review in Chapter 3 and the methodological review in Chapter 5, the framework combines structured human factors classification with AI-based analytical techniques to enable scalable, consistent, and interpretable analysis of both short incident narratives and detailed accident investigation reports. By adopting separate analytical pathways for low-information and high-information datasets, the framework provides a flexible approach capable of supporting a wide range of maritime and offshore safety applications.

The framework established in this chapter provides the methodological foundation for the remainder of the thesis. The following chapter implements and evaluates the high-information pathway through the development of SHIELD-NLP for automated human factors classification from detailed accident reports. The subsequent case study chapters then apply the low-information pathway to maritime near-miss and offshore wind incident datasets, demonstrating the framework's practical applicability, scalability, and transferability across different safety-critical domains.

5. Automated Human Factors Classification in Maritime Accident

Reports using the SHIELD-NLP Framework

5.1 Chapter Overview

This chapter presents the implementation and evaluation of the high-information pathway introduced in the methodological framework (Chapter 6). This pathway was designed for detailed accident investigation reports containing rich textual and contextual information and aims to automate the classification of human factors in maritime incidents using the SHIELD taxonomy. The development of SHIELD-NLP was preceded by an initial proof-of-concept investigation that explored the application of traditional Natural Language Processing techniques to the SHIELD taxonomy. A summary of this preliminary study is provided in Appendix B.

The research presented in this chapter forms the basis of the peer-reviewed publication *A domain-adapted NLP pipeline for human factors classification of maritime accidents* (Black et al., 2026b), with the chapter providing additional detail, analysis, and discussion beyond that presented in the published article. The SHIELD taxonomy serves as a structured framework for identifying and categorising human and organisational factors in maritime accidents, providing a foundation for integrating AI-driven methods. The chapter will first explain basic principles of SHIELD taxonomy and the changes that have been implemented for better use of AI. Next, results of the application of both supervised machine learning and then deep learning methods will be presented.

5.2 Data Acquisition and Development of the Database

The SHIELD taxonomy was selected as the underlying human factors framework for this research because it was specifically developed for maritime accident investigation and provides a comprehensive, hierarchical representation of contributory factors spanning organisational, supervisory, environmental, and individual levels. Unlike more generic frameworks, such as HFACS, SHIELD was designed to reflect the operational context of maritime and offshore industries,

incorporating sector-specific human and organisational factors relevant to accident causation. Furthermore, SHIELD supports multi-label classification, enabling multiple interacting contributory factors to be assigned to a single incident, thereby providing a richer representation of accident causation. A detailed comparison and justification for the selection of SHIELD over alternative human factors frameworks is provided in Chapter 3.

The data used in this research originates from the SHIELD (Safety Human Incident and Error Learning Database) system, which was developed as part of the EU-funded SAFEMODE project. SAFEMODE brought together a consortium of both industrial and academic partners across the aviation and maritime sectors, with the overarching goal of improving the integration of human factors into transport safety management. As part of my research centre's contribution to the project, we supported the development, population, and analysis of the SHIELD database. Through this involvement, I obtained direct access to the SHIELD dataset, enabling its use in this study to adapt the taxonomy and support the development of an NLP-assisted classification framework.

SHIELD contains structured analyses of incident reports from maritime and aviation safety investigation bodies. For the purposes of this research, only maritime incidents were considered. All reports included in SHIELD are publicly available and sourced from official investigation authorities worldwide, such as the UK Marine Accident Investigation Branch (MAIB), the Bahamas Maritime Authority, the Malta Marine Safety Investigation Unit, and the Australian Transport Safety Bureau. This global coverage ensured a diverse and representative dataset, capturing a wide range of vessel types, operational contexts and regulatory environments.

Each incident report in SHIELD underwent a detailed human-factors-oriented analysis performed by trained assessors. The process began with documenting high-level descriptors, including the actors involved, the type of occurrence, and a brief narrative summarising the sequence of events. This descriptive layer provided essential contextual grounding for subsequent human factors analysis.

Following the initial documentation, assessors classified each incident using the SHIELD taxonomy. This required identifying the corresponding layer, category, and specific human factor(s) reflected in the event. Crucially, assessors also recorded a short textual explanation justifying each classification, extracted directly from the incident report. These reasoned justifications ensured traceability, reduced subjectivity, and created an explicit link between the narrative evidence and the assigned human factors codes. This element of the dataset is particularly valuable for the current research, as it enables the alignment of text segments with human factor labels for later use in supervised machine learning models.

Incidents were analysed either individually or collectively by small teams, depending on their complexity. To ensure consistency and minimise assessor bias, all analyses underwent independent review by a second project partner. Any discrepancies between initial assessments and reviewer feedback were resolved through discussion until consensus was reached. This multi-stage validation process enhanced the reliability and quality of the dataset.

Overall, SHIELD provided a comprehensive, curated, and quality-assured set of incident analyses. Its combination of structured classification, narrative justifications, and validated human factors coding made it well-suited for developing the SHIELD-NLP dataset and supported the goals of adapting the taxonomy for automation and improved machine-assisted classification.

5.3 The Safety Human Incident and Error Learning Database (SHIELD) Taxonomy

The SHIELD taxonomy is a structured framework developed to classify human errors and causal factors in maritime and aviation incidents. The taxonomy was created during the SAFEMODE project, funded by the European Union. They brought together multiple concepts which have already been used like James Reason's Swiss Cheese Theory, HFACS (Human Factors Analysis and Classification System) and HERA (Human Error in Air Traffic Management). The taxonomy often references the human factors iceberg where each of the layers finds deeper and deeper reasonings that are not able to be seen from the surface level. Many stakeholders of the project from both aviation and maritime worked together

to determine classifications in the taxonomy. The stakeholders tried to use the taxonomy and provided feedback, leading to changes to the codes and also updates to factor definitions to ensure it is easily understood and applied.

The taxonomy has four layers: Acts, Preconditions, Operational Leadership and Organisational. The “acts” layer focuses on unsafe acts performed by individuals. These are the immediate actions or inaction that directly contributed to an incident. The Acts layer is then further split into five subsection which include the following:

- Perception: These are errors when individuals fail to correctly sense or interpret information in their environment this can be visual, auditory, kinaesthetic or any other means.
- Planning and decision making: This refers to when there is failure to implement a correct plan. This can be that the decision was late, the plan was incorrect or there was no plan at all.
- Intentional deviation: This is when individuals deliberately choose to act contrary to established procedures, rules, or best practices. They can stem from overconfidence or workarounds due to the procedure being outdated or perceived as “unnecessary”.
- Response execution: This is when individuals fail to carry out a planned action as intended. This means that the decision and planning was appropriate, but the actual implementation was not conducted correctly.
- Communicating: Failure to effectively convey or interpret information among team members. Again, here the planning is correct but is communicated in a way which is unclear or is not communicated at all.

The “preconditions” layer then examines the conditions that set the stage for the unsafe acts to occur. This layer refers to any environmental factors or individual operator factors that have affected human performance. There are many subfactors for this layer shown below:

- Physical environment: This addresses external environmental factors like visibility, noise, vibration, temperature, and bad weather.

- Equipment and workplace: This involves the design and conditions of machinery and workplaces. This includes automation, human-machine interface (HMI) and ergonomics.
- Interpersonal communication: This is the reasonings behind miscommunications which can be issues with different languages being spoken, non-standard language and differences in rank.
- Team/group: When poor coordination, lack of trust, no cross-checking or speaking up occur the likelihood of error increases as the team fail to work together and function properly.
- Misperception: Instances where individuals misinterpret sensory input like motion, visual or auditory cues.
- Awareness: An individual's understanding of their surroundings and the current situation. This can be from channelized attention, confusion, distraction or inattention.
- Memory: Failure to recall critical information or procedural steps.
- Mental workload: The cognitive demands placed on an individual during task performance.
- Personal factors: This includes individual characteristics such as emotional state, personality, motivation, performance pressure, psychological condition, confidence and complacency.
- Physiological condition: These are personal physical factors like injury, illness, fatigue, and burnout.
- Drugs and nutrition: Whether prescription, recreational or over the counter, medication, alcohol or insufficient nutrition which can affect the operator performance.
- Competence, skills and capability: This relates to an individual's training, experience, and proficiency in performing tasks.

The “operational leadership” layer then goes a step further by exploring the role of leadership and management in contributing to incidents. This layer includes the following sub-categories:

- Personnel leadership: The actions and decisions of leaders responsible for managing individuals. This can be no measures against risky behaviour, lack of feedback on safety reporting and personality conflicts.

- Operation planning: This is when foresight, preparation and coordination are inefficient e.g. inadequate risk assessment.
- Task Leadership: inadequate leadership of operational tasks including no enforcement of existing rules, allowing unwritten policies to become standard, lack of correction of unsafe practices and directed deviation from procedures.

Lastly, the “organisational” level considers the broader organisational influences such as:

- Culture: This refers to the collective values, beliefs and behaviours that define an organisation's approach to safety.
- Safety management: The systems, policies, and processes designed to maintain safety standards. This includes training, risk management, safety assurance, and safety promotion.
- Resources: When the organisation does not provide sufficient resources for safe operations. This includes personnel, equipment, training, support for suitable design of equipment and procedures.
- Economy and Business: Economy and business pressures such as cost reduction mandates and tight schedules can affect safety.

Importantly, SHIELD is inherently a multi-label human-factors taxonomy. Individual accidents are not assigned to a single mutually exclusive SHIELD code; rather, multiple codes can be assigned to the same accident to represent the combination of contributory factors identified during the investigation. These factors may occur across different levels of the SHIELD hierarchy, including Acts, Preconditions, Operational Leadership and Organisational Influences. Consequently, the number of SHIELD code assignments is substantially greater than the number of accidents in the dataset, reflecting the multi-factorial nature of accident causation rather than duplicate observations.

5.4 Refining and structuring of SHIELD-NLP

While the original SHIELD taxonomy provides a richly detailed framework for analysing human factors in maritime incidents, its granularity poses significant challenges for supervised machine learning. Many SHIELD codes occur infrequently, resulting in a long-tailed distribution with substantial data sparsity. For multi-label classification, each label requires a sufficient number of positive training examples to allow the model to learn a stable decision boundary and generate calibrated probability estimates. Preliminary exploratory analysis showed that a large proportion of SHIELD codes appeared fewer than ten times across the dataset, with some occurring only once. Early modelling experiments confirmed that classifiers trained on classes with fewer than approximately 20 positive examples exhibited highly unstable behaviours, including overfitting, inconsistent decision thresholds, and unreliable calibration.

To overcome these limitations, an adapted and condensed taxonomy, SHIELD-NLP, was developed. The aim of SHIELD-NLP was to retain the causal and theoretical intent of the original taxonomy while ensuring that each class had sufficient representation for machine learning. The construction of SHIELD-NLP followed a multi-stage process.

First, all SHIELD codes were analysed to determine their frequency across the dataset. A minimum support threshold of 20 positive examples was therefore adopted as a pragmatic, empirically informed screening criterion. This threshold was selected as a compromise between providing sufficient representation for supervised modelling and retaining as much of the original SHIELD taxonomy and associated incident information as possible. The choice is also consistent with published low-resource NLP research: Karamanolakis et al. (2024), for example, explicitly used 20 labelled examples per class to simulate a low-resource text-classification setting. Their results further demonstrate the challenges associated with conventional supervised learning at this level of data availability, reinforcing the need to avoid categories with still lower representation where possible.

Importantly, the threshold was not treated as a universal minimum sample-size required or an automatic exclusion criterion. Rather, it identified codes requiring further conceptual review. Codes falling below this threshold were individually reviewed without compromising clarity. Where merging was appropriate, for example, when two codes differed only in operational context but referred to the same underlying human factor, they were combined into a single SHIELD-NLP category. In cases where codes were both rare and conceptually isolated, they were removed from the taxonomy for the purpose of automated classification, as their inclusion would reduce model robustness.

Throughout the consolidation process, efforts were made to preserve the hierarchical logic and meaning of the original SHIELD taxonomy. Assessors and domain experts familiar with SHIELD reviewed the proposed mappings to verify that the newly formed categories remained interpretable and aligned with human factors theory. The final SHIELD-NLP taxonomy therefore reflects a balance between theoretical fidelity and practical learnability, meaning it reduced sparsity while maintaining clear distinctions between major causal mechanisms.

The first stage of the process looked to identify low-support codes. The analysis of this found that from the original 99 codes, 33% had less than 10 examples and 56% had fewer than 20 examples. Consequently, adopting the 20-example threshold already required more than half of the original taxonomy to undergo conceptual review. A higher threshold would have further increased the number of affected categories and risked unnecessary loss of taxonomic granularity and incident information. The threshold of 20 was therefore retained as the lowest empirically defensible screening point identified during preliminary modelling, maximising information retention while addressing the most severe instances of label sparsity.

A full account of how the original SHIELD codes were merged or removed is provided in Appendix C: SHIELD-NLP Development Tables within Table 41 and Table 42. These tables document all consolidation decision and the rational for excluding low-support or conceptually isolated categories.

After consolidation, a final set of 53 SHIELD-NLP labels remained. These labels provided sufficient positive examples for training and supported a more stable multi-label classification approach. Importantly, the multi-label structure of the dataset was preserved, allowing each incident to retain its association with multiple contributing factors. The resulting SHIELD-NLP dataset therefore serves as a machine-learning-ready version of the original SHIELD data, enabling automated classification experiments while maintaining meaningful alignment with human factors analysis. The finalised version of SHIELD-NLP is shown within Figure 21.

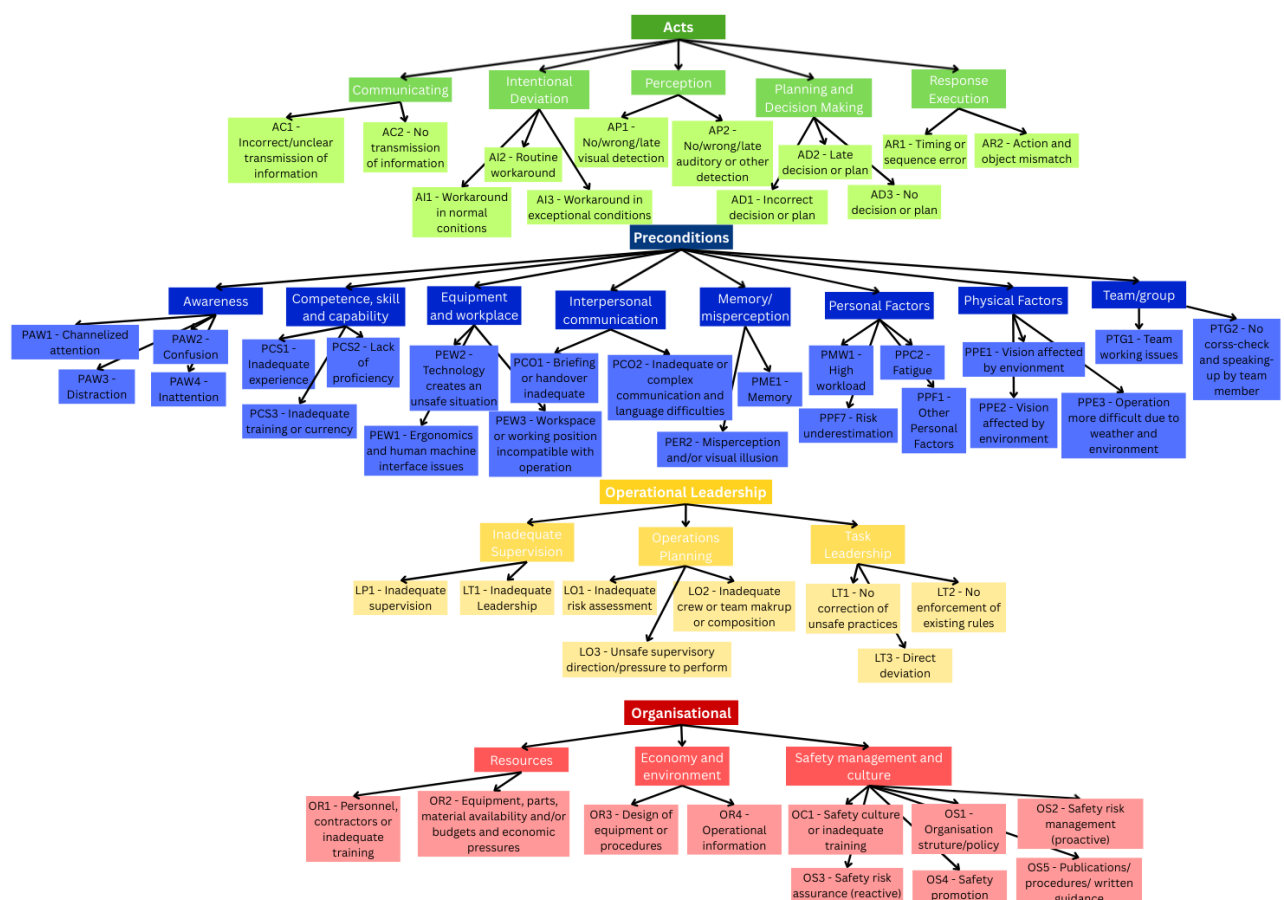


Figure 21. Final SHIELD-NLP Taxonomy.

5.5 Descriptive Statistics of SHIELD-NLP Dataset

5.5.1 General Statistics

The obtained and refined dataset was comprised on 330 maritime incidents, providing a rich foundation for analysing human factors using SHIELD taxonomy. However, before conducting descriptive analyses, significant data cleaning and preprocessing was necessary to ensure accuracy, consistency, and reliability of the dataset.

The final dataset captures incidents from a wide span of time periods, ranging from 1992 to 2022, allowing for analysis of trends in human factors over time. In addition to the SHIELD classifications, each of these incidents is characterised by several other key attributes, including the level of injury, damage level, and incident type, allowing for a broader characterisation of accident outcomes.

Analysis of severity revealed that the majority of the incidents had no injury (68%), some had minor (11%) or fatal (12%) injuries, and few had serious injuries (7%). With regards to damage to vessels around 52% of the incidents has substantial damage, 23% had minor damage and 14% had no damage. Notably, approximately 10% of the incidents analysed resulted in the total loss or destruction of the vessel, highlighting the potential severe consequences associated with maritime accidents.

Figure 22 illustrates the distribution of incident types within the analysed dataset. Collisions and groundings dominate the dataset, accounting for 38% and 36% of the incidents, respectively. These types of incidents are often emphasized in safety investigations due to their potential for significant operational and environmental impacts. Fire and explosion incidents follow, making up approximately 15% of the dataset, while confined space incidents and falls from height are reported less frequently. Other categories, such as contact, man overboard, capsizing, and oil leaks, appear in relatively small numbers.

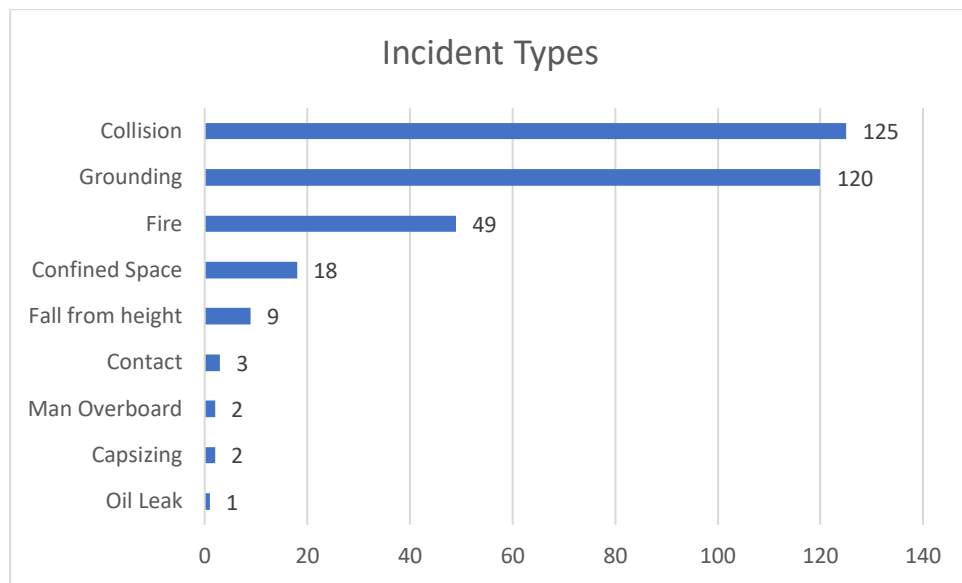


Figure 22. Bar chart of incident types from SHIELD analysed data.

It is important to note that the distribution of incident types shown in Figure 22 reflects the specific dataset used in this study and may not fully represent the actual distribution of maritime incidents occurring globally. For example, data from the European Maritime Safety Agency (EMSA) in their 2024 report highlights that the most common types of incidents from 2014 to the present have been loss of control (such as loss of propulsion or power), collisions, contact, damage/loss of equipment, and groundings.

One major limitation of this dataset is the absence of incidents classified as loss of control, despite their prominence in EMSA's data. This discrepancy may be attributed to the dataset's composition and the focus of the contributing agencies. Additionally, EMSA reports that loss of control incidents have been increasing in recent years. Since the dataset used in this study spans three decades, it is possible that these incidents were not as prevalent during earlier time periods and, therefore, are underrepresented or missing entirely from the database.

This limitation underscores the importance of contextualizing the findings within the scope and focus of the dataset. While the dataset provides valuable insights into certain types of incidents, such as collisions and groundings, it does not capture the full spectrum of incident types reported globally. This gap highlights the need for incorporating additional datasets, such as those from EMSA or other

global safety organizations, to provide a more comprehensive understanding of maritime safety trends.

5.5.2 Human Factors Statistics

From the 330 analysed incidents in the dataset, a total of 2842 human factors were identified, classified across the layers of the SHIELD taxonomy. These human factors reflect the underlying actions, conditions, and organisational influences that contributed to the incidents.

The distribution of these factors across the taxonomy layers reveals that 1,143 factors were classified under the Acts layer, representing unsafe actions performed by individuals. The Preconditions layer accounted for 1,078 factors, capturing the conditions that influenced these unsafe acts. Additionally, 619 factors were attributed to Operational Leadership, focusing on the role of leadership and supervisory influences, while 664 factors were categorised under Organisational Influences, highlighting the systemic and institutional factors impacting safety.

The following tables describe the frequency distribution of SHIELD-NLP codes across the four taxonomy layers following the consolidation process. An average of 8.6 contributing factors per incident was found.

Table 3 presents the distribution of codes within the Acts layer, which captures unsafe acts directly performed by 6 individuals. The most frequently occurring category in this layer was “incorrect decision or plan (AD1)” with 251 instances (8.8% of all annotations), indicating that decision-making failures were the most dominant immediate contribution to incidents. This was followed by “No transmission of information (AC2)” with 158 instances (5.6%), and “No/wrong/late visual detection (AP1)” with 113 instances (4%), highlighting the critical role of communication breakdown and perceptual failures in maritime accidents.

Overall, the acts layer is dominated by failures in decision-making, information transmission, and perceptual detection, indicating that front-line operators actions remain a central contributor to incident causation in the dataset.

Table 3. Frequency distribution of SHIELD-NLP codes within the Acts layer.

Code	N	%
AC1 - Incorrect/unclear transmission of information	72	2.5%
AC2 - No transmission of information	158	5.6%
AD1 - Incorrect decision or plan	251	8.8%
AD2 - Late decision or plan	48	1.7%
AD3 - No decision or plan	63	2.2%
AI1 - Workaround in normal conditions	59	2.1%
AI2 - Routine workaround	20	0.7%
AI3 - Workaround in exceptional conditions	33	1.2%
AP1 - No/wrong/late visual detection	113	4.0%
AP2 - No/wrong/late auditory or other detection	25	0.9%
AR1 - Timing or sequence error	66	2.3%
AR3 - Action and Object Mismatch	35	1.2%

Table 4 then summarises the distribution of codes within the Preconditions layer, which reflects environmental, cognitive, physiological, and team-related factors that influence performance. Within this layer, “Technology created an unsafe situation (PEW2)” was the most frequent category with 87 occurrences (3.1%), highlighting the substantial impact of technological and human-machine interface issues on operational safety.

Cognitive and attention factors were also prominent. “Fatigue (PPC2)” appeared 54 times (1.9%), while “Personal factors (PPF1)” and “Inadequate experience (PCS1)” accounted for 1.7% and 1.9% respectively. Visual and environmental stressors, including “Vision affected by environment (PPE1)” at 1.5% and “Operation more difficult due to weather and environment (PPE6)” at 1.6%, further illustrate the role of physical working conditions in degrading performance.

Team and communication-related preconditions, such as “Inadequate or complex communication (PCO2)”, “Team working issues (PTG1)”, and “No cross-check and speaking-up (PTG2)”, together

formed a consistent, though individually moderate, proportion of the dataset. This pattern suggests that while no single precondition dominates, a wide range of interacting cognitive, environmental, and team-related stressors contribute to unsafe acts.

Table 4. Frequency distribution of SHIELD-NLP codes within the Preconditions layer.

Code	N	%
PAW1 - Channelized attention	24	0.8%
PAW2 - Confusion	26	0.9%
PAW3 - Distraction	38	1.3%
PAW4 - Inattention	39	1.4%
PCO1 - Briefing or handover inadequate	24	0.8%
PCO2 - Inadequate or complex communication and language difficulties	27	1.0%
PCS1 - Inadequate experience	53	1.9%
PCS2 - Lack of proficiency	51	1.8%
PCS3 - Inadequate training or currency	31	1.1%
PER2 - Misperception and/or Visual illusion	34	1.2%
PEW1 - Ergonomics and human machine interface issues	29	1.0%
PEW2 - Technology creates an unsafe situation	87	3.1%
PEW3 - Workspace or working position incompatible with operation	32	1.1%
PME1 - Memory	31	1.1%
PMW1 - High workload	28	1.0%
PPC2 - Fatigue	54	1.9%
PPE1 - Vision affected by environment	42	1.5%
PPE2 - Environmental stressors on human performance	35	1.2%
PPE6 - Operation more difficult due to weather and environment	45	1.6%
PPF1 - Personal Factors	47	1.7%
PPF7 - Risk underestimation	39	1.4%
PTG1 - Team working issues	22	0.8%
PTG2 - No cross-check and speaking-up by team member	25	0.9%

The distribution of operational leadership factors is shown in Table 5. “Inadequate leadership or supervision (LT1)” represents the most prevalent leadership-related factor, occurring 123 times (4.3%), indicating that supervisory and leadership failures are a major upstream contributor to unsafe operations. This is closely followed by “No enforcement of existing rules (LT3)” with 100 occurrences (3.5%), suggesting that procedural non-compliance is often tolerated to inadequately controlled at the supervisory level.

Risk-related leadership failures were also significant, with “Inadequate risk assessment (LO1)” accounting for 108 instances (3.8%). In contrast, more direct forms of supervisory pressure, such as “Unsafe supervision direction/pressure to perform (LO3)” and “Direct deviation (LT5)”, occurred less frequently, each contributing under 1% of total annotations.

The dominance of supervision, risk assessment, and rule enforcement failures highlights the important role of middle management and operational leadership in shaping the conditions under which front-line errors occur.

Table 5. Frequency distribution of SHIELD-NLP codes within the Operational Leadership layer.

Code	N	%
LO1 - Inadequate risk assessment	108	3.8%
LO2 - Inadequate crew or team makeup or composition	43	1.5%
LO3 - Unsafe supervisory direction/pressure to perform	24	0.8%
LP1 - Inadequate Supervision	31	1.1%
LT1 - Inadequate leadership or supervision	123	4.3%
LT2 - No correction of unsafe practices	44	1.5%
LT3 - No enforcement of existing rules	100	3.5%
LT5 - Directed deviation	23	0.8%

Table 6 presents the distribution of organisational factors, which represent higher-level systemic influences. The most common organisational contributor was “Publications, procedures, and written guidance (OS5)” with 99 instances (3.5%), suggesting that deficiencies in documented procedures, guidance, or manuals play a substantial role in accident causation. This was closely followed by “Equipment, parts, materials availability and/or economic pressures (OR3)” at 90 occurrences (3.2%), reflecting the influence of resource constraints and economic pressures on safety performance.

Proactive and reactive safety management also features prominently. “Safety risk management (OS2)” accounted for 76 annotations (2.7%), while “Safety risk assurance (OS3)” and “Safety promotion (OS4)” together represented a further 2.4%. Structural and policy-related factors (OS1) comprised 1.8% of all annotations, demonstrating that formal organisational arrangements also contribute meaningfully to risk.

These findings indicate that a substantial proportion of incident causation can be traced to systemic weaknesses in organisational governance, resources, and safety management processes.

Table 6. Frequency distribution of SHIELD-NLP codes within the Organisational layer.

Code	N	%
OC1 - Safety culture and multi-cultural factors	31	1.1%
OR1 - Personnel, contractors or inadequate training	41	1.4%
OR3 - Equipment/parts/materials availability and/or budgets and economic pressures	90	3.2%
OR5 - Design of equipment or procedures	45	1.6%
OR6 - Operational information	41	1.4%
OS1 - Organisation structure / policy	50	1.8%
OS2 - Safety risk management (proactive)	76	2.7%
OS3 - Safety risk assurance (reactive)	37	1.3%
OS4 - Safety promotion	30	1.1%
OS5 - Publications / procedures / written guidance	99	3.5%

5.5.3 Text-based Analysis

The distribution of justification text lengths was analysed to characterise the textual input used for the NLP models. In total, 2842 individual textual justifications were available after data cleaning and preprocessing. The texts had a mean length of 30.2 tokens and a median of 28 tokens, indicating that most justifications consist of short to moderately sized explanatory statements rather than extended narratives, since they are based on small excerpts from full accident reports. The standard deviation of 13.4 tokens suggests a moderate degree of variability in justification length across the dataset.

The shortest justification contained 3 tokens, while the longest contained 103 tokens, showing that although most entities are concise, a small number of detailed explanations are also present. This distribution reflects the structured nature of the SHIELD coding process, where assessors typically provide brief, focused textual evidence to justify each assigned human factor code. From a modelling perspective, the relatively compact sequence lengths are well-suited to transformer-based language models, as they reduce the need for aggressive truncation while preserving the full semantic context of most inputs.

Figure 23 illustrates the distribution of justification text lengths in terms of token counts. The majority of texts fall within the 20-50 token range, with 943 justifications between 20 and 30 tokens and a further 1081 between 30 and 50 tokens, confirming that most SHIELD-NLP justifications are concise but sufficiently descriptive. Shorter texts and longer explanations are less common.

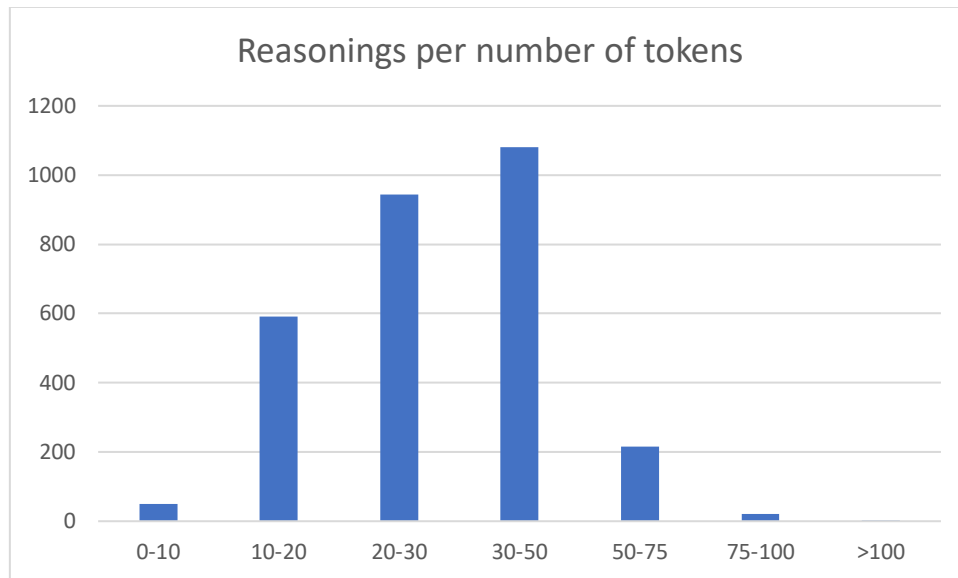


Figure 23. Distribution of SHIELD-NLP justification text lengths measures in tokens.

5.6 Performance Evaluation of the Two-Stage SHIELD-NLP Classification Framework

This section presents the performance evaluation of the proposed two-stage classification framework, which aims to automatically assign high-level categories (Stage 1) and fine-grained causal codes (Stage 2) to narrative text. Model performance is assessed using the area under the receiver operating characteristic curve (ROC-AUC) as a threshold-independent measure of discriminative capacity, and the F1 score as the primary measure of classification quality at the fixed operating threshold used for deployment.

5.6.1 Stage 1: High-Level Causal Category Performance

Table 7. Macro and Micro summary statistics of Stage 1 model performance.

Performance Metrics	Macro	Micro
Precision	0.2623 (0.0104)	0.2475 (0.0144)

Recall	0.4872 (0.0423)	0.5114 (0.0325)
F1 score	0.3084 (0.0255)	0.3335 (0.0198)
ROC-AUC	0.8168 (0.0090)	0.8219 (0.0177)

Stage 1 evaluates the model's ability to identify seven high-level causal categories derived from the SHIELD framework. Table 8 summarises the performance across all categories. Overall, as shown in Figure 24 and Table 7, the ROC-AUC results show considerable variation in discriminative performance across categories. Some categories show clear separation between positive and negative cases, whereas others exhibit substantially weaker discrimination. This variability indicates that the model's ability to distinguish between classes is dependent on the specific factor being classified, rather than being consistent across the taxonomy.

The highest classification quality, as measured by F1-score, was achieved for Physical Environment (F1=0.65), followed by Perception (F1=0.54) and Competence, skills and capability (F1=0.53). The Physical environment category in particular showed very high precision (0.81) combined with a strong recall (0.54), indicating that when the model identifies this category, it does so with high reliability. Similarly, Perception demonstrated the highest recall among the top-performing categories (0.57), indicating effective sensitivity in identifying perceptual failures in free-text explanations.

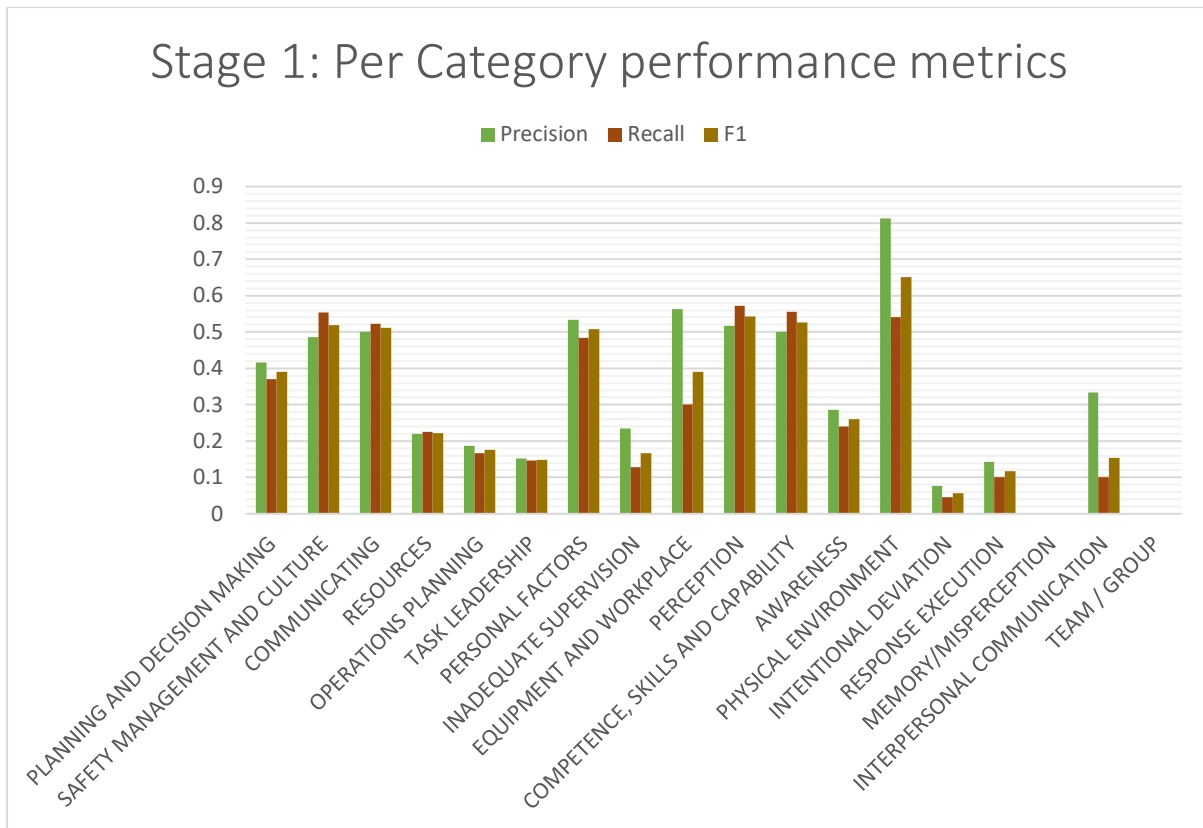


Figure 24. Bar chart of per-category performance results.

Several additional categories achieved robust and balanced performance, including Safety management and culture (F1=0.52), Communicating (F1=0.51), and Personal factors (F1=0.51). These results are particularly encouraging given the abstract and system-level nature of these constructs, suggesting that the model is able to capture high-order organisational concepts from narrative language.

Lower performance was observed in a small number of categories, including International deviation, Response execution, Memory/misperception, and Team/group. These categories also had the smallest number of positive test samples, and in several cases no correct positive predictions were made at the default classification threshold. This outcome is consistent with the known effects of class imbalance and sparse supervision, and does not indicate a lack of underlying signal, as several of these categories still demonstrated moderate to high ROC-AUC values, implying that more favourable operating thresholds or increasing training data may yield improved results.

Overall, Stage 1 results demonstrate that the model can reliably discriminate between the majority of high-level human and organisational causal domains, providing a strong foundation for automated pre-classification and expert decision support.

Table 8. Stage 1 classification performance for high-level SHIELD-NLP causal categories.

	n_pos_test	roc_auc	precision	recall	f1
PLANNING AND DECISION MAKING	73	0.808164	0.415385	0.369863	0.391304
SAFETY MANAGEMENT AND CULTURE	65	0.874958	0.486486	0.553846	0.517986
COMMUNICATING	46	0.849331	0.5	0.521739	0.510638
RESOURCES	40	0.646911	0.219512	0.225	0.222222
OPERATIONS PLANNING	36	0.683464	0.1875	0.166667	0.176471
TASK LEADERSHIP	34	0.669449	0.151515	0.147059	0.149254
PERSONAL FACTORS	33	0.799392	0.533333	0.484848	0.507937
INADEQUATE SUPERVISION	31	0.602321	0.235294	0.129032	0.166667
EQUIPMENT AND WORKPLACE	30	0.821362	0.5625	0.3	0.391304
PERCEPTION	28	0.856745	0.516129	0.571429	0.542373
COMPETENCE, SKILLS AND CAPABILITY	27	0.777778	0.5	0.555556	0.526316
AWARENESS	25	0.668762	0.285714	0.24	0.26087
PHYSICAL ENVIRONMENT	24	0.848901	0.8125	0.541667	0.65
INTENTIONAL DEVIATION	22	0.578961	0.076923	0.045455	0.057143
RESPONSE EXECUTION	20	0.580128	0.142857	0.1	0.117647
MEMORY/MISPERCEPTION	13	0.532063	0	0	0
INTERPERSONAL COMMUNICATION	10	0.597212	0.333333	0.1	0.153846
TEAM / GROUP	9	0.879513	0	0	0

5.6.2 Stage 2: Fine-grained SHIELD-NLP Code Performance

Table 9. Macro and Micro summary statistics of Stage 1 model performance.

	Macro	Micro
Precision	0.2262 (0.0225)	0.2579 (0.0132)
Recall	0.2202 (0.0142)	0.2579 (0.0132)
F1 score	0.2115 (0.0132)	0.2579 (0.0132)
ROC-AUC	0.6546 (0.0116)	0.6682 (0.0121)

Stage 2 evaluates the model's ability to predict the 53 fine-grained SHIELD-NLP causal codes nested beneath the high-level categories, shown in Table 9 and Figure 25. As expected for a real-world safety dataset, performance varied substantially across codes, reflecting differences in class frequency, linguistic distinctiveness, and conceptual abstraction. Overall performance decreased, which is expected due to the large number of labels, substantially lower support per-code and increased semantic overlap between finer-grained codes.

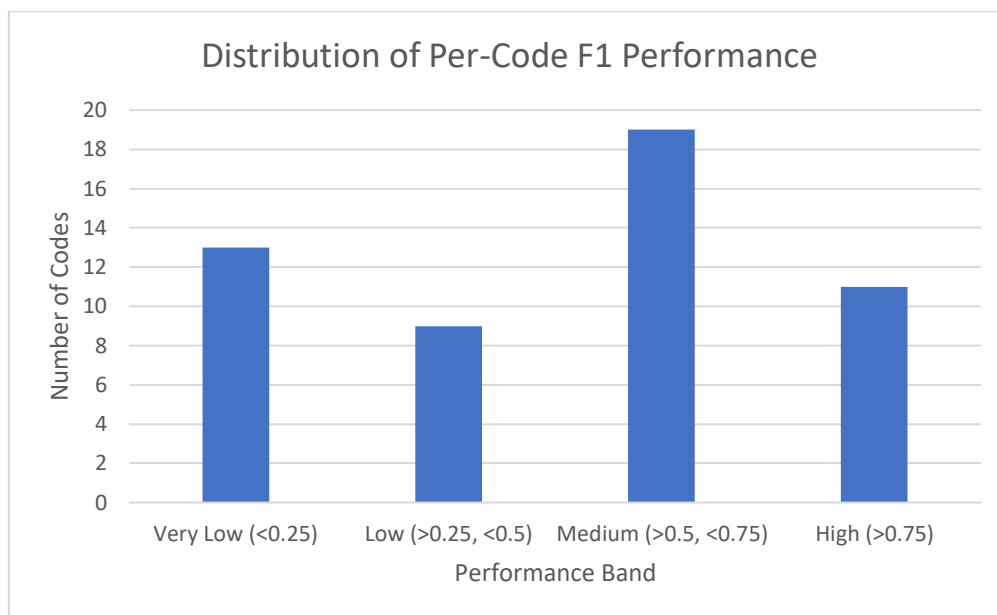


Figure 25. Bar chart showing the number of codes in each performance band with F1 score.

A substantial subset of codes exhibited exceptionally strong classification performance, demonstrating the feasibility of direct operational automation for these factors. Notable examples include:

- AP1 – No/wrong/late visual detection (F1=1),
- AP2 – No/wrong late auditory or other detection (F1=1)
- OR1 – Personnel/contractors/inadequate training (F1=1)
- LO1 – Inadequate risk assessment (F1=0.92)
- LT1 – Inadequate leadership or supervision (F1=0.85)
- AC2 – No transmission of information (F1=0.85)
- AD1 – Incorrect decision or plan (F1=0.82)

- PCO2 – Complex/inadequate communication and language difficulties (F1=0.83)
- PPC2 – Fatigue (F1=0.76)
- PCS1 – Inadequate experience (F1=0.72)
- PME1 – Memory (F1=0.71)

These results indicate that for many high-impact human and organisational risk factors, the language used in investigation justifications is highly distinctive and consistently learnable, enabling the model to achieve both strong precision and recall.

Several high-performing codes demonstrate informative precision-recall trade-offs. For example, ACS (No transmission of information) and LT1 (Inadequate leadership or supervision) achieved very high recall, 0.91 and 0.92 respectively. Indicating that the model successfully identified almost all true instances. While this results in slightly reduced precision due to some over-prediction, such behaviour is highly desirable for safety screening and causal discovery, where missing contributing factors are often more critical than generating additional candidates for expert review.

Conversely, codes such as PPE1 (Vision affected by environment) and PPE6 (Operating more difficulty due to environment) achieved perfect or near-perfect precision, meaning that when the model makes a positive prediction it is consistently correct, although some true instances remain undetected. This profile is well-suited for high-confidence automated flagging in operational workflows.

Behavioural adaptation and workload-related codes also demonstrated encouraging performance including:

- AI1 – Workaround in normal conditions (F1=0.54)
- PMW1 – High workload (F1=0.40)
- PER2 – Misperception/visual illusion (F1=0.67)
- PCO1 – Briefing or handover inadequate (F1=0.75)

These Results suggest that the model is able to capture subtle behavioural and cognitive patterns that are not explicitly procedural in nature, which is a significant achievement given the abstract nature of such constructs.

A small number of codes returned F1-scores of zero, including OS1, LP1, PPF1, PAW2, PAW4 and AI2. These codes share two important characteristics: (i) they have very small test-set support (typically fewer than 7 instances), and (ii) they are highly abstract semantic definitions, which are often expressed indirectly in narrative form.

Importantly, several of these low-F1 codes still exhibited moderate to high ROC-AUC values, indicating that the model has learned partial separation but that fixed-threshold decision-making limits measurable performance. This reinforced the conclusion that data scarcity rather than model capacity is the principal limitation for these categories.

Table 10. Stage 2 classification performance for fine-grained SHIELD-NLP causal codes.

	N test	roc_auc	precision	recall	f1
AD1 - Incorrect decision or plan	49	0.79	0.78	0.86	0.82
AD2 - Late decision or plan	9	0.76	0.50	0.11	0.18
AD3 - No decision or plan	15	0.85	0.50	0.33	0.40
OC1 - Safety culture and multi-cultural factors	6	0.82	0.50	0.17	0.25
OS4 - Safety promotion	10	0.80	0.00	0.00	0.00
OS2 - Safety risk management (proactive)	17	0.58	0.14	0.06	0.08
OS1 - Organisation structure / policy	11	0.50	0.00	0.00	0.00
OS5 - Publications / procedures / written guidance	15	0.77	0.50	0.47	0.48
OS3 - Safety risk assurance (reactive)	6	0.93	0.33	0.17	0.22
AC2 - No transmission of information	32	0.82	0.81	0.91	0.85
AC1 - Incorrect/unclear transmission of information	14	0.82	0.70	0.50	0.58
OR5 - Design of equipment or procedures	12	0.69	0.67	0.50	0.57
OR3 - Equipment/parts/materials availability and/or budgets and economic pressures	12	0.88	0.63	0.42	0.50
OR1 - Personnel, contractors or inadequate training	6	1.00	1.00	1.00	1.00
OR6 - Operational information	10	0.85	0.67	0.60	0.63
LO2 - Inadequate crew or team makeup or composition	10	0.86	0.64	0.90	0.75
LO1 - Inadequate risk assessment	18	0.96	0.89	0.94	0.92

LO3 - Unsafe supervisory direction/pressure to perform	6	0.69	0.00	0.00	0.00
LT3 - No enforcement of existing rules	16	0.65	0.56	0.88	0.68
LT5 - Directed deviation	6	0.84	1.00	0.33	0.50
LT2 - No correction of unsafe practices	10	0.63	0.33	0.10	0.15
PPF7 - Risk underestimation	5	0.72	0.25	0.40	0.31
PPF1 - Personal Factors	12	0.63	0.00	0.00	0.00
PPC2 - Fatigue	11	0.88	0.80	0.73	0.76
PMW1 - High workload	5	0.78	0.40	0.40	0.40
LT1 - Inadequate leadership or supervision	25	0.66	0.79	0.92	0.85
LP1 - Inadequate Supervision	6	0.73	0.00	0.00	0.00
PEW2 - Technology creates an unsafe situation	18	0.67	0.71	0.56	0.63
PEW1 - Ergonomics and human machine interface issues	7	0.63	0.40	0.29	0.33
PEW3 - Workspace or working position incompatible with operation	5	0.87	0.75	0.60	0.67
AP1 - No/wrong/late visual detection	25	1.00	1.00	1.00	1.00
AP2 - No/wrong/late auditory or other detection	3	1.00	1.00	1.00	1.00
PCS2 - Lack of proficiency	9	0.82	0.83	0.56	0.67
PCS3 - Inadequate training or currency	6	0.62	0.40	0.33	0.36
PCS1 - Inadequate experience	12	0.77	0.69	0.75	0.72
PAW4 - Inattention	7	0.63	0.00	0.00	0.00
PAW2 - Confusion	9	0.58	0.00	0.00	0.00
PAW3 - Distraction	7	0.89	0.75	0.43	0.55
PAW1 - Channelized attention	2	0.98	0.67	1.00	0.80
PPE6 - Operation more difficult due to weather and environment	10	0.96	1.00	0.50	0.67
PPE1 - Vision affected by environment	7	0.99	1.00	0.57	0.73
PPE2 - Environmental stressors on human performance	7	0.76	0.75	0.43	0.55
AI1 - Workaround in normal conditions	8	0.57	0.39	0.88	0.54
AI3 - Workaround in exceptional conditions	12	0.63	0.33	0.08	0.13
AI2 - Routine workaround	2	0.80	0.00	0.00	0.00
AR1 - Timing or sequence error	12	0.51	0.60	0.75	0.67
AR3 - Action and Object Mismatch	8	0.48	0.29	0.25	0.27
PER2 - Misperception and/or Visual illusion	5	0.89	0.57	0.80	0.67
PME1 - Memory	8	0.89	0.83	0.63	0.71
PCO1 - Briefing or handover inadequate	4	0.85	0.75	0.75	0.75
PCO2 - Inadequate or complex communication and language difficulties	6	0.85	0.83	0.83	0.83
PTG2 - No cross-check and speaking-up by team member	4	0.40	0.43	0.75	0.55
PTG1 - Team working issues	5	0.40	0.50	0.20	0.29

5.6.3 Model Performance Comparisons

To quantify the contribution of domain adaptation and hybrid lexical-contextual representations a comparison was conducted for the four feature representations: TF-IDF only, generic BERT, SeaSafeBERT, and the hybrid representation proposed. All models used the same multinomial logistic regression classifier and were evaluated using 5-fold stratified cross-validation, reporting mean and variability across folds.

Table 11. Model performance comparison across different feature representation methods.

Model	Macro-F1	Micro-F1	Macro-AUC	Micro-AUC
TF-IDF	0.0460 (± 0.0506)	0.0605 (± 0.0751)	0.7646 (± 0.0181)	0.6917 (± 0.0520)
Generic BERT	0.2943 (± 0.0236)	0.3147 (± 0.0173)	0.8070 (± 0.0100)	0.8150 (± 0.0151)
SeaSafeBERT	0.2958 (± 0.0114)	0.3207 (± 0.0096)	0.8080 (± 0.0080)	0.8167 (± 0.0035)
Hybrid Model	0.3084 (± 0.0254)	0.3335 (± 0.0198)	0.8168 (± 0.0090)	0.8219 (± 0.0176)

Across folds, contextual transformer embeddings substantially outperformed TF-IDF, as shown in Table 11. Among transformer variants, the hybrid representation model achieved the highest performance. Although the differences between the models were modest, suggesting that domain adaptation provides incremental gains under the current corpus size.

5.7 Deployment of the SHIELD-NLP Decision Support Tool

To demonstrate the practical applicability of the SHIELD-NLP classification framework, the trained models were integrated into a web-based decision support tool implemented using the Flask framework. The purpose of this tool is to allow users to upload real maritime investigation reports in PDF format, automatically extract and segment the text, identify causally relevant sentences, and

generate predicted SHIELD-NLP codes for expert review. The system supports a human-in-the-loop workflow, enabling both automated pre-classification and supervised validation of model outputs.

The application architecture consists of five main components:

1. Document ingestion and preprocessing
2. causal sentence extraction
3. automated SHIELD-NLP code prediction
4. interactive expert evaluation
5. persistent storage of validated classifications

5.7.1 Document Ingestion and Text Extraction

The tool allows users to upload investigation reports in PDF format via a web interface. Only PDF files are accepted to ensure consistency of document handling. Upon upload, the file is saved securely to a dedicated server directory using a sanitised filename. Text is then extracted automatically from all pages of the document using the PyMuPDF (fitz) library.

All page-level text is concatenated into a single clean text stream and stored in the user session shown in Figure 26. This ensures that the entire investigation narrative is available for downstream NLP processing without requiring manual page selection by the user.

This step enables the direct operational use of the system with authentic safety investigation reports, reflecting real-world deployment conditions.

Upload PDF for Analysis

Upload Your PDF

Choose a PDF File:

No file chosen

Figure 26. SHILED-NLP PDF upload interface.

5.7.2 Sentence Segmentation and Casual Sentence Selection

Once the full report text has been extracted, it is segmented into individual sentences using the NLTK sentence tokenizer. Given that full investigation reports contain large volumes of descriptive, legal, and administrative text that are not causally informative, an automated causal filtering mechanism is applied to prioritise analytically relevant content.

Each sentence is passed through a rule-based causal scoring function that evaluates the presence of:

- High-signal causal cues (e.g. “caused by”, “contributing factor”, “failure to”, “inadequate”)
- Medium-signal supporting cues (e.g. “due to”, “because”, “as a result”)
- Non-causal boilerplate indicators associated with legal, regulatory, or administrative content
- Additional heuristic detecting document numbers, legal references, extensive capitalisation, and abnormal sentence length.

To be retained, a sentence must contain at least one high-signal causal cue, ensuring that only sentences with explicit causal structure are forwarded to the machine learning stage. Sentences that are excessively short, excessively long, or dominated by legal front-matter are automatically rejected as non-causal boilerplate.

Sentences exceeding a minimum causal score threshold are retained directly. If too few sentences satisfy this threshold, the highest-scoring sentences are selected as a fallback to guarantee sufficient analytical coverage. This approach ensures a robust balance between precision (removing irrelevant text) and recall (retaining diverse causal statements).

5.7.3 Automated SHIELD-NLP Code Prediction

Each retained causal sentence then passed to the trained SHIELD-NLP classification model which are passed through a function. This function applied the full two-stage prediction pipeline and returns the top-K most probable SHIELD-NLP codes, together with their predicted probabilities and confidence status.

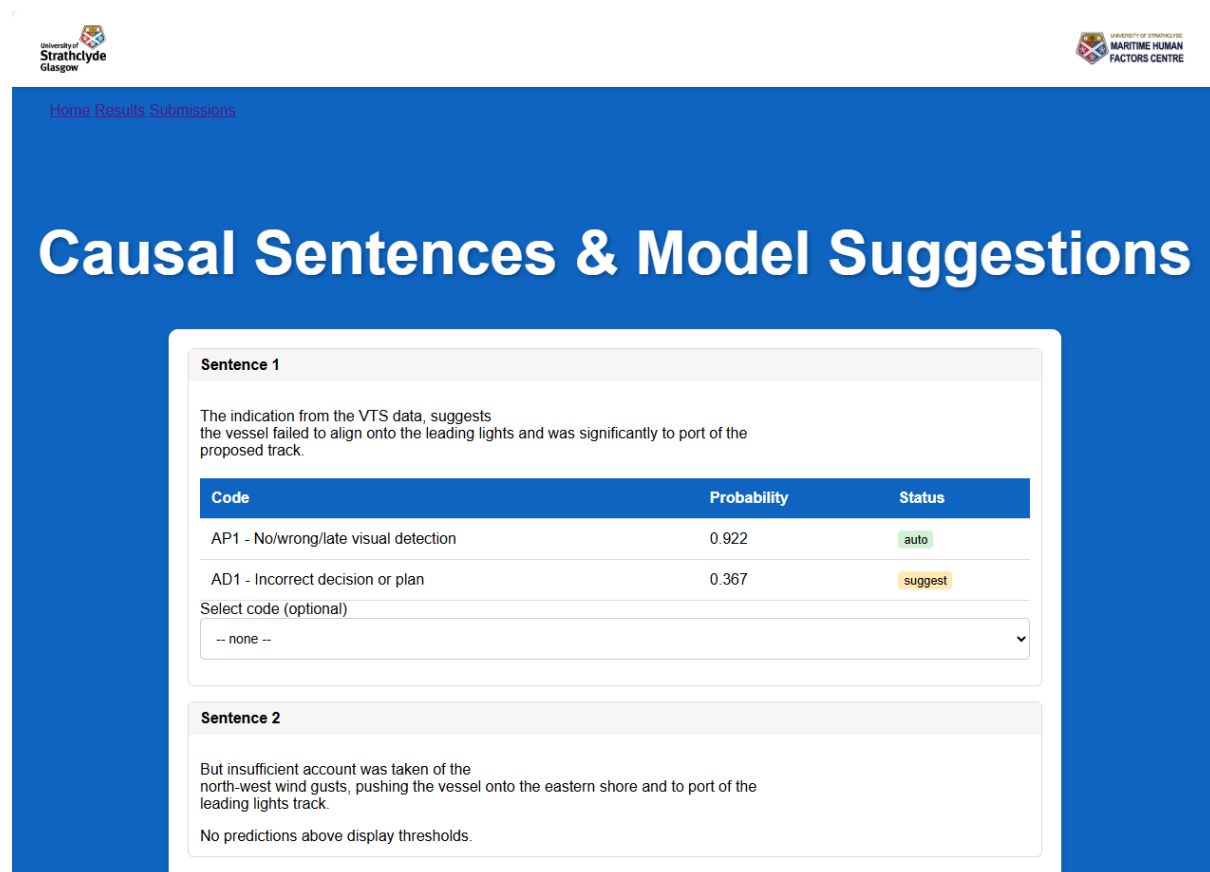
For each sentence, up to ten candidate causal codes are generated and filtered to retain only prediction assigned an “auto” or “suggest” confidence status. Predictions that are vetoed or assigned negligible probability are suppressed. The remaining predictions are ranked by descending probability and stored for user presentation.

This design enabled the system to function as a decision-support assistant rather than a fully autonomous classifier, ensuring that expert judgment remains central to the final classification.

5.7.4 Interactive Human-in-the-Loop Evaluation Interface

The filtered predictions are presented to the user through an interactive web interface shown in Figure 27. Each causally relevant sentence is displayed together with its ranked list of predicted SHIELD-NLP codes. The user is able to inspect the underlying sentence context, review the ranked code predictions, and select the most appropriate SHIELD-NLP label from the suggested set. This interface supports rapid

expert validation, allowing analysts to confirm correct predictions, correct misclassification, or override automated outputs entirely. This design reflects real operational requirements in safety analysis, where explainability and analysts control is essential.



The screenshot shows a web interface with a blue header and a white content area. The header contains the University of Strathclyde Glasgow logo on the left and the University of Strathclyde MARITIME HUMAN FACTORS CENTRE logo on the right. The main title is 'Causal Sentences & Model Suggestions'. Below the title, there are two sections, 'Sentence 1' and 'Sentence 2', each containing a text description and a table of model suggestions.

Sentence 1

The indication from the VTS data, suggests the vessel failed to align onto the leading lights and was significantly to port of the proposed track.

Code	Probability	Status
AP1 - No/wrong/late visual detection	0.922	auto
AD1 - Incorrect decision or plan	0.367	suggest

Select code (optional)

-- none --

Sentence 2

But insufficient account was taken of the north-west wind gusts, pushing the vessel onto the eastern shore and to port of the leading lights track.

No predictions above display thresholds.

Figure 27. Automatically extracted causal sentences from an investigation report.

5.7.5 Persistence and Database Storage of Expert-Validated Outputs

All user-confirmed classifications are stored in a structured SQLite database using SQLAlchemy. Each stored record includes the original sentence text, the user-selected SHIELD-NLP code, the model-assigned probability score, and a timestamp of the evaluation. This enables persistent storage of validated training examples and provides a mechanism for continuous dataset expansion. The stored outputs may be reused for future model retraining, auditability for automated classifications, and longitudinal analysis of classification behaviour.

A dedicated interface allows users to review all stored evaluations in chronological order, supporting transparency and traceability shown in Figure 28.

Evaluated Sentences

Sentence	Selected Prediction	Probability	Timestamp
The indication from the VTS data, suggests the vessel failed to align onto the leading lights and was significantly to port of the proposed track.	AP1 - No/wrong/late visual detection	92.2%	2025-12-07 19:02:04
The safety investigation concluded that the fire was most likely caused by fuel spilling accidentally onto a hot surface, near the entrance to the purifier room.	AD1 - Incorrect decision or plan	30.0%	2025-11-05 11:35:33
This resulted in a loss of electrical power, following which, the emergency generator automatically started ⁴ and supplied emergency electrical power to the emergency consumers.	PEW2 - Technology creates an unsafe situation	99.6%	2025-11-05 11:35:33

Figure 28. User validated and stored classification records.

5.7.6 Operational Significance of the Deployment

By embedding the SHIELD-NLP models into a fully functional web application, this work demonstrates the operational feasibility of automated human-factors classification within real investigative workflows. The system transforms unstructured investigation reports directly into structured, machine-interpretable SHIELD-NLP classification, while preserving expert oversight through a human-in-the-loop design.

The deployment validates that the high performance observed in the experimental evaluation can be translated into a practical, usable decision-support tool for safety analysts. This represents a critical step beyond offline model evaluation and demonstrates the system's readiness for real-world integration within maritime safety analysis organisations.

5.8 Association Rule Mining for Human Factors Interaction Analysis

5.8.1 Overall Database

The results from the association rule mining applied on the full dataset found 11 associations, shown in Table 12 and Figure 29. From these we were able to gain the following insights:

1. Workspace design and communication

There was an association found between workspace or working position incompatible with operation and no transmission of information. The results indicated that 61% of the time when there are workspace issues there will be an issue with information transmission. This means that poor ergonomic or physical workspace layout significantly increase the likelihood of communication failures, suggesting that physical design should be considered an integral component of safety management rather than solely an engineering concern. In practical terms, vessel designers and operators should evaluate bridge and workspace layouts to minimise visual obstructions, excessive noise, and physical separation between crew members. Human-centred design approaches, such as ergonomic bridge assessments and communication task analyses, could therefore be incorporated into vessel design reviews and safety audits to reduce the likelihood of information transmission failures.

2. Environmental & Personal Factors Drive Poor Decisions

Several personal/environmental antecedents are associated with incorrect decision or plan. Specifically, the following: vision affected, bad weather operations, unsafe technology, low confidence, risk underestimation, inadequate experience/training, misperception of changing environment. However, those with the highest confidence were misperception of changing environment (0.87), risk underestimation (0.80) and inadequate training or currency (0.72). This suggests that dynamic situational awareness represents a critical vulnerability in maritime operations. This finding indicates that traditional classroom-based training may be insufficient for preparing operators to respond effectively to rapidly evolving operational environments. Instead, greater emphasis should be placed on scenario-based simulation training that exposes crews to deteriorating weather, traffic congestion, equipment degradation, and other dynamic operating conditions. In addition, decision-support systems capable of integrating environmental,

navigational, and operational information could assist crews in maintaining situational awareness and reducing the likelihood of erroneous decisions.

3. Organisational procedure weakness leads to leadership gaps

When both safety risk management (proactive) and publications/procedures/written guidance are weak or unclear in an organisation it was found to frequently correlate with poor leadership/supervision in operations. It was found that when both of these organisational issues are apparent then 68% of the time there will be leadership or supervision issues. This shows a top-down weakness in documentation and systems can manifest in frontline command issues. This shows that top-down weaknesses in documentation and safety systems can manifest in frontline command issues. Importantly, this suggests that deficiencies in leadership and supervision may often represent symptoms of wider organisational shortcomings rather than isolated individual failings. In practice, supervisors rely on clear procedures and proactive risk management processes to support decision-making and maintain operational oversight. Consequently, shipping companies should routinely review the usability and operational relevance of procedures, ensure that safety management systems remain proactive rather than compliance-driven, and provide supervisors with adequate leadership training and decision-support resources.

Causal Association Rule Graph (All Accidents)

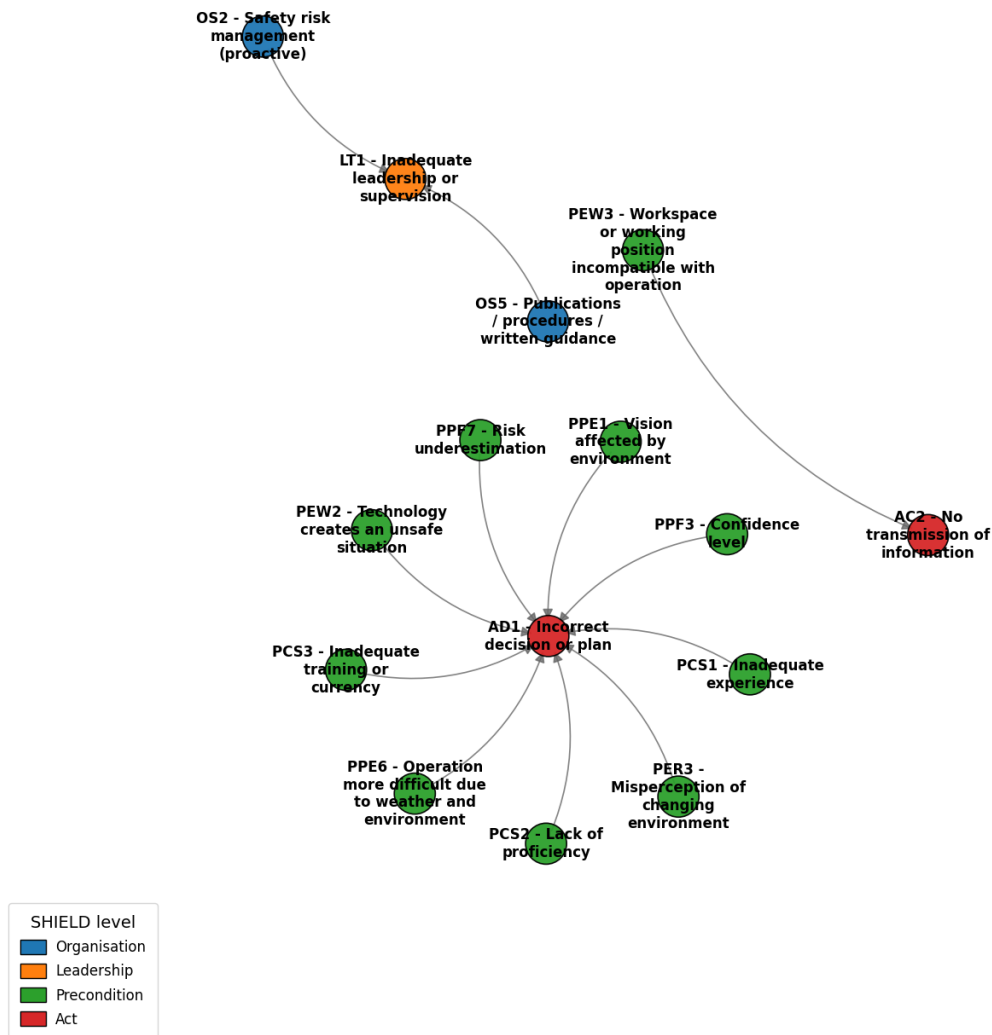


Figure 29. ARM network of contributory factors (full dataset).

Table 12. Full data association rule mining results.

antecedents	consequents	support	confidence	lift
['PEW3 - Workspace or working position incompatible with operation']	['AC2 - No transmission of information']	0.051	0.613	1.763
['PPE1 - Vision affected by environment']	['AD1 - Incorrect decision or plan']	0.070	0.619	1.196
['PPE6 - Operation more difficult due to weather and environment']	['AD1 - Incorrect decision or plan']	0.065	0.615	1.189
['PEW2 - Technology creates an unsafe situation']	['AD1 - Incorrect decision or plan']	0.089	0.600	1.159

['PPF3 - Confidence level']	['AD1 - Incorrect decision or plan']	0.054	0.800	1.546
['PPF7 - Risk underestimation']	['AD1 - Incorrect decision or plan']	0.078	0.763	1.475
['PCS1 - Inadequate experience']	['AD1 - Incorrect decision or plan']	0.111	0.672	1.299
['PCS2 - Lack of proficiency']	['AD1 - Incorrect decision or plan']	0.097	0.706	1.364
['PCS3 - Inadequate training or currency']	['AD1 - Incorrect decision or plan']	0.057	0.724	1.399
['PER3 - Misperception of changing environment']	['AD1 - Incorrect decision or plan']	0.054	0.870	1.680
['OS5 - Publications / procedures / written guidance', 'OS2 - Safety risk management (proactive)']	['LT1 - Inadequate leadership or supervision']	0.051	0.679	2.289

5.8.2 Accident Type Analysis

We then look at how these associations change when we look at specific accident types. Here we will compare the rules for collision, grounding, contact, fire and occupational accidents.

5.8.2.1 Collision Accidents

The analysis of collision-related data using association rule mining reveals a series of informative patterns that reflect the interaction of perceptual, environmental, cognitive, and organisational factors in the occurrence of these incidents. An overview of these results are shown in Table 13 and Figure 30

Causal Association Rule Graph (Collision Accidents)

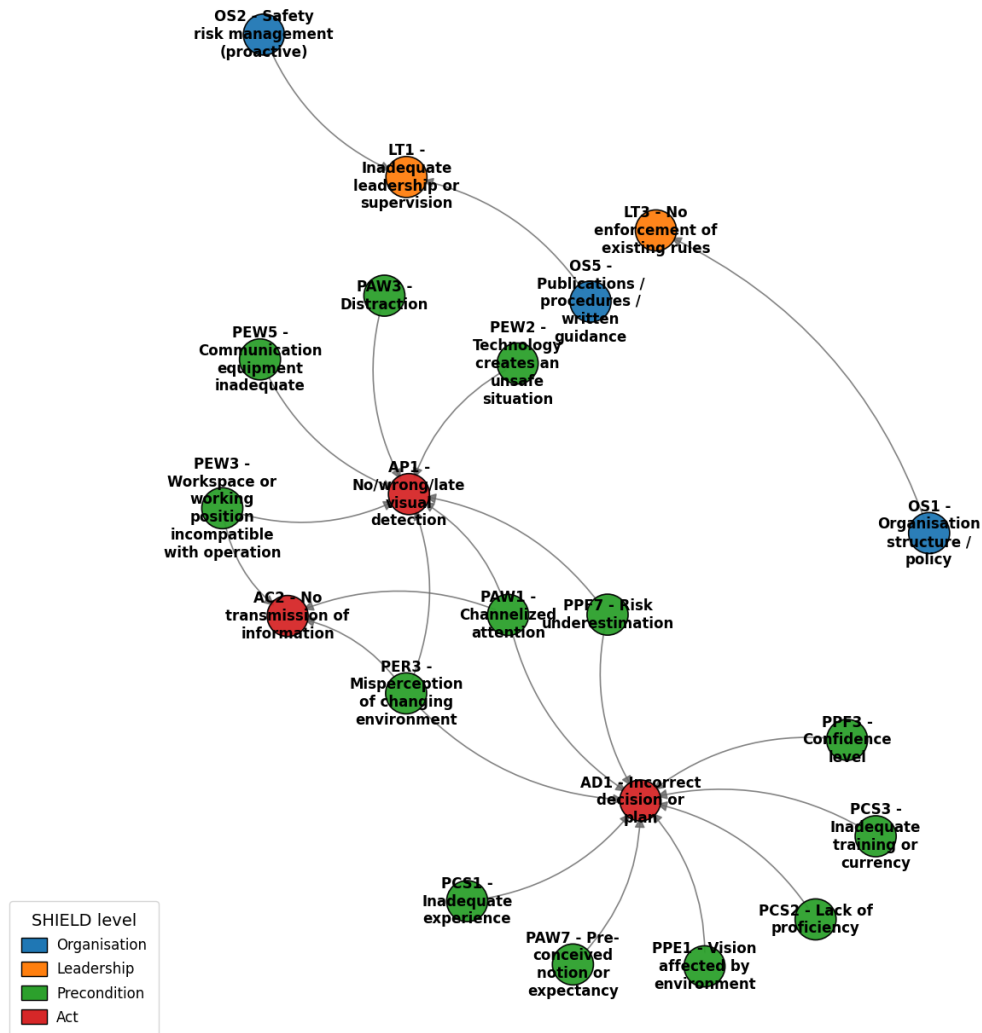


Figure 30. ARM network of contributory factors for collision incidents.

Table 13. Collision incidents association rule mining results.

antecedents	consequents	support	confidence	lift
['PEW2 - Technology creates an unsafe situation']	['AP1 - No/wrong/late visual detection']	0.10	0.68	1.46
['PEW3 - Workspace or working position incompatible with operation']	['AP1 - No/wrong/late visual detection']	0.07	1.00	2.14
['PEW5 - Communication equipment inadequate']	['AP1 - No/wrong/late visual detection']	0.06	0.88	1.87
['PAW1 - Channelized attention']	['AP1 - No/wrong/late visual detection']	0.06	0.64	1.36
['PAW3 - Distraction']	['AP1 - No/wrong/late visual detection']	0.11	0.61	1.30

['PPF7 - Risk underestimation']	['AP1 - No/wrong/late visual detection']	0.07	0.60	1.28
['PER3 - Misperception of changing environment']	['AP1 - No/wrong/late visual detection']	0.07	0.64	1.37
['PEW3 - Workspace or working position incompatible with operation']	['AC2 - No transmission of information']	0.06	0.78	1.79
['PAW1 - Channelized attention']	['AC2 - No transmission of information']	0.06	0.73	1.67
['PER3 - Misperception of changing environment']	['AC2 - No transmission of information']	0.07	0.64	1.48
['PPE1 - Vision affected by environment']	['AD1 - Incorrect decision or plan']	0.14	0.61	0.94
['PAW1 - Channelized attention']	['AD1 - Incorrect decision or plan']	0.06	0.64	0.99
['PAW7 - Pre-conceived notion or expectancy']	['AD1 - Incorrect decision or plan']	0.06	0.89	1.38
['PPF3 - Confidence level']	['AD1 - Incorrect decision or plan']	0.07	0.69	1.07
['PPF7 - Risk underestimation']	['AD1 - Incorrect decision or plan']	0.11	0.93	1.45
['PCS1 - Inadequate experience']	['AD1 - Incorrect decision or plan']	0.18	0.69	1.07
['PCS2 - Lack of proficiency']	['AD1 - Incorrect decision or plan']	0.13	0.80	1.24
['PCS3 - Inadequate training or currency']	['AD1 - Incorrect decision or plan']	0.06	0.67	1.03
['PER3 - Misperception of changing environment']	['AD1 - Incorrect decision or plan']	0.10	0.93	1.44
['OS1 - Organisation structure / policy']	['LT3 - No enforcement of existing rules']	0.08	0.77	2.58
['PEW3 - Workspace or working position incompatible with operation']	['AP1 - No/wrong/late visual detection', 'AC2 - No transmission of information']	0.06	0.78	3.21
['PPF7 - Risk underestimation']	['AP1 - No/wrong/late visual detection', 'AD1 - Incorrect decision or plan']	0.07	0.60	2.07
['PER3 - Misperception of changing environment']	['AP1 - No/wrong/late visual detection', 'AD1 - Incorrect decision or plan']	0.07	0.64	2.21
['PER3 - Misperception of changing environment']	['AC2 - No transmission of information', 'AD1 - Incorrect decision or plan']	0.07	0.64	2.10
['PPE1 - Vision affected by environment', 'PCS1 - Inadequate experience']	['AD1 - Incorrect decision or plan']	0.06	0.67	1.03
['OS2 - Safety risk management (proactive)', 'OS5 - Publications / procedures / written guidance']	['LT1 - Inadequate leadership or supervision']	0.06	0.78	2.19

A dominant theme is the relationship between visual perception failures and a variety of precursors. The consequence no/wrong/late visual detection (AP1) is frequently associated with antecedents such as:

- Technology created an unsafe situation (PEW2)
- Workspace or working position incompatible with operation (PEW3)
- Communication equipment inadequate (PEW5)
- Channelised attention (PAW1)

Of particular note PEW3 leading to AP1, with a confidence of 1 and lift of 2.14 suggests that poor workspace ergonomics almost invariably coincides with visual detection failures. Similarly, the high lift and confidence of PEW5 leading to AP1 underlines how communication equipment inadequacy severely hampers the visual awareness critical in navigation contexts. The deterministic relationship between incompatible workspace design and visual detection failures highlights the importance of bridge ergonomics within collision prevention. Vessel operators should therefore undertake periodic ergonomic reviews focusing on sightlines, equipment placement, alarm positioning, and communication pathways. Future vessel design standards may also benefit from incorporating formal human-centred design principles to minimise perceptual failures.

There is also strong evidence linking attentional and perceptual factors to failure in information transmission (AC2). Antecedents like channelised attention (PAW1) and misperception of changing environment (PER3) are frequently followed by AC2, demonstrating that focused attention or misinterpreted environmental cues reduced the flow of critical information during operations. This finding suggests that communication breakdowns in collision scenarios are often secondary consequences of degraded situational awareness rather than isolated communication failures. When bridge personnel become overly focused on a single vessel, navigational hazard, or operational task, their ability to maintain a shared understanding of the wider traffic situation may be compromised. Consequently, collision prevention strategies should not only emphasise communication procedures,

but also practices that support collective situational awareness, including closed-loop communication, challenge-and-response behaviours, and bridge resource management techniques designed to maintain shared mental models.

In terms of decision-making failures, the rule risk underestimation (PPF7) leads to incorrect decision or plan (AD1) has the highest confidence (0.93) and lift of 1.45, indicating that underestimating risks almost always leads to incorrect planning in these scenarios. This pattern is also consistent with other cognitive precursors such as confidence level (PPF3), pre-conceived notions (PAW7), and inadequate experience (PCS1) all feeding into AD1. Collectively, these findings suggest that collision accidents are frequently preceded by failures in dynamic risk assessment rather than deficiencies in ship-handling ability alone. Familiarity with routes, routine operations, or previous successful outcomes may lead operators to underestimate emerging hazards or delay corrective actions. This highlights the importance of training situations and encouraging continual reassessment of navigational risk rather than reliance on statistical assumptions or previous experience.

Importantly, the presence of multiple antecedents leading to the same consequence further strengthens the argument for systemic cognitive vulnerabilities. The rule PER3 leading to both AP1 and AD1 and similar compound consequences show the multi-faceted rule of misperception, resulting in both perceptual failure and poor decisions.

At the organisational level, two rules stand out:

- Organisation structure/policy (OS1) leads to no enforcement of existing rules (LT3) with a lift of 2.57, and
- Publications/procedures/written guidance (OS5) and safety risk management (OS2) leads to inadequate leadership or supervision (LT1) with a confidence of 0.78 and lift of 2.19.

These suggest that inadequate documentation, procedural clarity, and proactive risk management are highly predictive of supervisory failures, further affirming the layered nature of causal factors in

collisions. Importantly, these findings suggest that supervisory failures identified during collision investigations may often represent symptoms of wider organisational weaknesses rather than isolated individual shortcomings. Deficiencies in procedures and proactive risk management may reduce supervisors' ability to effectively monitor operations, challenge unsafe behaviours, and ensure compliance with navigational procedures. Consequently, organisations should routinely assess the operational usability of procedures and ensure that safety management systems actively support frontline decision-making rather than functioning solely as compliance mechanisms.

Lastly, rules with multiple consequents such as misperception or working position incompatible with operation (PEW3) leads to both no transmission of information (AC2) and no/wrong/late visual detection (PA1) reinforce the notion of “divergent causality”, where a single unsafe condition can propagate into several downstream effects, both in perception and communication breakdowns. This complexity should be addressed by multimodal interventions that combine technical, procedural, and behavioural components to improve maritime operational safety. The identification of these divergent causal pathways further suggests that certain contributory factors may function as high-level intervention points within the maritime system. For example, improvements in bridge ergonomics or situational awareness practices may simultaneously reduce both perceptual failures and communication breakdowns. Prioritising interventions that target these upstream contributors may therefore provide greater safety benefits than addressing downstream failures individually.

5.8.2.2 Grounding

The association rule mining for grounding incidents uncovered several meaningful patterns that reflect both human and organisational contributions to risk, shown in Table 14 and Figure 31. A prominent theme was the impact of cognitive and attentional failures on grounding-related decision-making errors. For instance, confusion (PAW2), risk underestimation (PPF7), and high workload (PMW1) were significantly associated with no transmission of information (AC2), a critical communication lapse. These rules had confidence scores of 0.66 and lift values exceeding 1.84, indicating that these mental

state conditions strongly co-occur with breakdowns in information flow. This finding suggests that communication failures during grounding events may often arise because operators become cognitively overloaded or uncertain about the evolving navigational situation. Under conditions of high workload or confusion, crews may prioritise immediate navigational tasks at the expense of information sharing, thereby compromising shared situational awareness within the bridge team. Consequently, grounding prevention strategies should place greater emphasis on workload management, effective bridge team coordination, and structured communication practice during periods of elevated operational complexity.

Causal Association Rule Graph (Grounding Accidents)

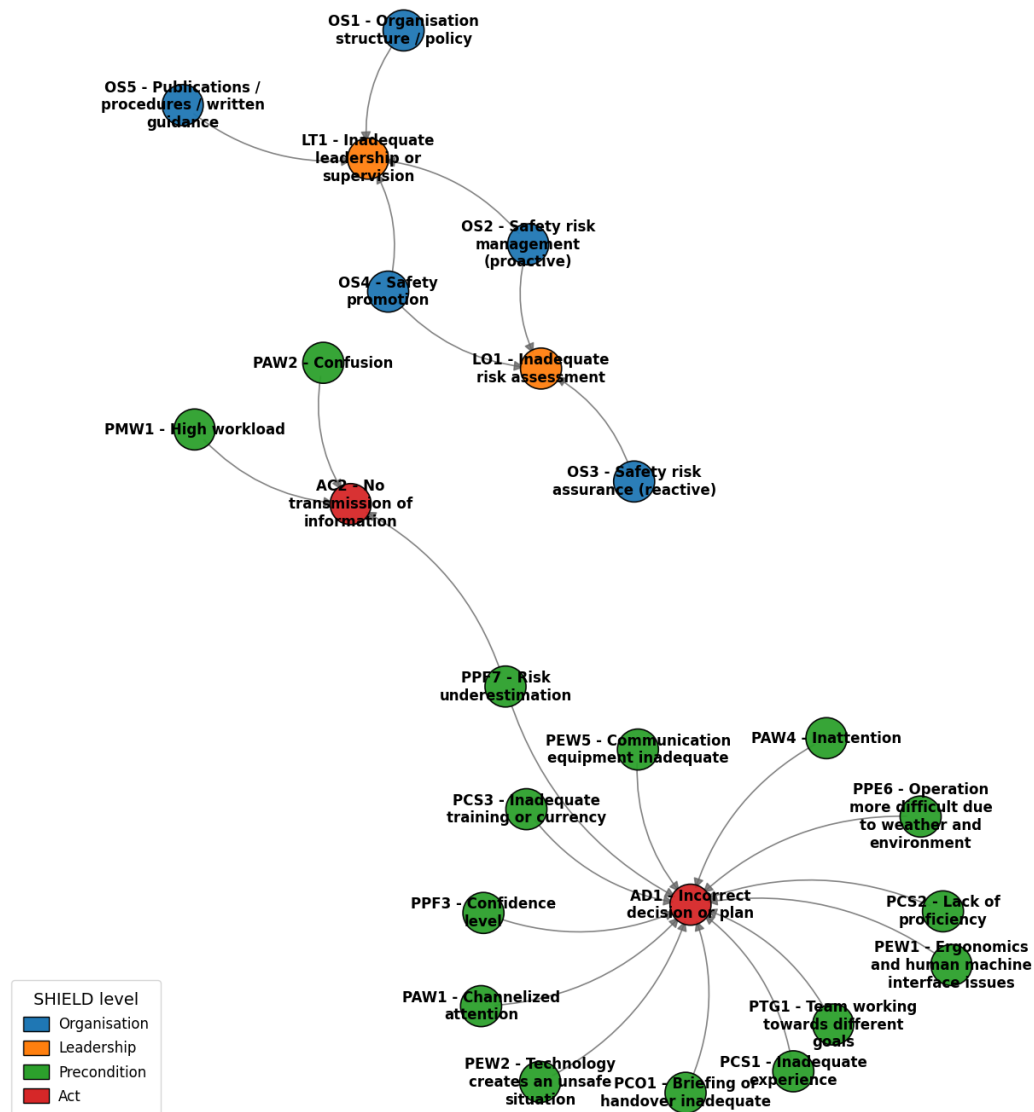


Figure 31. ARM network of contributory factors for grounding incidents.

Table 14. Grounding incidents association rule mining results.

antecedents	consequents	support	confidence	lift
['PAW2 - Confusion']	['AC2 - No transmission of information']	0.07	0.67	1.85
['PMW1 - High workload']	['AC2 - No transmission of information']	0.07	0.64	1.78
['PPF7 - Risk underestimation']	['AC2 - No transmission of information']	0.07	0.67	1.85
['PPE6 - Operation more difficult due to weather and environment']	['AD1 - Incorrect decision or plan']	0.12	0.65	1.06
['PEW1 - Ergonomics and human machine interface issues']	['AD1 - Incorrect decision or plan']	0.10	0.60	0.98

['PEW2 - Technology creates an unsafe situation']	['AD1 - Incorrect decision or plan']	0.08	0.67	1.08
['PEW5 - Communication equipment inadequate']	['AD1 - Incorrect decision or plan']	0.06	0.70	1.14
['PCO1 - Briefing or handover inadequate']	['AD1 - Incorrect decision or plan']	0.07	0.75	1.22
['PTG1 - Team working towards different goals']	['AD1 - Incorrect decision or plan']	0.07	0.75	1.22
['PAW1 - Channelized attention']	['AD1 - Incorrect decision or plan']	0.07	0.82	1.33
['PAW4 - Inattention']	['AD1 - Incorrect decision or plan']	0.07	0.60	0.98
['PPF3 - Confidence level']	['AD1 - Incorrect decision or plan']	0.08	0.91	1.48
['PPF7 - Risk underestimation']	['AD1 - Incorrect decision or plan']	0.08	0.83	1.36
['PCS1 - Inadequate experience']	['AD1 - Incorrect decision or plan']	0.11	0.67	1.08
['PCS2 - Lack of proficiency']	['AD1 - Incorrect decision or plan']	0.14	0.68	1.11
['PCS3 - Inadequate training or currency']	['AD1 - Incorrect decision or plan']	0.07	0.82	1.33
['OS4 - Safety promotion']	['LO1 - Inadequate risk assessment']	0.09	0.65	1.75
['PCS2 - Lack of proficiency', 'PCS3 - Inadequate training or currency']	['AD1 - Incorrect decision or plan']	0.06	1.00	1.63
['OS4 - Safety promotion', 'OS2 - Safety risk management (proactive)']	['LO1 - Inadequate risk assessment']	0.07	0.73	1.97
['OS3 - Safety risk assurance (reactive)', 'OS4 - Safety promotion']	['LO1 - Inadequate risk assessment']	0.06	0.88	2.37
['OS1 - Organisation structure / policy', 'OS5 - Publications / procedures / written guidance']	['LT1 - Inadequate leadership or supervision']	0.07	0.73	1.74
['OS4 - Safety promotion', 'OS2 - Safety risk management (proactive)']	['LT1 - Inadequate leadership or supervision']	0.07	0.73	1.74
['OS5 - Publications / procedures / written guidance', 'OS2 - Safety risk management (proactive)']	['LT1 - Inadequate leadership or supervision']	0.08	0.63	1.50
['OS4 - Safety promotion', 'OS5 - Publications / procedures / written guidance']	['LT1 - Inadequate leadership or supervision']	0.06	0.78	1.86
['OS4 - Safety promotion', 'OS5 - Publications / procedures / written guidance', 'OS2 - Safety risk management (proactive)']	['LT1 - Inadequate leadership or supervision']	0.06	0.88	2.09

Furthermore, factors such as operation more difficulty due to weather (PPE6) and technology creates an unsafe situation (PEW2) consistently appeared as antecedent to incorrect decision or plan (AD1).

Though the lift values for these rules were modest, the high support and confidence suggest these are recurring, reliable patterns in the dataset. This supports the view that environmental complexity and system usability issues directly impair effective decision-making in navigation. The prominence of weather-related and technology-related factors suggests that grounding accidents frequently occur when operators are required to interpret and respond to multiple, often competing, sources of information. This indicates that grounding prevention is not solely dependent on navigational competence, but also on the ability of crews to effectively manage complexity and uncertainty. From an operational perspective, decision-support systems, improved human-machine interfaces, and simulator exercises exposing crews to degraded environmental and equipment conditions may therefore provide important safety benefits.

Human factors related to inexperience also showed strong predictive power. Antecedents like inadequate experience (PCS1), lack of proficiency (PCS2) and inadequate training (PCS3) were associated with incorrect planning or decision. Notably, the rule combining PCS3 and PCS3 leading to AD1 had 100% confidence and a lift of 1.63, suggesting a nearly deterministic relationship in this context. Such patterns indicate that multiple compounding deficiencies in competence significantly raise the likelihood of grounding incidents. Importantly, the strong relationships between experience, proficiency, training, and decision-making errors suggest that grounding accidents are particularly sensitive to deficiencies in competency management. The findings imply that competence should not be viewed solely in terms of certification or sea time, but rather as the ability to apply knowledge effectively under dynamic operational conditions. Organisations should therefore supplement traditional training approaches with scenario-based exercises, mentoring programmes, and periodic competency assessments designed to evaluate decision-making performance under realistic navigational conditions.

The analysis also highlighted organisational factors, particularly relating to risk management and leadership. For example, safety promotion (OS4) and safety risk management (OS2) frequently led to

inadequate risk assessment (LO1). The three-way combination (OS4, OS5 and OS2) leading to inadequate leadership or supervision (LT1) stood out with a confidence of 0.875, lift of 2.09, and conviction of 4.66, indicating a highly robust relationship. These patterns illustrate how systemic management practices, including quality of procedures, promotion of safety culture, and proactive risk planning, directly influence supervisory capacity and frontline leadership effectiveness. These findings further suggest that weaknesses in safety promotion and proactive risk management may act as leading indicators of future grounding risk. Organisations should therefore routinely evaluate the effectiveness of safety communication, hazard reporting systems, and risk assessment processes to ensure that emerging navigational hazards are identified and mitigated before they contribute to operational failures.

Finally, the strength of multi-factor rules involving both individual capability and organisational policy underscores the complex, layered nature of risk in grounding incidents. The interplay between inattentional antecedents, procedural weakness, and inadequate team coordination reflects the multifactorial conditions often preceding maritime errors. These findings suggest a need for holistic safety interventions that integrate training, procedural, reform, and organisational safety culture enhancements. The findings indicate that grounding prevention requires interventions across multiple system levels. Improvements in individual competence alone are unlikely to be sufficient unless accompanied by effective bridge coordination, usable procedures, proactive risk management practices, and organisational mechanisms that support continuous learning. This reinforces the view that grounding accidents emerge from interactions between human, technological, and organisational factors rather than isolated failures at a single system level.

5.8.2.3 Fire

The application of association rule mining to fire-related incidents revealed distinct patterns in the way multiple latent conditions interact to produce unsafe outcomes, shown in Table 15 and Figure 32. To enhance interpretability, only representative rules from each dominant causal pathway were retained

in the summarisation table. Notably, several rules highlighted the significant influence of cognitive factors, such as forgetfulness and inaccurate recall, which frequently preceded timing errors and incorrect decisions or plans. For example, the rule forget actions/intentions leads to timing error has a confidence of 0.67 and a lift of 4.17, demonstrating a strong probabilistic relationship. When these cognitive factors were combined (forget actions/intentions and no/inaccurate information recall) leading to timing errors, confidence increased to 0.75, further underscoring the compounded risk posed by simultaneous memory-related deficiencies. These findings suggest that fire scenarios place substantial cognitive demands on operators, particularly when rapid decision-making and action execution are required under stressful conditions. The prominence of memory-related failures indicates that reliance on recall alone may be inappropriate during emergency situations. Consequently, emergency procedures should be designed to minimise memory dependence through the use of clear checklists, decision aids, and repeated realistic drills that reinforce critical actions under time pressure.

Causal Association Rule Graph (Fire Accidents)

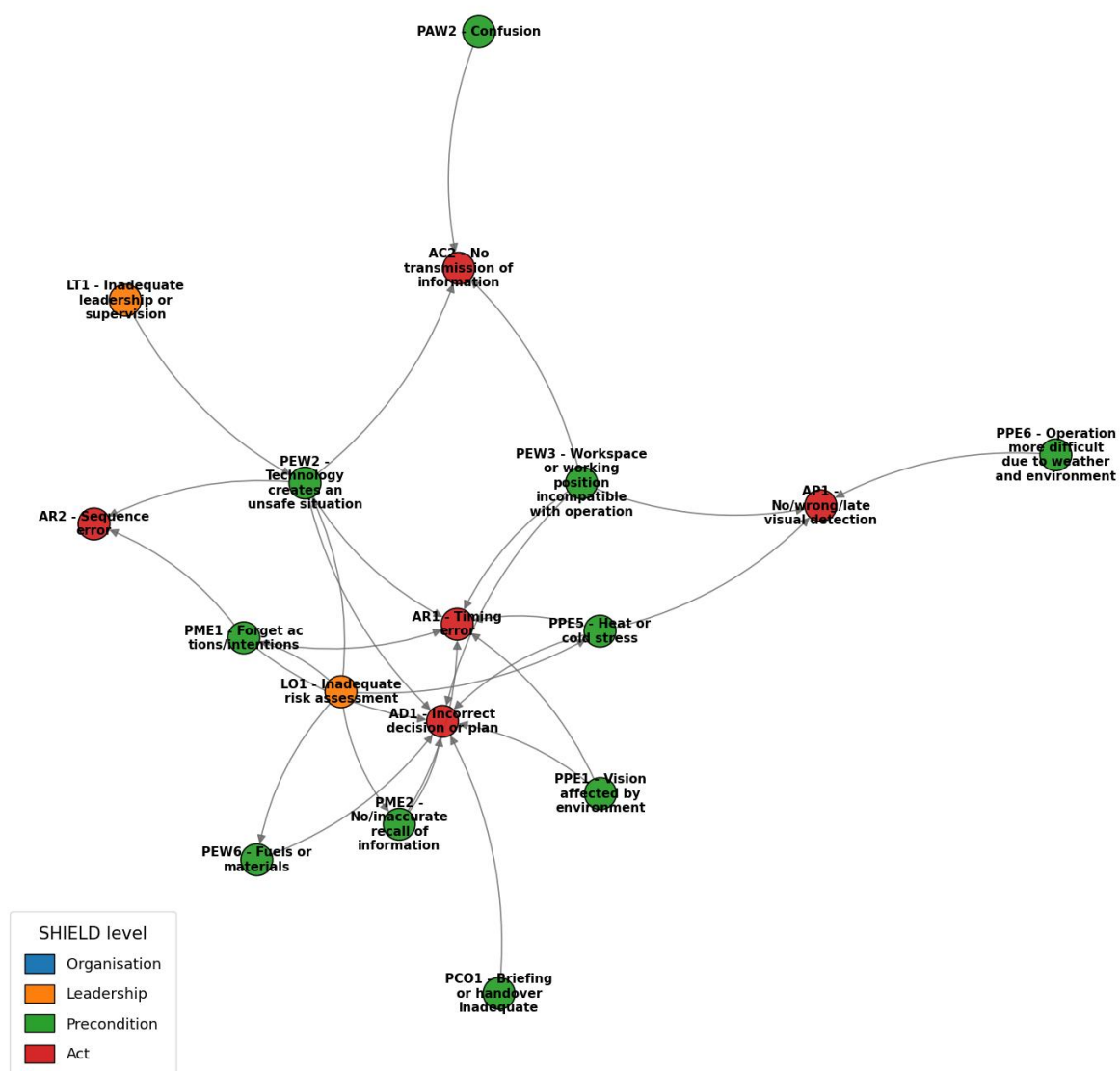


Figure 32. ARM network of contributory factors in fire incidents.

Table 15. Fire incidents association rule mining results.

antecedents	consequents	support	confidence	lift
['PME1 - Forget actions/intentions']	['AR1 - Timing error']	0.08	0.67	4.17
['PME1 - Forget actions/intentions']	['AD1 - Incorrect decision or plan']	0.1	0.83	1.89
['PME2 - No/inaccurate recall of information']	['AD1 - Incorrect decision or plan']	0.1	0.71	1.62
['LO1 - Inadequate risk assessment']	['PPE5 - Heat or cold stress']	0.08	0.8	3.33
['LT1 - Inadequate leadership or supervision']	['PEW2 - Technology creates an unsafe situation']	0.08	0.67	1.85

['PPE5 - Heat or cold stress', 'PEW3 - Workspace or working position incompatible with operation']	['AR1 - Timing error']	0.08	0.8	5
['PPE5 - Heat or cold stress', 'PME2 - No/inaccurate recall of information']	['AR1 - Timing error']	0.08	0.67	4.17
['PME2 - No/inaccurate recall of information', 'PEW3 - Workspace or working position incompatible with operation']	['AR1 - Timing error']	0.08	0.8	5
['PEW2 - Technology creates an unsafe situation', 'PEW3 - Workspace or working position incompatible with operation']	['AC2 - No transmission of information']	0.1	0.63	1.36
['PEW2 - Technology creates an unsafe situation', 'PPE5 - Heat or cold stress']	['AD1 - Incorrect decision or plan']	0.12	0.6	1.36
['PPE5 - Heat or cold stress', 'PEW3 - Workspace or working position incompatible with operation']	['AD1 - Incorrect decision or plan']	0.08	0.8	1.82
['PPE5 - Heat or cold stress', 'PME2 - No/inaccurate recall of information']	['AD1 - Incorrect decision or plan']	0.1	0.83	1.89
['PEW2 - Technology creates an unsafe situation', 'PME1 - Forget actions/intentions']	['AD1 - Incorrect decision or plan']	0.1	1	2.27
['PEW2 - Technology creates an unsafe situation', 'PME2 - No/inaccurate recall of information']	['AD1 - Incorrect decision or plan']	0.1	0.83	1.89
['PME1 - Forget actions/intentions', 'PEW6 - Fuels or materials']	['AD1 - Incorrect decision or plan']	0.08	1	2.27
['PPE5 - Heat or cold stress', 'PME2 - No/inaccurate recall of information', 'PEW3 - Workspace or working position incompatible with operation']	['AR1 - Timing error']	0.08	1	6.25
['PEW2 - Technology creates an unsafe situation', 'PME2 - No/inaccurate recall of information', 'PPE5 - Heat or cold stress']	['AD1 - Incorrect decision or plan']	0.1	1	2.27
['PEW2 - Technology creates an unsafe situation', 'PME1 - Forget actions/intentions', 'PEW6 - Fuels or materials']	['AD1 - Incorrect decision or plan']	0.08	1	2.27

Environmental and ergonomic stressors were also shown to significantly impact fire-related risks.

Factors like operation more difficult due to weather (PPE6) and heat or cold stress (PPE5) appeared

frequently in high-confidence rules. For instance, PPE6's association with no/wrong/late visual decision (AP1) had a confidence of 1 and a lift of 5.56, suggesting a deterministic relationship in many cases. Such findings point to the critical role of external conditions, such as temperature, visibility, and weather, which may overwhelm an individual's ability to perceive and respond effectively, particularly in high-risk fire situations. The findings further suggest that emergency response effectiveness may deteriorate significantly when physiological and environmental stressors occur simultaneously. Reduced visibility, heat exposure, and degraded working conditions may impair both perception and decision-making, limiting the ability of personnel to respond effectively. Organisations should therefore ensure that emergency training exercises are conducted under realistic, degraded conditions, including low visibility, high workload, and simulated environmental stressors, to improve operational resilience during fire events.

Additionally, a number of rules revealed interdependencies between organisational shortcomings and downstream safety failures. Inadequate risk assessment (LO1) was a recurring antecedent in rules predicting both environmental and cognitive issues. For example, LO1 leads to technology creating unsafe situations and head or cold stress has a lift of 4.29 and a confidence of 0.6, indicating that poor risk planning often precedes both technical and physiological difficulties. These results support the view that upstream organisational weaknesses can initiate a cascade of failures that compromise human and technical systems alike. Importantly, these results indicate that deficiencies in risk assessment may act as upstream drivers of both technical and human failures. Inadequate hazard identification and risk planning may leave organisations insufficiently prepared for foreseeable fire scenarios, increasing exposure to environmental stressors and equipment-related issues during emergency response. This highlights the importance of robust fire risk assessment, scenario-based emergency planning, and periodic review of emergency preparedness arrangements.

The rules also emphasised the multi-consequence nature of fire incidents, where antecedents frequently led to more than one failure type. Several rules involved compound consequents, such as

PPE5, Workspace or working position incompatible with operation (PEW3) leads to both timing errors (AR1) and incorrect decision-making (AD1), with confidence and lift values indicating near-deterministic relationships. These rules illustrate the interconnectedness of decision-making and action execution under fire conditions – failures are rarely isolated, and timing and planning errors tend to co-occur. This duality suggests a need for integrated risk assessment approaches that consider both cognitive and operational vulnerabilities in tandem.

Finally, combinations of three or more antecedents demonstrated the highest predictive power, often with lift values exceeding 6 and perfect confidence scores of 1. For example, PPE5+PME2+PEW3+PEW2 leading to AR1 yielded a lift of 6.25 and confidence of 1. These multivariate rules reveal that certain configurations of latent conditions almost invariably lead to error. This further reinforces the concept of latent risk accumulation, when enough stressors are present concurrently, safety barriers are overwhelmed, and failure becomes highly probable. The prevalence of high-confidence multi-antecedent rules suggests that fire accidents rarely result from single-point failures. Instead, they emerge through the accumulation and interaction of multiple latent conditions spanning human, technical, environmental, and organisational domains. From a safety management perspective, this finding emphasises the need for integrated barrier management approaches capable of identifying and controlling combinations of risk factors before they align. Monitoring individual hazards in isolation may therefore be insufficient, as safety degradation may only become apparent when several seemingly minor deficiencies occur concurrently.

5.8.2.4 Occupational

The association rule mining applied to occupational incidents highlights key relationships between behavioural, procedural, and organisational shortcomings and the occurrence of workarounds, decision-making failures, and communication breakdowns, shown in Table 16 and Figure 33.

Causal Association Rule Graph (Occupational Accidents)

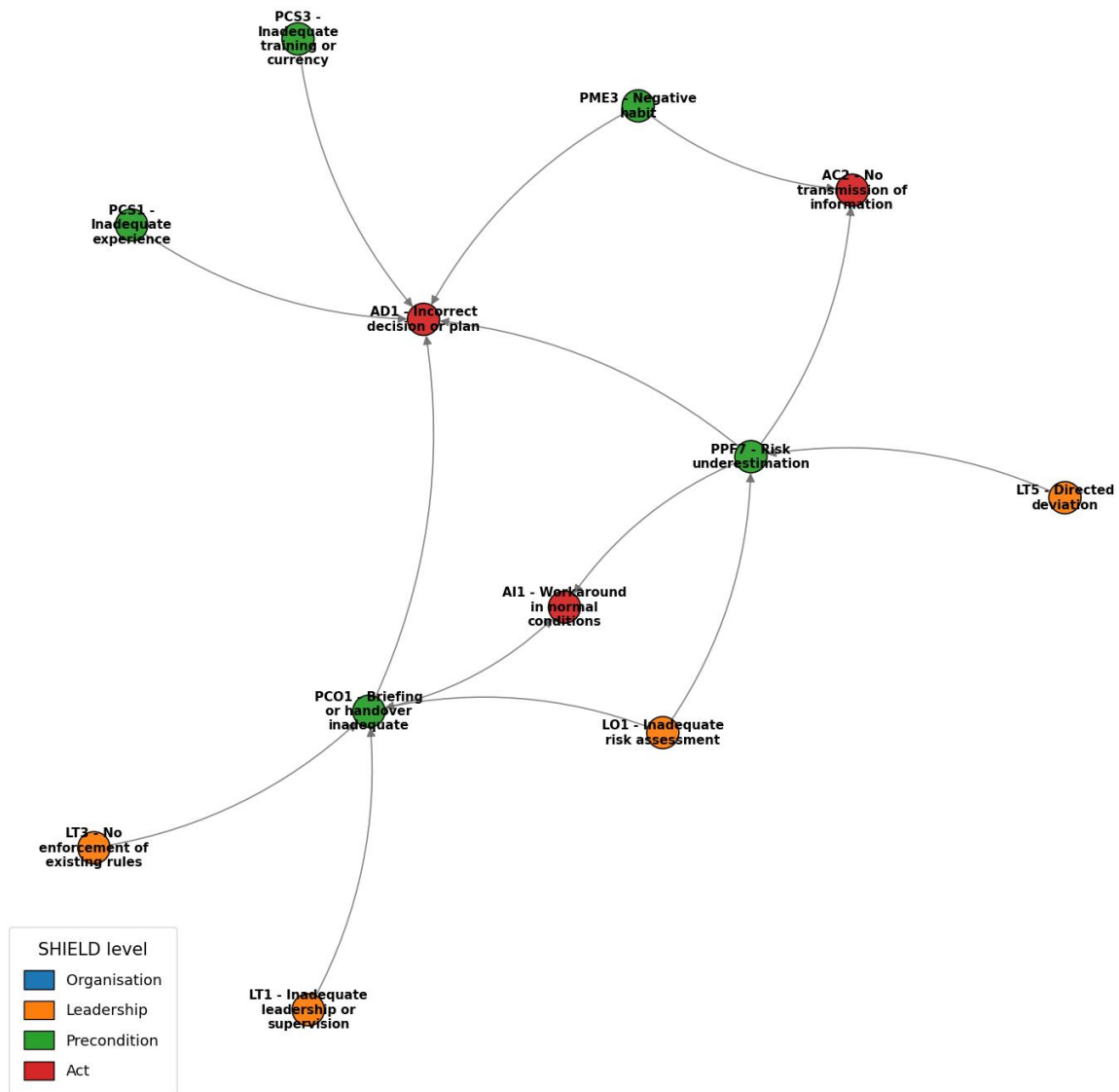


Figure 33. ARM network of contributory factors in occupational incidents.

Table 16. Occupational incidents association rule mining results.

antecedents	consequents	support	confidence	lift
['PEW3 - Workspace or working position incompatible with operation']	['AC2 - No transmission of information']	0.05	0.61	1.76
['PPE1 - Vision affected by environment']	['AD1 - Incorrect decision or plan']	0.07	0.62	1.20
['PPE6 - Operation more difficult due to weather and environment']	['AD1 - Incorrect decision or plan']	0.06	0.62	1.19
['PEW2 - Technology creates an unsafe situation']	['AD1 - Incorrect decision or plan']	0.09	0.60	1.16
['PPF3 - Confidence level']	['AD1 - Incorrect decision or plan']	0.05	0.80	1.55

['PPF7 - Risk underestimation']	['AD1 - Incorrect decision or plan']	0.08	0.76	1.47
['PCS1 - Inadequate experience']	['AD1 - Incorrect decision or plan']	0.11	0.67	1.30
['PCS2 - Lack of proficiency']	['AD1 - Incorrect decision or plan']	0.10	0.71	1.36
['PCS3 - Inadequate training or currency']	['AD1 - Incorrect decision or plan']	0.06	0.72	1.40
['PER3 - Misperception of changing environment']	['AD1 - Incorrect decision or plan']	0.05	0.87	1.68
['OS2 - Safety risk management (proactive)', 'OS5 - Publications / procedures / written guidance']	['LT1 - Inadequate leadership or supervision']	0.05	0.68	2.29

A prominent insight is the strong impact of habitual behaviours as shown in the fact that negative habit (PME3) will often lead to incorrect decision or plan (AD1) and no transmission of information (AC2). Both rules have a confidence of 1 and lift above 2.4 and 4.1 respectively and when combined together raises the lift to 9.7. This suggests that entrenched negative habits strongly correlate with both poor decisions and communication breakdowns. This effect compounds when paired with risk underestimation (PPF7), which further increases confidence and lift, results in a lift of 4.14. Such a high lift value indicates a strong positive association, suggesting these factors interact synergistically rather than occurring by chance. These findings suggest that occupational accidents may often emerge from the gradual normalisation of unsafe behaviours rather than isolated mistakes. Repeated exposure to routine tasks may encourage personnel to bypass established procedures, underestimate risks, or rely on habitual practices that deviate from formal operating requirements. Consequently, organisations should focus not only on procedural compliance but also on identifying and addressing behavioural drift through safety observations, supervisory engagement, and continuous reinforcement of safe working practices.

Cognitive precursors such as:

- Inadequate experience (PCS1)
- Inadequate training or currency (PCS3)
- Risk underestimation (PPF7)

Are repeatedly shown as strong antecedents to incorrect decision or plan (AD1). The confidence of 1 and lift of 2.4 across many of these rules indicates a consistent causal pattern, reinforcing that insufficient preparedness directly contributes to faulty decisions. Unlike collision and grounding incidents, where environmental complexity plays a significant role, the occupational accident rules suggest that deficiencies in competence and preparedness alone may be sufficient to initiate unsafe actions. This highlights the importance of competency assurance processes, including refresher training, mentoring programmes, and practical assessment designed to ensure that personnel maintain proficiency throughout their operational careers.

Workarounds (AI1) are another significant consequence, linked to both briefing or handover inadequate (PCO1) and risk underestimation (PPF7). Both individual and joint rules show moderate confidence and lift values exceeding 2.7, illustrating that procedural drift and informal practices often stem from gaps in communication and planning. These findings support the resilience engineering perspective that workarounds often represent adaptive responses to deficiencies in system readiness, enabling operators to adjust their performance to match the demands of real operational conditions rather than those assumed during system design (Hollnagel, 2017). Importantly, these findings suggest that workarounds should not necessarily be interpreted as deliberate violations, but rather as adaptive responses to deficiencies in planning, communication, or operational support. While such adaptations may enable work to contribute, they may also erode established safety barriers over time. Organisations should therefore treat the emergence of workarounds as a leading indicator of underlying system weaknesses and investigate why personnel feel compelled to deviate from prescribed procedures.

The rules for systemic issues only went back to leadership factors finding that inadequate risk assessment (LO1), no enforcement of existing rules (LT3), and inadequate leadership (LT1) lead to inadequate briefing or handover (PCO1). Indicating that failures at leadership level in control and oversight are closely tied to poor briefings, which in turn increase the likelihood of procedural and

planning failures. One of the most telling rules is a multi-layer rule combining all three of these. The presence of this multi-layer relationship reinforces the systemic nature of occupational accidents, demonstrating how weakness in leadership, rule enforcement, and risk assessment can propagate downward to influence communication quality and operational performance. This finding suggests that interventions aimed solely at frontline workers are unlikely to produce sustained improvements unless accompanied by stronger supervisory practices, effective safety assurance activities, and leadership development initiatives. Furthermore, recurring deficiencies in briefing and handovers may serve as leading indicators of broader organisational weaknesses and should therefore be actively monitored within safety management systems.

5.8.2.5 Summary

Across all four accident types common themes emerge that point to the systemic nature of maritime safety failures. A primary similarity lies in the prominence of cognitive and attentional precursors, such as risk underestimation, inattention and memory-related errors, consistently leading to incorrect decision-making. In all accident types, the confidence of rules linking these factors to decision failures is high, often close to or equal to q , indicating a robust and reliable relationship. This underscores the centrality of human error, particularly cognitive overload or misjudgement, as a cross-cutting causal mechanism in maritime accidents.

Despite this commonality, the mechanisms by which cognitive limitation manifest different by accident type. In a collision-scenarios, cognitive failures are often intertwined with perceptual issues, particularly failures in visual detection. This is distinct from grounding accidents, where the emphasis is more on decision-making under complex or uncertain environment conditions, such as poor weather or technological confusion. Fire-related incidents, on the other hand, are unique in showing strong deterministic patterns involving memory-based antecedents, and physiological stress, often resulting in both timing and planning failures simultaneously. This dual-error phenomenon was not as prominent in other accident types.

Another important similarity is the recurrent role of organisational factors in escalating risk. In all four types, rules involving inadequate leadership, safety management, and lack of enforcement or structure appear frequently as antecedents to high-level consequences like inadequate risk assessment or inadequate briefing or handover. However, the way these organisational deficiencies play out varies. In collision and grounding, organisational issues tend to result in supervisory or procedural lapses, often indirectly contributing to perceptual or planning errors. In fire and occupational accidents, these issues more directly result in breakdowns of communication and informal adaptations, such as workarounds, which suggest a behavioural response to systemic inadequacies.

A notable difference is the degree of complexity and convergence present in the rules. Fire and occupational accidents display the highest prevalence of multi-antecedent rules, often with perfect confidence and elevated lift values. This suggests a latent risk accumulation mode, where multiple collision and grounding accidents tend to be explained by fewer, more direct associations, often involving one or two key antecedents leading to single outcome. This could reflect the more situational and dynamic nature of these accidents, where failure occurs more rapidly and is less dependent on long-term latent conditions.

Another key difference lies in the role of environmental stressors. These were especially prominent in fire and grounding incidents contributing to both cognitive and perceptual failures. In contrast, such environmental factors were almost absent in occupational accidents, where behavioural habits and procedural breakdowns were the dominant themes. Lastly, only collision and fire accidents displayed rules with multiple consequences, suggesting divergent causality where single antecedents lead to both decision and perceptual failures. This suggests that certain unsafe conditions can initiate a cascade of failures affecting multiple operational layers. This was evident in grounding and occupational incidents, where causal pathways were more linear and targeted.

In summary, while all accident types share a foundational vulnerability rooted in cognitive limitations and organisational deficiencies, their expressions differ based on context. Collisions are characterised

by perceptual breakdowns, groundings by environmental complexity and inexperience, fires by cognitive overload under stress and occupational accidents by entrenched behavioural patterns and weak procedural oversight. Collectively, these findings demonstrate that maritime accidents rarely arise from isolated failures. Instead, they emerge through interactions between human, technological, environmental, and organisational factors, with the specific pathways varying according to accident context. The results further suggest that effective safety management should move beyond single-factor interventions and adopt a systems perspective that addresses upstream organisational conditions, competency management, human-centred design, and operational resilience simultaneously. Importantly, several antecedents identified through ARM, including deficiencies in training, communication, risk management, and procedural compliance, may function as leading indicators of emerging safety issues and therefore provide opportunities for proactive intervention before accidents occur.

5.9 Discussion of Model Performance, ARM Results and Practical Implications

5.9.1 Classification Performance

Overall, the findings indicate that an NLP pipeline combining domain-adapted language representations (SeaSafeBERT), hierarchical logistic regression with calibration, and human-in-the-loop decision logic, demonstrates meaningful discriminative capacity for many contributory factors from narrative reports and surface meaningful interaction patterns between them.

The stage 1 modelling results showed that high-level categories tied to concrete scene descriptors (e.g. Physical Environment, Perception, Competence/Capability) are detected with comparatively strong F1 and ROC-AUC, whereas abstract or sparsely represented categories (e.g. International Deviation, Response Execution) remain challenging.

At the code level, several factors exhibit operationally useful accuracy, notably no wrong/late visual detection (AP1), no transmission of information (AC2), incorrect decision or plan (AD1), briefing or handover inadequate (PCO1), inadequate leadership or supervision (LT1), inadequate risk assessment

(LO1), crew composition (LO2) and personnel/training resources (OR1) achieved very high scores. These codes achieved strong precision-recall balance (many with $F1 \geq 0.80$), suggesting that they may be suitable candidates for high-confidence automated flagging within a supervised, human-in-the-loop workflow.

The precision-recall trade-off was evident between other strong performers, suggesting complementary deployment modes since high recall models can drive broad triage and discovery, while high precision models support low-volume, high-confidence automation. These help support a mixed workflow where analysts receive both auto-coded factors and calibrated suggestions.

The tool demonstrates that a calibrated, hierarchical NLP pipeline can be operationalised in a way that aligns with real investigative workflow, not just algorithmic performance metrics. The Auto/Suggest mechanism, in particular, provides a structured compromise between automation and expert judgement, maintaining human oversight while reducing analytical burden.

5.9.2 Accident Patterns Identified Through ARM

The ARM analyses added explanatory context by highlighting how factors cluster within and across accident types. One key similarity between accidents lies in the prominence of cognitive and attentional precursors, such as risk underestimation, inattention and memory-related errors, consistently leading to incorrect decision-making. Across all accident types, the confidence of rules linking these factors to decision failures is high, indicating a consistently observed co-occurrence pattern within the analysed dataset. This pattern is consistent with prior research emphasising the importance of cognitive and organisational factors in maritime incidents.

Despite this commonality, the mechanisms by which cognitive limitation manifest different by accident type. In a collision-scenarios, cognitive failures are often intertwined with perceptual issues, particularly failures in visual detection. This is distinct from grounding accidents, where the emphasis is more on decision-making under complex or uncertain environment conditions, such as poor weather or technological confusion. Fire-related incidents, on the other hand, are unique in showing

comparatively stronger multi-antecedent associations involving memory-based antecedents, and physiological stress, often resulting in both timing and planning failures simultaneously. This dual-error phenomenon was not as prominent in other accident types.

Another important similarity is the recurrent role of organisational factors in escalating risk. In all four types, rules involving inadequate leadership, safety management, and lack of enforcement or structure appear frequently as antecedents to high-level consequences like inadequate risk assessment or inadequate briefing or handover. However, the way these organisational deficiencies play out varies. In collision and grounding, organisational issues tend to result in supervisory or procedural lapses, often indirectly contributing to perceptual or planning errors. In fire and occupational accidents, these issues more directly result in breakdowns of communication and informal adaptations, such as workarounds, which suggest a behavioural response to systemic inadequacies.

A notable difference is the degree of complexity and convergence present in the rules. Fire and occupational accidents display the highest prevalence of multi-antecedent rules, often with perfect confidence and elevated lift values. This suggests a latent risk accumulation mode, where multiple collision and grounding accidents tend to be explained by fewer, more direct associations, often involving one or two key antecedents leading to a single outcome. This could reflect the more situational and dynamic nature of these accidents, where failure occurs more rapidly and is less dependent on long-term latent conditions.

Another key difference lies in the role of environmental stressors. These were especially prominent in fire and grounding incidents, contributing to both cognitive and perceptual failures. In contrast, such environmental factors were almost absent in occupational accidents, where behavioural habits and procedural breakdowns were the dominant themes. Lastly, only collision and fire accidents displayed rules with multiple consequences, suggesting divergent causality where single antecedents lead to both decision and perceptual failures. This suggests that certain unsafe conditions can initiate a

cascade of failures affecting multiple operational layers. This was evident in grounding and occupational incidents, where causal pathways were more linear and targeted.

Overall, all accident types share a foundational vulnerability rooted in cognitive limitations and organisational deficiencies, their expressions differ based on context. Collisions are characterised by perceptual breakdowns, groundings by environmental complexity and inexperience, fires by cognitive overload under stress and occupational accidents by entrenched behavioural patterns and weak procedural oversight. These insights suggest that while certain interventions may benefit all domains, each accident type also requires tailored strategies that reflect its unique risk dynamics.

Important, the ARM findings demonstrate that many contributory factors identified within accident investigations may also serve as leading indicators of emerging safety issues. Deficiencies in training, communication, risk management, procedural compliance, and supervisory practices repeatedly appeared as upstream contributors across accident types. Monitoring these factors within safety management systems may therefore provide organisations with opportunities to intervene proactively before unsafe conditions escalate into accidents. This highlights the value of combining accident investigation data with data-driven analytics techniques to support predictive and preventative safety management.

5.9.3 Methodological Implications

Beyond classification accuracy, the methodology provides a structured and potentially deployable framework, a suitable pathway for AI-assisted investigation. The SHIELD-NLP label set balanced causal factor complexity with learnability, preserving the intent of SHIELD while ensuring each class has sufficient support for stable training. The hybrid representation (SeaSafeBERT embeddings with TF-IDF) captures maritime semantics and the high-precision operational cues. The painting consistently improved precision for codes whose evidence is expressed through distinctive phrasing.

The hierarchical, calibrated OVR elastic net framework emphasises interpretability and control. Analysts can inspect coefficients and through probability calibration using the auto/suggest method

can understand the extent to which the model is sure of the classification. Future work on embedding these into an interface tool, turns the model into a learning system where submission of classifications is re-added to the model's database. Coupling the modelling process with rule-based coding also helped with very low sample sizes, allowing us not to remove them.

ARM then adds explanatory context of how these factors co-occur and breaking down to incident types allows us to differentiate and create actionable hypotheses for design, training, and governance interventions. Together these methods demonstrate the potential to support more consistent, and scalable learning from accidents at sea, subject to continued validation and oversight.

More broadly, the proposed framework demonstrates how artificial intelligence and machine learning techniques can be integrated with established accident investigation methodologies to support scalable and evidence-based safety analysis. Unlike traditional approaches that rely on manual coding qualitative interpretation, the framework enables systematic extraction of contributory factors and their interactions across large accident datasets. This provides a foundation for more proactive, data-driven safety management and offers a transferable methodology that may be applied across other safety-critical domains beyond maritime transport.

The deployment of AI-assisted classification in safety-critical domains requires careful governance. Automated outputs should be interpreted as probabilistic suggestions rather than authoritative determinations. Over-reliance on automated coding risks reinforcing historical attribution biases embedded in past investigations. For this reason, the proposed system is designed as a human-in-the-loop decision-support tool, where final coding authority remains with trained analysts.

5.10 Chapter Summary

This chapter presented the implementation and evaluation of the high-information pathway of the proposed methodological framework through the development of SHIELD-NLP, a machine-learning-ready adaptation of the SHIELD taxonomy. The chapter demonstrated how a structured human factors taxonomy can be combined with natural language processing, domain-specific language models, and

supervised machine learning to automate the classification of human and organisational factors from detailed maritime accident investigation reports. The consolidation of the original SHIELD taxonomy into SHIELD-NLP addressed challenges associated with class imbalance while maintaining theoretical consistency, enabling robust multi-label classification of contributory factors.

The results demonstrated that the proposed two-stage classification framework achieved strong performance across many human, technical, leadership, and organisational categories, particularly for frequently occurring and operationally significant factors. The development of the SeaSafeBERT language model and the deployment of the SHIELD-NLP decision support tool further illustrated the practical feasibility of integrating automated human factors classification into real-world accident investigation workflows. In addition, the association rule mining analysis demonstrated the framework's ability to move beyond classification by identifying meaningful relationships between contributory factors, providing insights into the complex interactions that underpin maritime accidents.

Having established the effectiveness of the high-information pathway for analysing detailed investigation reports, the remainder of the thesis focuses on the low-information pathway introduced in Chapter 6. The following chapter applies this pathway to a large ferry near-miss dataset, demonstrating how topic modelling, large language models, temporal analysis, and association rule mining can be used to extract actionable safety insights from short incident narratives where detailed investigation reports are not available.

6. Uncovering patterns in Ferry Near misses using topic modelling and GPT-assisted labelling

6.1 Chapter Overview

Near-miss reporting systems represent a critical but unexplored source of safety intelligence in ferry operations. Unlike accident investigations, near-miss narratives capture both hazardous conditions and the adaptive actions that prevent escalation, providing direct insight into how operational safety is actively maintained. This case study therefore, adopts a Safety-II perspective, focusing not only on what could have gone wrong but also on how frontline practices, coordination, and technical controls enable safe outcomes under variable and often constrained operating conditions. The research presented in this chapter forms the basis of the peer-reviewed publication *Uncovering patterns in ferry near misses: A topic modelling approach with GPT-assisted labelling* (Black et al., 2026a), with this chapter providing additional detail, analysis, and discussion beyond that presented in the published article.

This chapter explores the application of the framework, specifically to short-text narratives of ferry near-misses. Using BERTopic, incident descriptions were classified to identify key themes in incident types and prevention strategies, and an LLM was used to interpret topics. A cross-analysis of these categories highlights critical areas for improvement while also showcasing instances where skills and teamwork or operational safety play a vital role in effective prevention.

6.2 Ferry Near Miss Data and Modelling Approach

This section of the research utilises a comprehensive dataset of 4,360 near-miss incidents sourced from a Ferry operator, covering a range operational scenarios across marine and harbour environments. The data provides rich contextual, environmental, and procedural information which enables a multifaceted exploration of safety incidents including the reason the near miss did not turn into an accident. The dataset contained the following information:

1. **Incident meta data:** including a unique near miss ID, incident date and time, and detailed vessel-specific information including asset name and region.
2. **Open Text:** “Near miss description” detailing what happened, “what in your opinion prevented it from developing into an accident?” detailing the prevention. These fields are central to topic modelling efforts and allow for the extraction of themes and contextual relationships through NLP.
3. **Categorisations:** “Near miss category” and “Near miss sub category” offer structured classification of the vent type, which were used as a validation of the BERTopic classifications.
4. **Other details:** the database also contains information such as the weather condition, visibility, wind direction and strength, sea state and impact score. As well as this there are further details, however, the majority of other details are only completed for very small amounts of incidents.

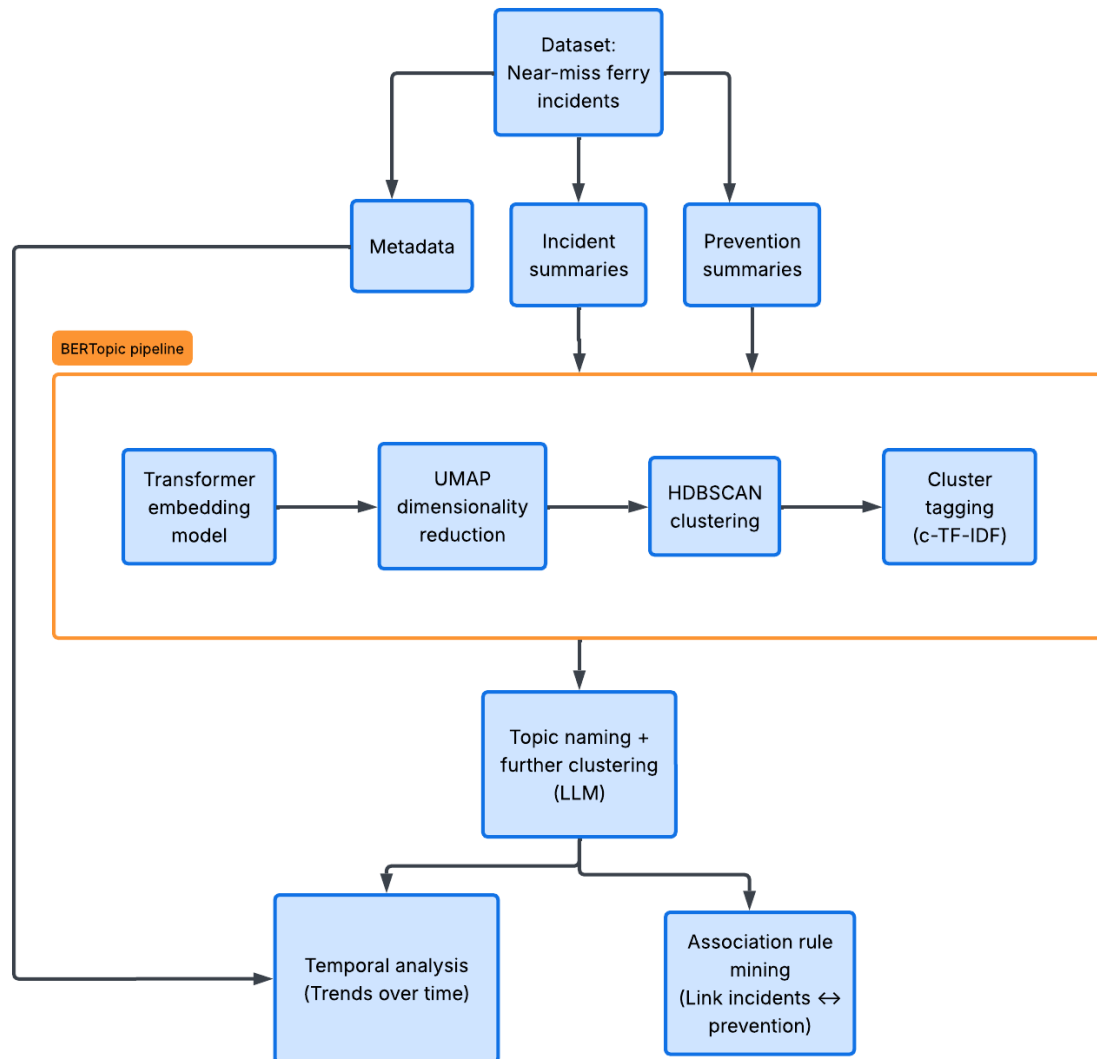


Figure 34. Flow diagram of the methodology used in the ferry near-miss case study.

Figure 34 showcases the overall approach used in this case study; more details are discussed in Chapter 4. As BERTopic relies on transformer-based embeddings, traditional text preprocessing (e.g. stop word removal or lemmatisation) is unnecessary and may even reduce semantic quality. We therefore used only minimal cleaning prior to modelling. This included the removal of exact duplicate records, correction of character encoding issues, standardisation of whitespace and punctuation artefacts, and the exclusion of records with missing narrative contexts.

Abbreviations, spelling inconsistencies, and maritime-specific terminology were retained in their original form. This decision was made because transformer-based language models are generally

robust to minor spelling variations and domain-specific vocabulary, and excessive normalisation may remove operational nuances present within safety narratives. Similarly, no manual synonym mapping or maritime terminology standardisation was applied. Very short reports were not removed; however, many of these records were subsequently identified as outliers by the HDBSCAN clustering stage due to insufficient semantic content for reliable clustering. This approach ensured that the original meaning and operational context of the reports were preserved while reducing obvious sources of textual noise.

6.3 Results of the Application of BERTopic to Ferry Near-misses

6.3.1 Near Miss Incident Summary Classification and Topic Naming

Application of BERTopic to the near-miss incident summaries generated 76 distinct topics. One large outlier cluster contained 1,656 (~38%) reports that were too generic or ambiguous for meaningful interpretation and was excluded from further analysis. The remaining 75 topics (shown in Table 17) spanned a diverse set of operational, environmental, safety, and human-factors issues. Topic sizes ranged from 10 to 161 reports, with a median of 28, indicating a balanced distribution across the dataset.

Comparison with the company's existing classification system showed that 66% of the BERTopic-derived clusters could be mapped directly onto predefined categories. The remaining 34% represented either (a) edge cases not explicitly covered by the taxonomy, (b) overlapping themes that cut across categories, or (c) novel patterns not previously recognised. This demonstrated that BERTopic can replicate established taxonomies while also surfacing under-specified or emergent issues, thereby extending the analytical value of incident reports beyond manual classification.

Table 17. Clustered topic groups of incident summary topic modelling.

Cluster	Topic numbers	Topic Names	N (% of Reports)

Passenger & Public Safety	2, 5, 9, 10, 11, 14, 15, 22, 35, 45, 55, 62, 66, 74	Walkway and Marshalling Lane Barriers, Slippery Surface and Growth, Maritime Equipment and Safety, Pier Safety and Public Access, Door and Gate Operation, Cyclist Safety and Bike Rental, Child Safety Around Railings, Maritime Accidents and Damages, Pet-related Incidents, Docking and Disembarkation Incidents, Passenger Vessel Toilet Issues, Seat and Chair Construction and Repair, Customer Service Desk Screen Incident, Police on Male Passenger Behaviour	642 (23.74%)
Mooring & Berthing Operations	0, 1, 16, 19, 34, 40, 41, 58, 61, 69	Maritime Mooring and Rope Maintenance, Waste Oil Disposal and Management, Maritime Navigation and Vessel Movements, Ferry Ramp Operations, Harbour Berthing Challenges and Vessel Safety, Maritime Engine Troubleshooting, Vehicle Discharge and Traffic Issues, Maritime Navigation Incidents, Marine Winch and Mooring Operations, Maritime Equipment Failures and Troubleshooting	508 (18.79%)
Traffic & Vehicle Operations	6, 7, 24, 26, 27, 29, 32, 33, 42, 44, 46, 65, 71, 73, 75	Road and Vehicle Incidents, Pedestrian and Vehicle Interaction, Car Handbrake Issues in Vehicles, HGV Trailer Traffic and Marshalling, Traffic Congestion at Roundabouts, Ferry Terminal Traffic Management, Public Transportation Queues and Disruptions, Vehicle Deck and Mezzanine Loading, Vehicle Parking and Traffic Issues, Coach Travel and Traffic Regulations, Vehicle Transportation and Safety, Passenger Vessel Incidents at	444 (16.42%)

		Gourock, Workplace Car Park Incidents, Traffic Congestion at Kilchoan Pier, Speed Limit Violation at Pier	
Navigation & Bridge Management	3, 13, 21, 23, 31, 37, 38, 43, 54, 59	Marine Vessel Operations and Emergency Drills, Ship Thruster Failures, Mallaig Harbour Vessel Navigation, Marine Vessel and Yachting Operations, Yachting Incident at Rothesay Harbour, Boat/Ship Navigation Control, Marine Vessel Maintenance and Navigation, Autopilot System Issues, Schottel Steering System Issues and Solutions, Yacht Navigation and Safety Protocols	376 (13.91%)
Mechanical & Power Systems	4, 12, 20, 48, 53, 56, 64, 68	Generator and Power System Management, Engine Maintenance and Troubleshooting, Electrical Overheating and Safety Issues, Fire Safety Equipment Maintenance, Welding and Fire Safety, Fire Door and Fuse Safety, Fuel and Water Filter Maintenance, Fuel Pipe Leak and Repair in Ship Engines	266 (9.84%)
Fire & Emergency Systems	17, 18, 39, 47, 57, 67	Lifejacket Use and Safety, Smoking and Fire incidents, Passenger and Vehicle Disembarking Issues, Covid-19 and Ambulance Operations, Seaside Activities and Locations, Vehicle Activity at Lock Slipway	156 (5.77%)
Workplace & Crew Practices	25, 36, 49, 50, 63, 70, 72	Construction and Safety at Seaports, Night Watchman Duties and Challenges, Safety Equipment Regulations at Work Sites, Laundry and Dryer Equipment, Maritime Supervision and Vessel Management, UK Marine Infrastructure Maintenance, Intermittent Fault Investigation	136 (5.03%)

Cargo & Logistics	8, 30	Gas Shipment and Transport, Loading and Safety Procedures in Transport	92 (3.40%)
Environmental & Infrastructure	28, 51, 52, 60	Harbour and Maritime Activities, Fishing Activities at Loch Shira in Largs, Pier Lighting Issues, Seaman Booking and Organisation Issues	84 (3.11%)

6.3.2 Time-series Analysis

To explore how incident frequencies varied over time, a five-year dataset (2019-2023, plus early 2024) was analysed using monthly and annual decompositions (Figure 35 and Figure 36). The results reveal clear seasonal dynamics as well as longer-term shifts associated with operational stressors such as the COVID-19 pandemic.

Across months, incident patterns show clear seasonality. Passenger and traffic-related categories exhibit sustained peaks during the summer months (May-July), aligning with increased travel demand and peak tourist and commuting periods. The Traffic and Vehicle operations and Passenger and Public safety groups consistently intensify from late spring into summer, reflecting congestion pressures, higher passenger volumes, and increased pedestrian-vehicle interactions at terminals, car parks, and loading ramps. Similarly, cargo and logistics issues also show mid-year increases, potentially linked to heightened freight and baggage-handling activity during high-demand seasons.

Conversely, winter months demonstrate elevated rates of Mechanical and power system incidents, environmental and infrastructure incidents, and workplace and crew practices incidents. This pattern reflects environmental and operational conditions such as adverse weather, reduced daylight, tidal constraints, and maintenance-season activity. These groups show January and late-year concentrations consistent with maintenance cycles, asset wear, and environmental stress on infrastructure and equipment. Navigation-related events display more distributed, year-round patterns, indicating persistent operational exposure rather than seasonal concentration.

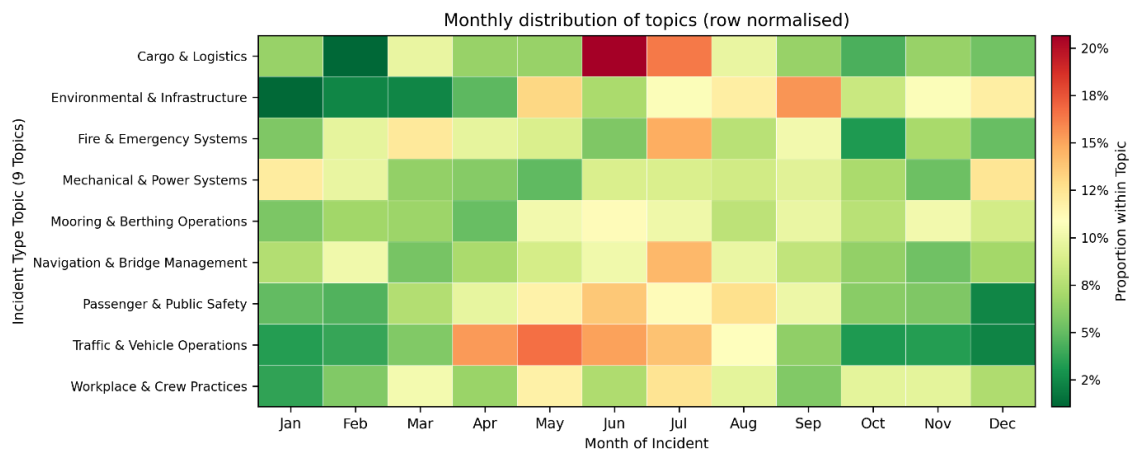


Figure 35. Seasonal patterns of incidents by incident topic and month.

Annual trends further emphasise the impact of broader systemic factors. The pandemic period (2020-2021) was marked by relative reductions in passenger and traffic-related categories, aligning with lower passenger and transport activity. These categories then demonstrate strong rebounds in 2022, consistent with post-restriction recovery and renewed pressures. By contrast, mechanical and power systems and workplace and crew practices show relative increases in 2023, potentially reflecting deferred maintenance, heightened inspection regimes, or a return to normal operational tempo, revealing latent system issues.

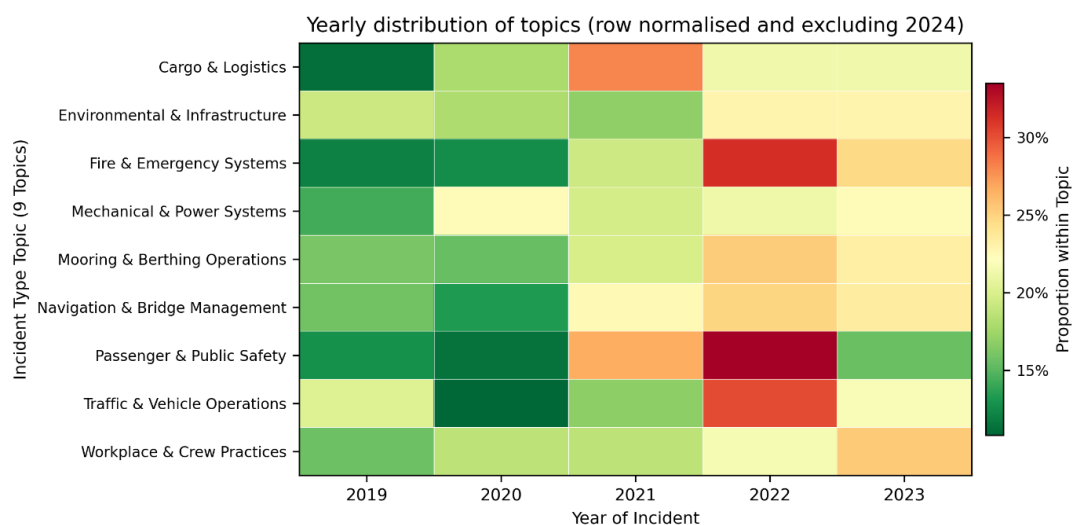


Figure 36. Heatmap of yearly incidents by topic from 2019 to 2023.

6.3.3 Classification Preventions

Topic modelling of the free-text prevention measures yields 87 distinct topics after excluding 1,287 (~30%) outlier entries that were too generic or ambiguous for meaningful interpretation. These were grouped into eight higher-level clusters (Table 18), each capturing a coherent set of prevention practices. The distribution of reports across clusters was relatively balanced, with no single cluster dominating, reflecting the breadth of operational risk management strategies in place.

Table 18. Clustered topic groups of near miss prevention

Cluster	Topic numbers	Topic Names	N (% of Reports)
Vehicle & Road Safety	6, 8, 10, 15, 17, 21, 24, 36, 38, 48, 49, 56, 60, 61, 62, 83, 86	Staff directing traffic and park safety, Car marshalling and vehicle control, Traffic speed limitations, Stationary vehicles, Vehicle patrol and vigilance, Driver awareness and attention, Road traffic incidents, Driver safety and reaction time, Vehicle-pedestrian falls, Vehicle horn usage, Pedestrian and vehicle crossings, Vigilant vehicle drivers, Vehicle loading and unloading procedures, Customer safety and moving vehicles, Quiet morning road traffic, Traffic management in carparks, Traffic and pedestrian safety	565 (18.38%)
Crew & Team Management	0, 5, 12, 27, 31, 41, 66, 71, 72, 78, 82	Bridge Team Management and Communication, Maritime watchkeeping and machinery observations, Crew observation and awareness, Crew intervention and guidance, Crew vigilance at work, Staff reactions and responses, CFL staff communication and safety procedures, Master's decision support and development, Crew	550 (17.91%)

		response and direction, Staff vigilance and resilience, Crew management and problem resolution	
Maritime Operations & Vessel Handling	4, 23, 26, 29, 33, 34, 53, 57, 58, 73, 77, 85, 87	Vessel service safety, Boat manoeuvring and thrust control, Boating and rope handling, Skipper experience and vessel operation issues, Maritime mooring techniques, Marine steering and control systems, Ship handling and seamanship knowledge, Ferry master's situational awareness, Maritime safety and vigilance, Tug operation and safety, Maritime berthing procedures, FRC recovery and launching procedures	384 (12.49%)
Passenger, Crew & Public Safety	9, 13, 16, 18, 25, 52, 59, 63, 64, 65, 81	Crew members' quick actions and reactions, Security and cleanliness in public places, Passenger and crew safety alerts, Describing favourable weather conditions, Work interruption and escalation, Desolate surroundings, Passenger and crew safety situations, Marine safety procedures, Marine staff operations, Marine alarm system and monitoring, Clothing storage and laundry management	368 (11.97%)
Maintenance & Technical Safety	3, 19, 30, 32, 37, 39, 47, 51, 54, 70, 74	Routine maintenance and inspections, Manual chain positioning on platforms, Deck equipment safety, Oil spill management and prevention, Generator Capacity and Functionality, Preventing injuries by maintenance balance, Preventing pedestrian falls, Door and gate security, Electrical power supply issues, Electrical circuit safety and maintenance, Valve operation and maintenance	354 (11.53%)

Risk, Luck & Uncertainty	2, 14, 20, 22, 67, 68, 69, 75, 79, 84	Luck and Chance, Good fortune and timing, Scottish isles and lochs, Undefined topic, Accidental falls and reactions, Managing customer balance and falls, Uncertainty and negation, Corrective actions and appreciation events, Discussions on luck, Knowledge and uncertainty	324 (10.55%)
Port, Pier & Shore Operations	1, 35, 40, 44, 45, 46, 50, 76, 80	Pier Workers, Port Assistant Observations and Actions, Port management and interventions, Port staff process management and intervention, Officers' intervention and observations, Port staff training and vigilance, Marine ramp operations and maintenance, Quiet pier and reduced sailings, Ferry traffic and disembarkation	320 (10.40%)
Emergency Preparedness & Response	7, 11, 28, 42, 43, 55	Fire prevention and detection, Emergency response and awareness, Potential accident risks, Emergency drill discoveries, Fast problem resolution & accident response, Emergency procedures and equipment operations	211 (7.18%)

The preventative clusters show that safety measures were distributed across technical, operational, and human domains. The largest share related to vehicle and road safety, particularly traffic marshalling and pedestrian-vehicle interactions in terminal and ferry deck settings. Crew and team management was also highly represented, underscoring the centrality of vigilance, communication, and leadership in incident prevention. Other major clusters included passenger and public safety, maritime operations and vessel handling, and maintenance and technical safety, reflecting the breadth of frontline operational practices. Smaller but significant clusters captured port and shore operations, emergency preparedness. Together, these categories illustrate a comprehensive landscape of preventative strategies, ranging from routine vigilance and technical reliability to readiness for rare but high-consequence events.

The risk, luck and uncertainty category differed from the other preventative categories in that some narratives did not describe an identifiable safety barrier or active preventative intervention. Instead, the avoidance of harm was attributed to favourable circumstances or the non-occurrence of further hazardous events, indicating an absence of effective risk controls rather than the successful operation of a preventative barrier.

6.3.4 Linking Incident Types to Prevention Actions

The network diagram in Figure 37 illustrated the relationships between ferry incident types (blue nodes) and their associated prevention measures (green nodes) identified through association-rule mining. The edge width represents the confidence of each rule, thicker lines denote stronger and more consistent relationships, while edge colour intensity indicates lift, showing the relative strength of the association beyond random chance.

Several clear clusters and associations were found linking technical incidents with asset-management controls. One notable association was the relationship between generator and power system management and generator capacity and functionality. This link reflects the strong dependence of power-related incidents on the adequacy and reliability of onboard generator systems. Examination of underlying reports revealed that these near misses were associated with a diverse range of technical failures, including generator trips, failed starts, cooling system deficiencies, fuel supply restrictions, electrical protection faults, and load management issues. Despite the diversity of failure mechanisms, the corresponding restoration actions consistently focused on maintaining adequate generating capacity, preventing redundancy through standby generators, and ensuring reserve power availability. This finding suggests that the primary safety mechanism was not the prevention of individual component failures, but maintenance of power system resilience. In many cases, vessel operations continued safely because reserve generating capacity was available to absorb load fluctuations or compensate for equipment failures before they affected operational performance.

Associations related to passenger and vehicle movements highlighted the importance of managing shared operational spaces. The relationship between passenger-vehicle interaction incidents and pedestrian and vehicle crossing controls revealed that these near misses frequently occurred during embarkation and disembarkation, when pedestrians crossed active vehicle lanes, moved between marshalling and terminal areas, or entered linkspan zones while vehicles remained in motion. The corresponding prevention measures consistently involved marshals stopping or slowing traffic, controlling vehicle movements, and facilitating designated pedestrian crossings. These findings suggest that the principal risk arose from the coexistence of pedestrian and vehicle flows rather than isolated unsafe acts. Consequently, traffic segregation, designated crossing points, and active marshalling emerged as the primary mechanisms through which passenger-vehicle interaction incidents were prevented. Related clusters involving walkway and marshalling lane barriers, customer safety, moving vehicles and traffic speed limitations reinforced this pattern. A further association linked marine vessel operations and emergency drills prevented by FRC (Fast Rescue Craft) Recovery and Launching Procedures. Many of these near misses occurred during training exercises, drills, inspections, or routine testing of FRC systems. The incidents exposed a range of conditions, including damaged wires, defective hooks, hydraulic system faults, limit switch failures, equipment wear, and operator handling errors. Prevention actions consistently involved controlled recovery procedures, equipment inspections, functional testing, servicing activities, and operational restrictions when defects were identified. These findings indicate that emergency drills served a dual safety function. Beyond preparing crews for emergency response, they also provided opportunities to identify equipment defects and procedural weaknesses before they contributed to operational failures.

Similarly, door and gate operations were strongly associated with door and gate security. The incidents frequently involved access-control failures, defective lock mechanisms, jammed doors, and situations where access to operational spaces was either unintentionally restructured or insufficiently secured. Prevention actions consistently involved prompt reporting by crew members, corrective maintenance, lock replacement and active intervention to restore safe access. Effective control, therefore, depended

on both the integrity of locking systems and the vigilance of personnel who identified and responded to defects before they escalated.

Fire safety and environmental management linkages also emerge clearly. The relationship between smoking and fire incidents and fire prevention and detection demonstrated that many of the incidents originated from relatively minor ignition sources, including discarded smoking materials, hot-work activities, smouldering waste, overheating equipment, and vehicle-related defects. In most cases, the incidents were identified during the early stages of fire development when only smoke, unusual odours, or localised heating were present. The corresponding prevention actions consistently involved early detection, prompt reporting, rapid investigation, and immediate intervention by passengers, crew members, or other personnel. Several incidents were successfully interrupted because individuals recognised weak warning signs, such as smoke burning smells, or abnormal equipment behaviour, before the situation escalated into a more significant fire event. These findings suggest that effective fire prevention relied primarily on early detection, reporting, and intervention rather than the activation of firefighting systems after escalation has occurred. Likewise, the association between waste oil disposal and management and oil spill management and prevention reflects the environmental dimension of ferry safety operations, illustrating how improper waste handling directly contributes to spill events.

Infrastructure-related controls were evident through the association between ferry ramp operations and marine ramp operations and maintenance. Many near misses occurred during ramp deployment, recovery, or alignment with shore infrastructure. The incidents frequently involved equipment degradation, ramp-linkspan interface issues, damaged components, and operational anomalies identified during routine use. Prevention actions consistently involved inspections, early defect identification, immediate maintenance interventions, and suspension of operations when abnormal conditions were observed. These findings suggest that ramp safety is influenced not only by the mechanical condition of the equipment but also by effective management of the vessel-shore

interface, with incidents often prevented through the timely recognition of developing defects before operational failure occurred.

Navigational and bridge management relations form one of the largest and most coherent sub-clusters within the network. The connection between yacht navigation and safety protocols and ferry masters' situational awareness indicated that near-miss and collision risks involving smaller crafts were strongly influenced by the master's ability to maintain dynamic situational awareness and adapt to unpredictable movements of non-commercial vessels. Examination of the reports revealed that many of these events involved recreational vessels making unexpected manoeuvres, crossing the bows of large vessels, impeding traffic in constrained waterways, or failing to follow established navigation rules. In contrast, the corresponding prevention narratives consistently attributed successful avoidance of escalation to the situational awareness of the ferry master. The ability of bridge teams to continuously monitor vessel movement, anticipate developing conflicts, and intervene early therefore emerged as the primary mechanism preventing close-quarters situations involving recreational and non-commercial vessels.

A Similar pattern was evident in the association linking yachting incidents at Rothesay Harbour, Mallaig Harbour vessel navigation, and maritime navigation and vessel movements, with bridge team management and communication. The Rothesay incidents frequently occurred within congested harbour environments involving multiple interacting waterway users, including yachts, kayaks, recreational craft, harbour authorities, and commercial vessels. These situations involved unexpected manoeuvres, route conflicts, and interactions requiring rapid adaptation by ferry crews. The corresponding prevention actions consistently highlighted effective bridge-team communication, adherence to established navigational procedures, coordination with harbour authorities, and professional watchkeeping practices. Successful outcomes were associated with information sharing, collective decision-making, and proactive traffic management by bridge teams.

Other maritime operations also show similar relational clusters. For example, the linkage between maritime mooring and rope maintenance with both boasting and rope handling and pier workers. Reports revealed that many near misses involved rope entanglement, snagging, unexpected tensioning, rope degradation, and interactions between mooring systems and harbour infrastructure. Several incidents involved conditions that could have resulted in snap-back events, rope failure, or personnel exposure to stored-energy hazards. The corresponding prevention actions consistently highlighted crew vigilance, continuous monitoring of rope behaviour, routine inspections, and the timely identification of developing hazards before operations continued. In many cases, incidents were prevented because abnormal rope movement, deterioration or tension changes were recognised early and corrective actions were taken before failure occurred. These findings suggest that safe mooring operations depend not only on the physical condition of ropes and equipment but also on the ability of personnel to recognise weak signals of deterioration and intervene before escalation. With the association to pier workers, prevention actions highlighted communication between vessel and shore personnel, rapid intervention, and the identification of unsafe conditions before operations continued. In several cases, incidents were prevented because either vessel crews or pier staff recognised developing hazards and communicated this information before personnel entered danger areas or line tensions increased further.

Taken together, these findings indicate that mooring safety depended on both equipment integrity and coordinated vessel-shore operations. Prevention was achieved not solely through procedural compliance, but through shared situational awareness, timely communication, and the collective ability of shipboard and shoreside personnel to recognise and respond to evolving hazards before they developed into injury events or equipment failures.

Finally, the association between public transportation queues and disruptions and luck and chance revealed a markedly different pattern from the other relationships identified within the network. Examination of the reports highlighted a recurring problem whereby passengers were required to

queue on public roads, often without designated waiting areas, physical barrier, or effective separation from vehicle traffic. Several narratives explicitly noted that these situations occurred repeatedly during busy periods and had been raised on multiple occasions over several years. Unlike most other associations identified in this study, the corresponding prevention narratives did not describe active interventions, procedural safeguards, or established operational controls. Instead, incidents were frequently avoided simply because no vehicle happened to enter the occupied space at the time or because potentially hazardous circumstances did not materialise. In these cases, the successful prevention of harm appeared to depend largely on favourable circumstances rather than identifiable protective mechanisms. This finding suggests the absence of effective risk controls rather than the presence of successful preventative measures. It further demonstrates how near-miss analysis can reveal persistent vulnerabilities that may remain hidden within traditional incident classification systems. Whereas many of the other associations identified adaptive responses that interrupted escalation pathways, this relationship exposed situations in which organisational resilience appeared weak, and safety performance relied heavily on change rather than robust defences.

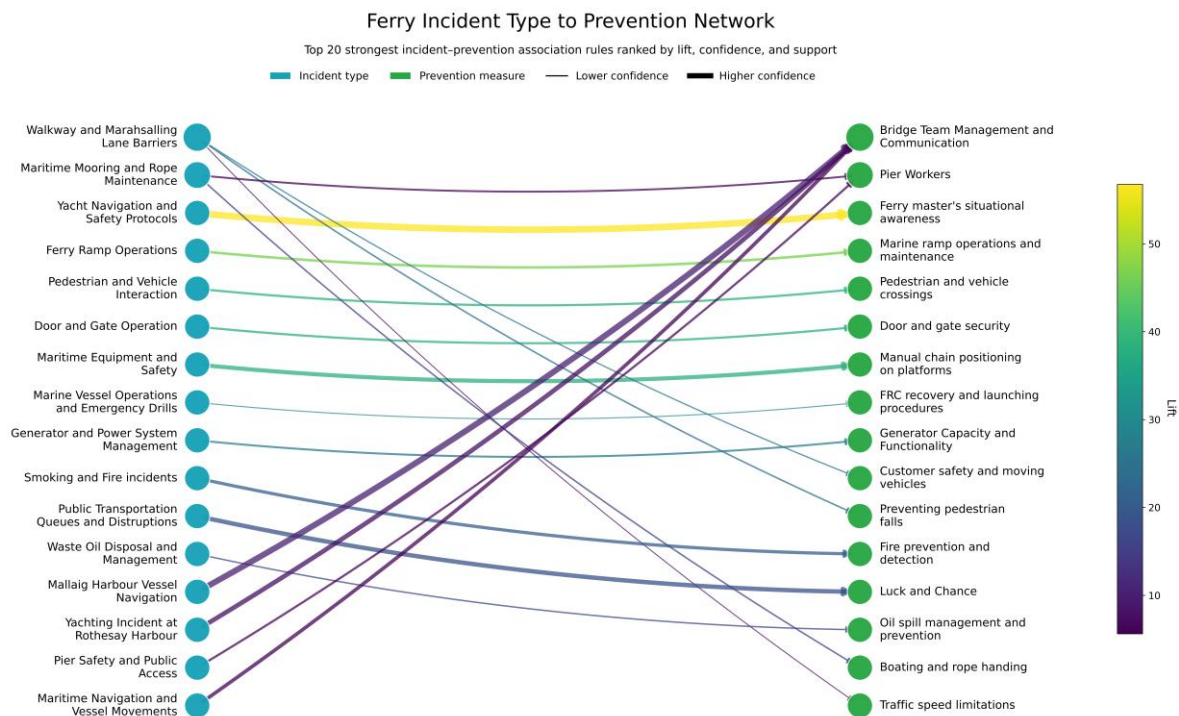


Figure 37. Network of incident type and prevention association rule mining results.

Overall, the ARM results demonstrate the prevention in ferry operations is achieved through a distributed set of technical, operational, and organisational controls rather than through isolated interventions. Technical incidents were primarily mitigated through proactive maintenance, redundancy, and system resilience, while passenger and traffic-related incidents relied heavily on movement segregation, active marshalling, and traffic management. Navigational and mooring incidents were consistently prevented through effective bridge resource management, shared situational awareness, and coordinated ship-shore communication. Importantly, not all associations reflected robust safety defences. Certain passenger-management scenarios appeared to depend largely on favourable circumstances rather than established controls, highlighting persistent vulnerabilities within the ferry system. Collectively, these findings demonstrate how near-miss analysis can identify both resilient operational practices and areas where existing safety barriers remain weak, thereby providing a foundation for more targeted and evidence-based safety management interventions.

6.3.5 Expert Validation

6.3.5.1 Incident Summary Validation

To assess the interpretability of BERTopic clusters, four domain experts independently evaluated topic quality across three dimensions: keyword coherence, GPT-generated label fit, and document relevance. Agreement was assessed using multiple reliability metrics.

Krippendorff's α and Kendall's W indicated relatively low ordinal agreement ($\alpha=0.24-0.39$, $W=0.09-0.12$), suggesting that experts did not consistently converge on identical interpretations of topic quality. The alpha values were substantially below the 0.667 minimum for drawing even tentative reliability conclusions and therefore do not demonstrate satisfactory inter-rater reliability. This indicates that, although experts could independently evaluate the generated topics, they did not consistently converge on the same assessment of topic quality. Given that near-miss narratives often contain overlapping operational, technical, and contextual elements, some variation in semantic interpretation was expected. This observation also reflects a broader challenge in incident analysis, whereby even experienced practitioners may differ in how they categorise and interpret complex events.

Cronbach's α scores were consistently above 0.72 (range: 0.72-0.83), showing that individual raters applied their own standards consistently rather than randomly. Intraclass correlation coefficients (ICC) provided further evidence of reliability: keyword coherence achieved good agreement (ICC = 0.79, $p<0.001$), while GPT label fit (ICC=0.69) and document relevance (ICC=0.70) demonstrated moderate agreement (Table 19). These results suggest that the identified topics were generally interpretable, although the precise boundaries and labels assigned to some topics remained subject to expert judgement.

Table 19. ICC scores per evaluation dimension.

Dimension	ICC(2,k)	Interpretation	p-value
-----------	----------	----------------	---------

Keyword Coherence	0.791	Good agreement	<0.001
GPT Name Fit	0.689	Moderate Agreement	<0.001
Document Relevance	0.700	Moderate Agreement	<0.001

Mean ratings across experts were positive for all three dimensions (Table 20). GPT-generated labels and document relevance were scored slightly higher than keyword coherence, suggesting that while topics were generally interpretable, the specific keyword groupings were less intuitive.

Table 20. Mean expert scores across topic validation dimensions.

Expert	Mean Keyword Coherence	Mean GPT Label fit	Mean label-document fit
Expert 1	4.029	3.943	3.743
Expert 2	3.632	3.868	4.022
Expert 3	3.575	3.725	3.875
Expert 4	3.707	3.697	3.973
Average	3.73	3.81	3.90

Qualitative reviewer comments provided further nuance to these findings. Some topics were judged to have overly narrow or overly specific labels (e.g. “waste oil disposal” applied to broader engine-related issues; “emergency drills” covering routine operational practices). Such observations reinforce the subjective nature of assigning semantic labels to emerging themes and illustrate the challenges inherent in achieving complete consensus when interpreting operational narratives. Despite these inconsistencies, most clusters were considered interpretable, relevant, and aligned with their underlying reports.

Taken together, the validation demonstrated that the majority of BERTopic-derived incident clusters were perceived as both meaningful and usable for safety analysis, although experts differed in their interpretation of some themes. The combination of low agreement coefficients and moderate reliability measures reflects the inherently subjective nature of interpreting complex near-miss narratives. These findings suggest that the extracted topics provide a plausible and operationally useful representation of the underlying reports, while acknowledging that alternative interpretations may also exist.

6.3.5.2 Prevention Validation

As with incident classifications, prevention clusters were evaluated by four domain experts. Agreement was low when assessed with Krippendorff's α (0.13-0.21), substantially below the 0.667 minimum for tentative reliability, reflecting variability in how experts interpreted qualitative prevention themes. The broad behavioural adaptations, overlapping operational practices and context-dependent actions represented within prevention narratives may have contributed to this variability, although the validation exercise did not specifically test the causes of expert disagreement. Consequently, the prevention-topic evaluations should be interpreted as individual expert assessments rather than evidence of a reproducible expert consensus.

Nevertheless, Intraclass correlation coefficients (ICC) indicated moderate agreement from keyword coherence (ICC=0.62, $p<0.001$), GPT label fit (ICC=0.61, $p<0.001$) and document relevance (ICC=0.47, $p=0.005$) (Table 4).

Table 21. ICC scores per evaluation dimension.

Dimension	ICC(2,k)	Interpretation	p-value
Keyword Coherence	0.620	Moderate agreement	<0.001
GPT Name Fit	0.612	Moderate Agreement	<0.001
Document Relevance	0.474	Moderate Agreement	0.005

Internal consistency was high for three of the four raters (Cronbach's $\alpha=0.80, 0.90, 0.92$), with one expert showing lower reliability ($\alpha=0.37$). This suggests that while individual interpretations varied, most experts applied scoring criteria consistently. These findings suggest that, although experts did not always agree on the exact interpretation of individual prevention topics, most applied their evaluation criteria consistently. The resulting prevention clusters should therefore be viewed as exploratory yet operationally meaningful representations of the data rather than definitive classifications.

Mean ratings across dimensions (Table 22) were generally positive, with GPT-generated labels receiving slightly higher scores than keyword coherence or document relevance.

Table 22. Mean expert scores across topic validation dimensions.

Expert	Mean Keyword Coherence	Mean GPT Label fit	Mean label-document fit
Expert 1	4.366	4.268	4.146
Expert 2	3.357	3.571	4.048
Expert 3	4.05	4.025	3.45
Expert 4	3.195	3.268	3.268
Average	3.742	3.783	3.728

Taken together, these results mirror those for incident topics: although experts varied in their detailed interpretations, overall ratings were favourable, and agreement was sufficient to confirm that the prevention clusters were interpretable and practically meaningful.

6.4 Discussion of Near-Miss Analysis Findings

This study applied BERTopic, supported by GPT-generated labels, to a large corpus of near-miss reports from ferry operations. The results demonstrate that topic modelling can extract coherent and

interpretable trends from unstructured safety data, replicate known taxonomies, and reveal emergent themes not captured in existing reporting frameworks. Validation with domain experts confirmed the practical interpretability of most clusters, while temporal and association analyses provided new insights.

Existing maritime safety studies have predominantly relied on accident investigations and predefined taxonomies to identify risk factors. In contrast, this study demonstrates how near-miss narratives can be systematically analysed to identify both emerging hazards and the operational mechanisms that prevent escalation. By linking incidents directly to the preventative actions documented by frontline personnel, the findings provide insights into how safety is actively maintained within ferry operations.

The application of BERTopic produced 75 interpretable incident topics and 87 prevention topics that captured a wide spectrum of operational, technical, and human factors in ferry near misses. When comparing to the company taxonomy for incident types, it was found that 66% of incident topics aligned with the existing taxonomy, while the remainder surfaced gaps and cross-cutting patterns that current labels do not capture. On the prevention side, staff routinely documented positive adaptations, crew vigilance, emergency preparedness, pier team interventions, and seamanship actions, yet these measures are seldom analysed in conventional Safety-I reporting. Together, the findings support a Safety-II perspective: near-miss narratives contain rich evidence of what enables operations to go right, not only what goes wrong.

One of the most common incident clusters was passenger and public safety, accounting for near-misses involving walkways, boarding/disembarking, and pedestrian-vehicle proximity. Similarly, traffic and vehicle operations were found to be the third most common incident type cluster, including HGV marshalling, roundabout congestion issues, and deck loading. Seasonal analysis showed clear summer peaks in these categories. The high volume of tourists and seasonal passengers, many of whom are unfamiliar with port layouts, onboarding procedures, and driving within ferry-terminal environments, may contribute to the increase in these incidents. Kim et al. (2022) highlighted how ferry-terminal

walkability is influenced by the design and organisation of passenger movement routes, particularly during embarkation and disembarkation. The present findings complement this perspective by demonstrating that these environments are also associated with increased operational safety challenges, particularly during periods of elevated passenger demand. While Kim et al. focused on passenger perceptions and satisfaction, the near-miss data analysed here provide evidence of how pedestrian-vehicle conflicts and movement-related hazards emerge in practice and require active intervention from ferry personnel.

Mooring and berthing operations were also found to be a recurring incident type including berthing, ramp operations, and line handling. Navigation and bridge management formed the fourth largest cluster, covering harbour-approach navigation, bridge resource management, and control-system issues such as thruster failures, autopilot anomalies and Schottel steering faults. Findings from ARM show that these type of harbour-specific topics were strongly associated with pier-team actions and bridge communication. This supports previous maritime human-factors research highlighting communication, coordination, and shared situational awareness as a critical safeguard during port approaches and other high-workload operational phases (Hetherington et al., 2006). While previous studies have largely identified communication breakdowns and bridge-team deficiencies as contributors to accidents, the present findings extend this evidence by demonstrating that effective coordination between vessel and shore personnel contributes to the successful management of navigational and berthing hazards before they escalate.

In addition to these themes, we also see clear patterns related to mechanical and power-system reliability, which show noticeably higher incident and near-miss rates during the winter months. This aligns with the operational reality that winter is the primary period for major vessel overhauls, maintenance windows, and engineering interventions. This interpretation is consistent with common ferry industry practice, whereby major maintenance and refit programmes are strategically scheduled during quieter winter periods to maximise fleet availability during the summer seasons. While these

activities are essential to ensure the fleet is fully available during the peak summer season, they may temporarily increase the likelihood of mechanical or power-system issues as vessels return to service after refit periods or operate on tighter engineering schedules. These findings suggest that increases in near-miss reporting do not necessarily indicate deteriorating safety performance. Instead, they may reflect periods of heightened operational sensitivity associated with maintenance transitions and engineering interventions. Ferry operators should therefore interpret near-miss trends alongside maintenance schedules and operational context when prioritising safety interventions and allocating engineering resources. However, as maintenance activities were not explicitly captured within the dataset, further investigation would be required to confirm this relationship.

Annual patterns revealed a marked decline in near-misses during the COVID-19 period. The reduction in near misses during the pandemic reflects temporary simplification of operations, reduced system load, and lower coordination demands. This was followed by a rebound in 2022 which illustrates how risk re-emerges and complexity returns with the quick recovery of demand. Although previous studies have primarily examined accident trends and safety challenges in ferry systems, rather than near-miss reporting specifically, they similarly emphasise that changes in operational conditions, traffic levels, and system demands can alter the nature and frequency of safety events (Golden and Weisbrod, 2016, Hebbar et al., 2023). Consequently, safety-management systems should pay particular attention during periods of organisational or operational change, including demand recovery, route expansion, and service restructuring, as these phases may introduce new interactions, coordination requirements, and emerging risk pathways.

The prevention topics reveal a distributed set of effective control functions: traffic marshalling and pedestrian management, crew/bridge teamwork and vigilance, seamanship-based stabilisation, and routine technical reliability work, with additional contributions from port/shore operations and emergency preparedness. Collectively, these prevention themes provide empirical evidence of the adaptive capacities that enable ferry operations to remain resilient under varying conditions.

Importantly, the findings demonstrate that such adaptations can be systematically identified from large-scale narrative datasets rather than solely through qualitative case studies.

The ARM results reinforce this perspective by connecting specific incident types directly to corresponding prevention actions. For example, the linkage between generator and power system management and generator capacity and functionality highlights the importance of proactive reliability assurance and load verification in maintaining power integrity. Other relationships emphasise the value of operational resilience achieved through situational awareness, procedural competence, and effective team communication and coordination. This finding complements previous accident-based research by Sandhåland et al. (2015) who found that deficiencies in situational awareness frequently preceded collisions and navigational incidents. In contrast, the present study reveals the opposite perspective, demonstrating how effective bridge teamwork, communication, and situational awareness repeatedly prevented conflicts involving yachts, fishing vessels, harbour traffic, and technical anomalies from escalating into accidents. Whereas accident investigations primarily document the absence of effective controls, the present findings highlight the adaptive actions ferry crews use to manage operational disturbances and maintain safety in everyday operations.

The practical value of the ARM approach lies in its ability to identify where specific preventive mechanisms are most likely to interrupt escalation pathways. For ferry operators, this provides an evidence base for integrating near-miss information into maintenance planning, targeted training initiatives, and resource allocation. For port authorities, the findings highlight opportunities to strengthen passenger segregation measures, terminal layouts, and vessel-shore coordination practices. At the regulatory level, the systematic analysis of near misses provides a complementary source of leading safety indicators that can identify emerging vulnerabilities before they manifest as accidents. Rather than replacing existing safety-management frameworks, these approaches may enhance their predictive capability and support more proactive forms of risk governance.

More broadly, this study demonstrated the wider value of near-miss reporting beyond traditional hazard identification. First, the analysis showed that larger-scale narrative datasets can be systematically analysed to identify recurrent operational trends, seasonal patterns, and emerging risk areas that may not be evident through conventional reporting summaries. Second, by examining documented prevention actions, the study revealed the adaptive practices and successful interventions that prevented escalation, thereby extending near-miss learning beyond a solely failure-oriented perspective. These are the findings exposed limitations within existing reporting frameworks, including the inability of predefined taxonomies to fully capture contextual information and preventative activities recorded by frontline personnel. These limitations suggest that improvements in report structure, taxonomy design, and automated analytical support may enhance organisational learning. Finally, the proposed BERTopic-ARM framework provides a scalable approach for transforming large volumes of narrative near-miss reports into actionable safety intelligence, supporting more proactive and systematic learning across ferry operations.

The expert validation process found that model outputs were coherent, applicable and useful for analysis. Average ratings across four domain experts were positive for keyword coherence, label fit, and document relevance (means ~3.7-3.9), and Intraclass correlation coefficients indicated moderate-good agreement. While ordinal agreement (Krippendorff's α , Kendalls' W) was low, reflecting differences in rankings, internal consistency was high for most raters, and comments indicated only minimal issues (for example, overly narrow labels). The validation therefore provides preliminary evidence that the generated topics were perceived as coherent and relevant by the participating experts, but does not establish that these judgements were consistently reproducible across individual assessors. Accordingly, the BERTopic-GPT outputs were considered suitable for exploratory analysis within this study, but individual topic interpretations and downstream findings should be treated cautiously and should not be interpreted as reflecting validated expert consensus.

6.5 Practical Implications of the Findings and Method

For ferry operators, these insights highlight the need for proactive resourcing that reflects seasonal patterns of demand. This includes scaling up traffic management and passenger safety measures during peak summer months, while using the quieter winter season for pre-emptive infrastructure inspections and maintenance. The associations identified through ARM further suggest that near-miss information can be integrated more directly into operational decision-making. For example, ramp-related near misses may act as early indicators of developing asset degradation at the vessel-shore interface, supporting more targeted inspections and preventative maintenance. Similarly, generator-related incidents emphasise the importance of maintaining reserve capacity and redundancy within critical systems. Training initiatives can also be aligned with recurrent operational risks, for example, improving bridge resource management in congested waterways involving recreational or non-compliant vessels, communication, reinforcing effective mooring and vessel-shore communication practices and improving the management of passenger-vehicle interfaces and fall prevention measures. Emergency drills should also be viewed not solely as compliance activities, but as opportunities to identify latent equipment and procedural weaknesses before failures occur in operational settings.

For port authorities, the findings highlight opportunities to strengthen the management of shared operational spaces. Passenger-vehicle interaction incidents frequently occurred where pedestrians and vehicle movements overlapped and were successfully mitigated through active marshalling, designated crossing, and traffic control measures. In contrast, the association between passenger queues and luck and chance identified situations in which safety appeared to depend on favourable circumstances rather than robust preventative controls. These findings suggest that terminal layouts, designated waiting areas, physical segregation measures, and traffic-flow arrangements should be reviewed to minimise reliance on frontline intervention and improve resilience during periods of peak demand.

At an industry level, our findings show that NLP-based approaches can add significant value to existing reporting frameworks. By surfacing emergent risks, linking incidents to preventative strategies, and enhancing the predictive use of near-miss data, operators can move from reactive safety management to a more proactive, system-wide approach. More specifically, narrative-based analyses of near misses can complement traditional lagging indicators by identifying both emerging vulnerabilities and the adaptive mechanisms through which escalation is prevented. This provides regulators and operators with an evidence base to support targeted interventions, facilitate cross-organisational learning, and strengthen the predictive capabilities of existing safety management systems. Ultimately, this strengthens resilience, reduces operational disruptions, and builds greater passenger confidence in ferry services.

Despite the value demonstrated in this study, the quality of learning obtainable from near-miss reports remains dependent on the completeness and consistency of the narratives submitted. Many reports contained limited contextual information, variable levels of detail, and inconsistent descriptions of prevention actions. To improve organisational learning, operators should encourage richer narrative reporting through clearer reporting guidance, structured prompts, and training on effective near-miss reporting. The inclusion of dedicated fields capturing contextual factors, successful interventions, and lessons learned may further strengthen the ability of organisations to extract actionable knowledge from near miss datasets. In addition, integrating NLP-based decision-support tools within reporting systems could assist in automatically identifying emerging trends, highlighting missing information, and supporting more timely safety interventions.

6.6 Chapter Summary

This chapter demonstrated the application of the low-information pathway of the proposed analytical framework to a large-scale dataset of ferry near-miss reports. Using BERTopic, large language model-assisted topic labelling, temporal analysis, and association rule mining, the framework successfully transformed thousands of short narrative reports into structured safety intelligence. The analysis

identified recurrent operational themes, seasonal trends, and relationships between incident types and the preventative actions that successfully interrupted accident escalation. By analysing both incident and prevention narratives, the framework provided insights not only into the hazards encountered during ferry operations but also into the adaptive practices and operational behaviours that contribute to resilient safety performance, supporting a Safety-II perspective.

The findings demonstrated that large volumes of routinely collected near-miss reports, which are often impractical to analyse manually, can be systematically processed to generate meaningful and actionable safety knowledge. This illustrates the practical value of the low-information pathway for supporting proactive safety management and organisational learning from near-miss reporting systems.

Having demonstrated the framework within the maritime ferry sector, the following chapter applies the same low-information pathway to offshore wind incident data. This enables an assessment of the framework's transferability to a different safety-critical industry and evaluates its ability to identify trends and emerging safety issues within a rapidly expanding operational domain.

7. Offshore Wind Farm Occupational Incident Analysis using topic modelling and GPT-assisted labelling

7.1 Chapter Overview

Occupational accidents remain a significant safety challenge within the offshore wind industry due to the hazardous nature of activities undertaken during construction, development, and operational phases. Personnel are routinely exposed to risks associated with lifting operations, vessel transfers, working at heights, adverse environmental conditions, and interactions with heavy equipment. Despite the continued growth of the offshore wind sector, understanding of the underlying patterns and causal mechanisms associated with these incidents remains limited, particularly when relying on large volumes of unstructured safety reports.

This chapter extends the application of the proposed methodological framework beyond the maritime domain to the offshore wind sector, demonstrating its transferability to another safety-critical industry. Using incident data collected by G+, the chapter applies the same analytical principles to examine offshore wind safety events across construction, development, and operational phases. The analysis identifies dominant incident types, injury patterns, and causal factors, illustrating how unstructured narratives can be systematically transformed into actionable safety insights.

7.2 Offshore Wind Data and Modelling Approach

The data utilised in this study was provided by G+, a global health and safety organisation that serves as a unifying force within the offshore wind industry. G+ collects and publishes safety incident data statistics from its member companies, which include leading offshore wind operators across Europe, North America and Asia. While the dataset does not represent all offshore wind operators, it provides valuable insights into safety and trends across a significant amount of the industry (G+, 2023b, G+, 2024). As such, it offers a practical and realistic view of operational incidents across various phases. This dataset comprises incident descriptions and their corresponding classifications for 5403 offshore

wind incidents spanning years 2017 to 2021. These incidents encompass all three pivotal stages in the lifecycle of an offshore wind platform: construction, development, and operational sites. Each incident is annotated with multiple binary classification labels, including whether the event was a High Potential incident, whether it involved a Dropped Object, and whether it required an Emergency Response or Medical Evaluation (ERME). As such, the classification task represents a multi-label problem, where a single incident can belong to multiple categories simultaneously. The incidents take place across the world, within the following countries: UK, US, Ireland, Denmark, France, Germany, Netherlands, Poland, Sweden, and Taiwan. Of the 5403 incidents, around 20% were near miss incidents and the remaining 80% were accidents resulting in damage or injury.

This chapter applies the low-information branch of the proposed methodological framework to offshore wind incident reports, shown in Figure 38. Given that the G+ dataset primarily comprises short unstructured narrative descriptions accompanied by a limited set of structured labels, the analytical approach combines unsupervised and supervised NLP techniques to extract both descriptive and predictive insights.

First, BERTopic is applied to identify dominant incident themes and recurring operational patterns within the narrative reports. Association Rule Mining (ARM) is then used to explore relationships between identified topics, incident characteristics, and injury outcomes, providing insights into common co-occurring risk factors. Finally, supervised multi-label classification is undertaken to predict key incident outcomes, including High Potential events, Dropped Objects, and Emergency Response or Medical Evaluation (ERME) requirements. Together, these analyses demonstrate how unstructured offshore wind incident narratives can be transformed into actionable safety intelligence while assessing the transferability of the proposed framework beyond the maritime domain.

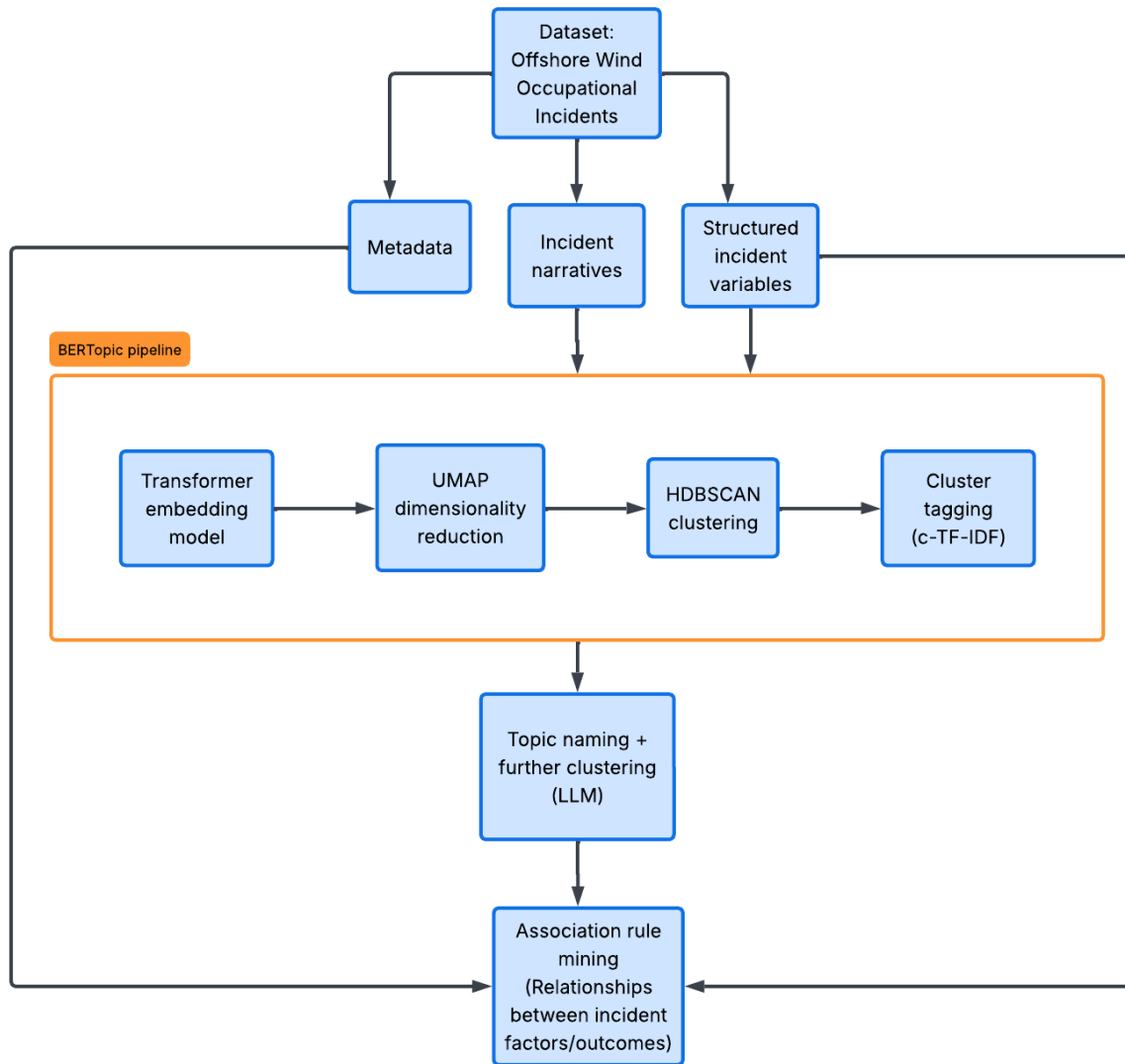


Figure 38. Flow diagram of the methodology used in the offshore wind case study.

7.3 Statistical Overview of the Offshore Wind Dataset

Before applying the proposed NLP framework, a descriptive statistical analysis was undertaken to characterise the structured offshore wind incident database. Table 23 summarises the principal characteristics of the dataset, including the distribution of incidents across project phases and the predefined safety classifications available within the G+ reporting system.

Table 23. Summary of the structured offshore wind incident dataset (2017-2021)

Variable	Category	Number of incidents	Percentage (%)
Dataset	Total incidents	5,403	100.0

	Countries represented	12	–
Site type	Construction	970	18.0
	Development	155	2.9
	Operation	2,942	54.5
	Project site	1,334	24.7
High Potential Incident	Yes	1,205	22.3
	No	4,198	77.7
Dropped Object	Yes	513	9.5
	No	4,890	90.5
ERME	Yes	185	3.4
	No	5,218	96.6
Hazardous liquid release	Major/Severe	2	<0.1
	Moderate	8	0.1
	Minor	45	0.8
	Negligible	53	1.0
	None / Not applicable	5,295	98.0

More than half of all reported incidents (54.5%) occurred during the operational phase, while construction and project-site activities accounted for a further 18.0% and 24.7%, respectively. Development activities represented only a small proportion of reports (2.9%), reflecting the comparatively shorter duration and smaller workforce typically associated with this phase.

The structured classifications indicate that approximately one in five incidents (22.3%) were designated as High Potential Incidents, demonstrating that a substantial proportion of reported events had the potential to result in significantly more severe outcomes. In comparison, dropped objects accounted for 9.5% of incidents, while only 3.4% required an Emergency Response or Medical Evaluation. Releases of hazardous liquids were uncommon, with fewer than 2% of incidents involving any reported release and only two events classified as major or severe.

Table 24 summarises the recorded consequences associated with the incident reports. Hazard observations constituted the largest category (**32.2%**), followed by first aid injuries (**21.9%**) and near hits or near misses (**20.7%**). More severe outcomes, including lost workday injuries, medical treatment injuries, and restricted workday injuries, collectively accounted for less than **11%** of all reports. This distribution demonstrates that the database contains a large proportion of lower-consequence events

and hazard observations, providing valuable opportunities for proactive safety learning rather than solely retrospective analysis of serious accidents.

Table 24. Distribution of recorded incident consequences

Consequence	Number of incidents	Percentage (%)
Hazard	1,742	32.2
First aid injury	1,183	21.9
Near hit / Near miss	1,117	20.7
Asset damage	770	14.3
Lost workday injury	240	4.4
Medical treatment injury	213	3.9
Restricted workday injury	138	2.6

7.3.1 Temporal Analysis of Incidents

Figure 39 illustrates the monthly distribution of reported offshore wind incidents between 2017 and 2021. The largest number of reports occurred in 2017, with monthly incident totals remaining consistently above those recorded in subsequent years. From 2018 onwards, reporting volumes remained relatively stable, with monthly incident counts typically ranging between approximately 40 and 90 incidents.

Across all years, a seasonal pattern is evident. Incident frequencies generally increase during the spring and summer months, reaching their highest levels between July and August before declining towards the end of the year. December consistently recorded the lowest number of incidents across the study period. Although this seasonal trend may reflect changes in operational activity, weather conditions, or workforce exposure, the descriptive statistics alone do not allow these factors to be distinguished and further investigation is therefore required.

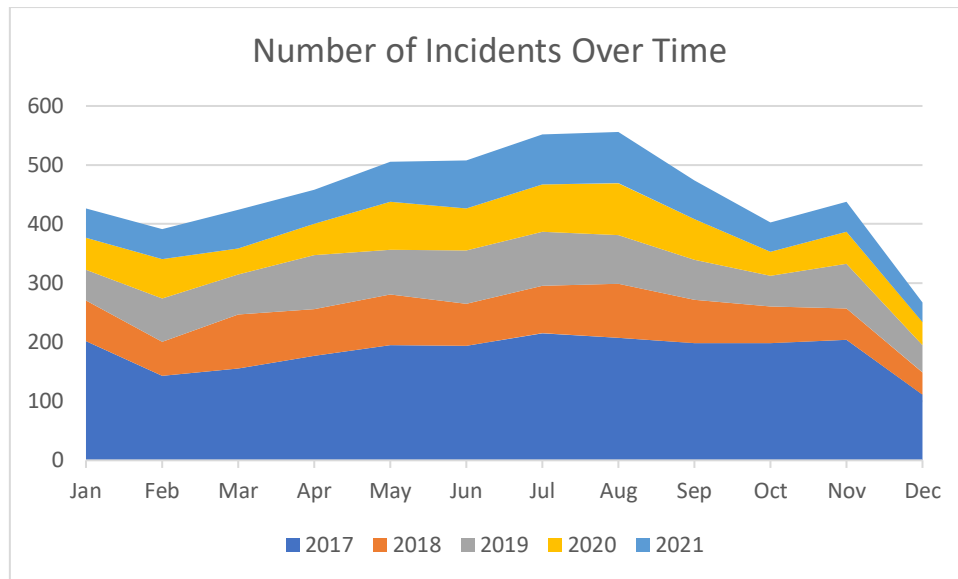


Figure 39. Monthly distribution of offshore wind incidents by year.

7.3.2 Geographical Distribution

Figure 39 presents the geographical distribution of the offshore wind incidents included within the G+ database. Incident reports originated from ten identified countries, although country information was unavailable for approximately 41% of records. Consequently, the geographical analysis should be interpreted as representing only those incidents where location information was recorded.

Among the identified locations, the United Kingdom accounted for the largest proportion of reported incidents, with 2,211 records (40.9% of the complete dataset), substantially exceeding all other countries. Germany recorded the second highest number of incidents (393), followed by Denmark (232), Taiwan (108), the Netherlands (91), the United States (76), and France (71). Sweden contributed relatively few incidents (18), while Ireland and Poland recorded only two and one incidents respectively.

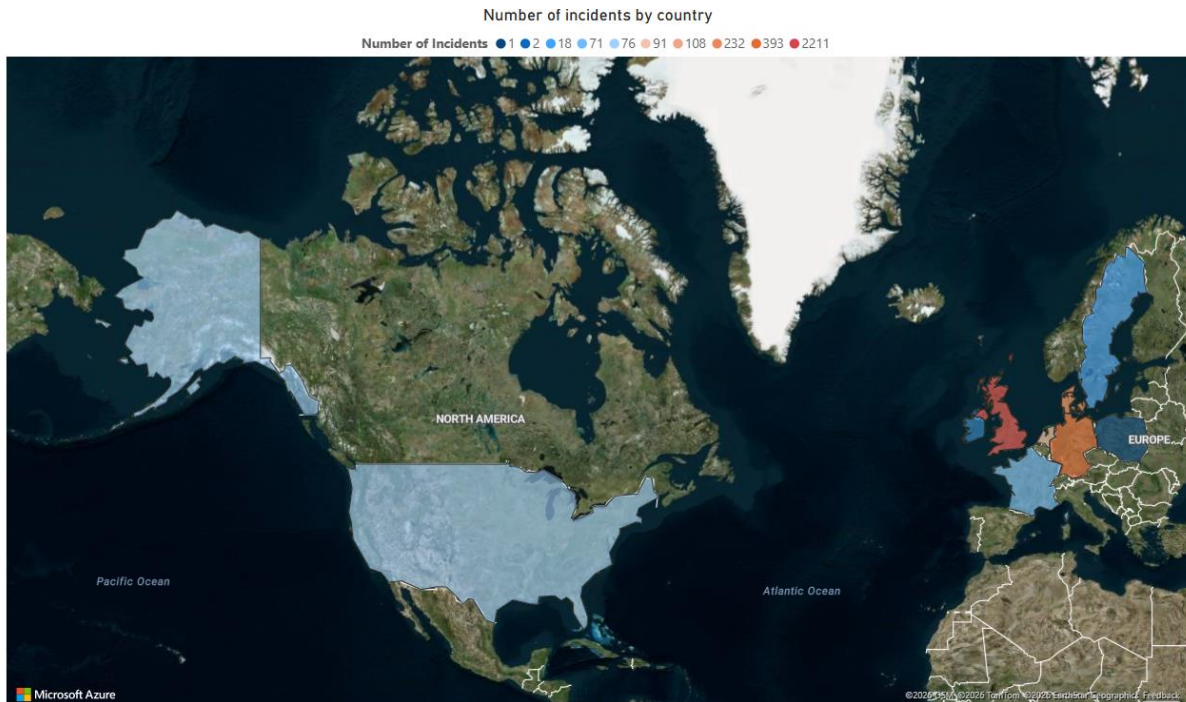


Figure 40. Geographic distribution of offshore wind incidents.

The dominance of reports from the United Kingdom is consistent with its position as one of the world's most mature offshore wind markets, possessing extensive installed capacity and a large operational workforce throughout the study period. Similarly, Germany and Denmark represent long-established offshore wind sectors with substantial operational activity. However, these frequencies should not be interpreted as direct indicators of national safety performance, as they are influenced by factors including the number of participating organisations contributing data to G+, the scale of offshore wind deployment within each country, and differences in reporting practices.

7.3.3 High-Potential Incidents Across Site Types

To investigate whether the likelihood of a high-potential incident differed between project phases, a chi-squared test of independence and univariate logistic regression were performed. Both analyses identified a statistically significant relationship between site type and the occurrence of high-potential incidents (χ^2 , $p < 0.00001$). Logistic regression demonstrated that, compared with construction sites, incidents occurring at operational sites were significantly less likely to be classified as high-potential (OR = 0.71, 95% CI: 0.60–0.84, $p < 0.001$), corresponding to approximately 29% lower odds of being classified as high-potential. Similarly, incidents occurring at project sites were associated with 23%

lower odds of being classified as high-potential than construction sites (OR = 0.77, 95% CI: 0.64–0.94, $p = 0.009$). No statistically significant difference was observed between construction and development sites (OR = 0.91, 95% CI: 0.62–1.35, $p = 0.648$).

Table 25 compares the observed and expected frequencies of high-potential incidents across the four site types. Construction sites recorded 261 high-potential incidents compared with 216 expected, indicating an excess of 45 incidents. Conversely, operational sites recorded 608 high-potential incidents compared with 656 expected, representing 48 fewer incidents than expected. Project sites closely matched the expected distribution (296 observed versus 297 expected), while development sites also differed only marginally from expectation (39 observed versus 35 expected).

Table 25. Observed and Expected number of incidents by site type.

	Construction site	Development site	Operation site	Project site
No	709 (754)	116 (120)	2334 (2286)	1038 (1037)
Yes	261 (216)	39 (35)	608 (656)	296 (297)

These findings suggest that although operational sites account for the largest proportion of reported incidents, incidents occurring during the construction phase are proportionally more likely to be classified as high-potential. This is likely attributable to the nature of construction activities, which frequently involve heavy lifting operations, working at height, large-scale installation activities, and dynamic work environments that inherently present greater opportunities for severe consequences should an incident occur. Conversely, operational activities are generally characterised by more established procedures, stable infrastructure, and mature safety management practices, which may contribute to the lower proportion of high-potential events observed within the operational phase.

7.3.4 Dropped object vs High-Potential

Given the significance of dropped objects as a leading hazard within the offshore wind industry, an additional statistical analysis was undertaken to examine their relationship with high-potential incidents. A chi-squared test of independence identified a statistically significant association between the occurrence of a dropped object incident and whether an incident was classified as high-potential

($\chi^2(1) = 112.27, p < 0.001$). Univariate logistic regression further demonstrated that incidents involving dropped objects were associated with significantly greater odds of being classified as high-potential than incidents not involving dropped objects (OR = 2.72, 95% CI: 2.25–3.29, $p < 0.001$).

These findings indicate that, although dropped object incidents accounted for only 9.7% of all reported incidents (513 of 5,294), they represented 17.8% of all high-potential incidents (208 of 1,166). The logistic regression indicated that incidents involving dropped objects had approximately 2.7 times greater odds of being classified as high-potential than incidents not involving dropped objects. This finding highlights dropped objects as a disproportionately severe hazard within offshore wind operations and reinforces the industry's continued emphasis on dropped object prevention through robust equipment inspections, effective tool tethering and securing practices, exclusion zones during lifting operations, and improved controls for working at height and suspended loads.

7.3.5 Limitations of this Data

The preceding statistical analyses demonstrate that the structured fields contained within the G+ database provide valuable insights into the overall characteristics of offshore wind incidents. Information such as the geographical distribution of incidents, project phase, temporal trends, consequence classifications, and indicators such as High Potential Incidents (HPI), Dropped Objects, and Emergency Response or Medical Evaluation (ERME) enable organisations to monitor safety performance, identify broad trends, and prioritise areas requiring further investigation. These structured variables therefore form an important component of organisational safety management and performance monitoring.

However, while these predefined classifications describe **what** occurred, they provide only limited understanding of **how** and **why** incidents occurred. For example, identifying that construction sites experience a greater proportion of high-potential incidents does not explain the specific work activities, equipment, environmental conditions, or human behaviours that contributed to those events. Similarly, although dropped object incidents can be identified from the structured fields, the

data does not indicate whether these occurred during lifting operations, maintenance activities, vessel transfers, or other operational tasks.

The structured dataset also lacks detailed information regarding the circumstances surrounding incidents, including the specific body parts injured, mechanisms of injury, equipment involved, environmental conditions, workplace hazards, and the human and organisational factors that contributed to the event. Understanding these contextual elements is essential for designing targeted interventions. For example, knowing that a first aid injury occurred provides relatively little guidance for prevention compared with understanding that the injury resulted from manual handling during turbine maintenance, involved the hand or lower back, and was influenced by inadequate housekeeping or restricted workspace.

Consequently, while structured statistical analysis provides an essential overview of offshore wind safety performance, it cannot fully support evidence-based intervention planning or organisational learning. To obtain this deeper operational understanding, the following section applies BERTopic to the free-text incident narratives. By analysing the narrative descriptions, the proposed approach automatically identifies recurring operational activities, injury mechanisms, equipment failures, environmental conditions, and causal factors that are not explicitly represented within the structured database. This richer contextual information complements the structured statistics and provides a more comprehensive foundation for proactive safety management within the offshore wind industry.

7.4 Offshore Wind Incident Topic Modelling Results

BERTopic produced a total of 87 topics from the 5403 incident descriptions it was given. Among these, 2508 incidents were classified as outliers due to insufficient information for categorisation. The identified topics encompass a range of aspects, including the type of incident, its outcomes, and the nature of injuries sustained. This variability arises from the diverse descriptions provided by different individuals and companies who are reporting these incidents. Consequently, some reports focus solely on the injuries resulting from the incidents, while others detail the events leading up to them.

However, the majority of descriptions lack sufficient information to ascertain the underlying causes of these incidents.

The topics named specified by GPT-4 were then used to further group the topics into incident types, injury types and causal factors. The grouping was also completed by GPT4 where the overwhelming majority of topics were related to the type of incident with 71.3% of the incident classified into this category. However, 18.4% focused on the type of injury sustained, and another 10.3% addressed human factors or a higher-level reasoning behind the incidents.

7.4.1 Incident Types

Since there were many incident types broken down into many different topics during the clustering process, incident types were further grouped into 7 clusters shown in Table 26. The topic with the highest number of incidents was incidents involving equipment, machinery and lifting operations. These incidents refer to crane, lifting gear, load handling, and warehouse machinery incidents, encompassing both operational errors and mechanical faults, such as dropped objects, sling failures, and unsafe mooring or lifting practices.

The next most common was maritime and vessel operation incidents. This covers all incidents related to vessel use, docking, personnel transfers, and offshore marine operations. This includes ladder access, mooring failures, subsea equipment work, and dropped object at sea. Such incidents are particularly hazardous due to the offshore environment, where emergency response is more complex and conditions can rapidly deteriorate.

Mechanical, hydraulic and electrical systems incidents then involved operational or maintenance issues. These were incidents that include electric shock, cable faults, oil leaks, turbine malfunctions, and generator issues. Failure in any of these systems can halt operations, cause environmental harm and lead to serious injury.

Structural and access safety incidents covers hazards linked to fixed site infrastructure like gates, ladders, handrails, scaffolding, and lighting systems. Failures in these areas increase the risk of falls,

restricted access, or collisions. They also include issues with site-specific structures like nacelle hatches and drainage systems that impact safe navigation and operational access.

Tools, workshop and maintenance safety incidents were also common with 190 occurrences. These stem from the handling, storage, and operation of tools and workshop equipment. They include risks from cutting tools, misplaced or faulty equipment, and maintenance tasks involving construction or installation processes. This also includes emergency kit maintenance and life jacket and survival equipment maintenance.

Fire, chemical and hazardous substance incidents involve allergic reactions, improper storage or disposal, exposure to harmful materials, and hot surface contact. These incidents can cause immediate harm to individuals, create long-term health risks, or escalate into major emergencies if uncontrolled.

Lastly, there were 147 vehicle and land related incidents. This involves operational safety concerns and mechanical faults of land-based vehicles. These can range from traffic-related hazards to risk during on-site driving and transport, posing threats to both personnel and equipment integrity.

Table 26. Incident type clusters with associated topics and their percentage of total classified reports.

Cluster	Representative topic numbers	GPT Topic Labels	% of total reports
Equipment, Machinery and Lifting Operations	8, 9, 10, 11, 13, 16, 20, 39, 46, 65	Crane Operation and safety, LIDAR Rod Detachment and Dropped Object Risk, Lift and cable equipment inspection and issues, Safety in load lifting operations, warehouse equipment and damage, Davit crane operations, lifting bag equipment, nacelle craning hatch operation, lifting and slings	16%

		safety incidents, lifting and mooring operations safety	
Maritime and Vessel Operations	0, 19, 21, 23, 30, 43, 49, 57, 59, 61, 72, 73, 76, 82	Maritime Vessel and Personnel Transfer Incidents, Maritime Operations, Subsea Equipment Operation and Maintenance, Maritime Safety Operations, Boat Docking Equipment and Safety, Oil Rig Equipment and Operations, Marine Operations and Seasickness, Maritime Accidents and Damages, Buoys Maintenance and Recovery Operations, Maritime Accidents and Investigations, Boat landing and ladder safety, Underwater drilling equipment failures and investigations, Boating safety and equipment review, recovering dropped bags from sea	14%
Mechanical, Hydraulic, and Electrical Systems	1, 24, 34, 38, 41, 60, 62, 66, 67, 71, 78, 84	Oil and hydraulic leak issues, electrical cable safety and identification, remote control and switching of turbines. Electrical shock and circuit safety, UPS Battery issues and maintenance, onshore cable excavation and installation, generator and motor maintenance, switchgear replacement process issues, radio and utility equipment accidents, turbine incidents, cable and tray	11%

		equipment safety, mobile communication interruptions and updates	
Structural and Access Safety	12, 15, 26, 27, 36, 37, 54, 64	Roller shutter doors and safety, ladder safety and fall prevention, lighting and illumination issues, water pipe and drainage works, scaffolding safety and supervision, gate maintenance and safety, nacelle hatch maintenance and safety, handrail and railing damage	10%
Fire, Chemical, and Hazardous Substances	5, 42, 53, 63, 70, 74, 85	Fire safety and prevention, Handling hot substances, hazardous waste management, workplace safety and hazard management, hazardous substance storage and safety, Allergic reactions to substances, Diesel refuelling safety and environmental impact	6%
Tools, Workshop & Maintenance Safety	44, 47, 48, 75, 77, 81, 83	Hardware tools and maintenance, construction excavation process, workplace accidents during installation process, construction and equipment handling incidents, misplaced equipment and tools, inspection and maintenance incidents, industrial equipment maintenance and troubleshooting	6%
Vehicle and Land Related Incidents	2, 18	Vehicle and road safety incidents, Missing or loose vehicle parts	5%

7.4.2 Injury Types and Body Parts Affected

Figure 41 shows a heatmap of body part injuries. Hand-related injuries formed the largest combined category when grouped (hand injuries from cutting tools, Workplace hand injury and maintenance, finger and thumb injuries, and trapped fingers). The majority were cuts, crush injuries, and pinches during tool use, handling sharp materials, or working around moving parts. Several incidents were linked to incorrect tool use or missing safety guards. Simple actions like ensuring gloves match the hazard, improving guarding and reinforcing correct handling techniques could significantly reduce frequency.

Manual work-related back injured were also a common type recorded. These were primarily linked to incorrect lifting techniques, handling loads without mechanical assistance, and working in awkward positions for prolonged periods. Many cases arose during routine maintenance or moving equipment where job planning was limited. In several reports, workers attempted to “push through” discomfort, leading to aggravated strains. Preventative measures including reinforcing correct manual handling training, scheduling adequate manpower, and encouraging the use of lifting aids or mechanical devices.

Knee, ankle and leg injuries were largely the result of slips, trips, and falls. Twisted ankles, bruised knees, ligament strains, and shin impacts were common, with some cases involving fractures. Contributing factors included wet or icy conditions, unsecured staircases, hidden trip hazards, and loss of balance during transfers between vessels and platforms.

Eye-related cases often involved dust, grit, or small debris entering the eye during grinding, drilling, or work in windy conditions. A notable proportion involved inadequate or incorrectly fitted eye protection, with some cases caused by workers removing PPE mid-task. There were also incidents of chemical splashes and impacts from small projectiles.

Head injuries were the least common but still occurred 54 times. These incidents most often involved striking the head on low clearance structures, moving machinery, or protruding fixtures. A recurring pattern was workers removing helmets in tight spaces, resulting in direct contact with hard or sharp surfaces. Even when PPE was worn, poor fit sometimes reduced protection. Other cases involved equipment falling onto head or being hit by swinging objects.

Some topics also referred to incidents but did not specify the injured body part but still provides clear insight into common hazard types and likely injured. These include slips, trips and falls, often leading to strains, sprains, or bruising. In several cases from these topics no injury occurred but the circumstances posed a clear risk. Together, these cases highlight recurring unsafe actions that could affect multiple body parts.

Other workplace health incidents and acute medical responses were also recorded. These ranged from heat exhaustion, dehydration, and dizziness during offshore work to more serious medical emergencies such as suspected stroke, heart attack symptoms, and loss of consciousness. While these incidents were not directly caused by workplace hazards, each required a coordinated emergency response, underlining the importance of early symptom recognition, clear communication protocols, and readiness to act quickly when health issues arise. Details of these topics are shown in Table 3.

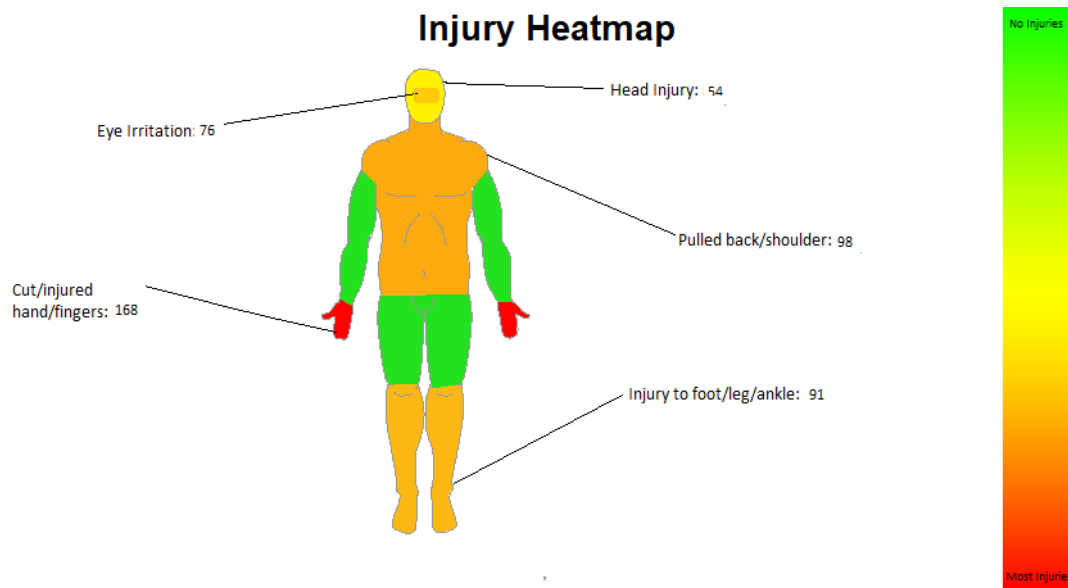


Figure 41. Heatmap of body part injured, with number of injuries to each body part given.

Table 27. Injury type clusters with associated topics and their percentage of total classified reports.

Cluster	Representative topic numbers	GPT Topic Labels	% of total reports
Cut/injured hands and fingers	7, 17, 31, 68	Hand injuries from cutting tools, Workplace hand injury and maintenance incidents, Finger and Thumb Injuries, Accidents involving trapped fingers in doors and Locking Mechanisms	6%
General Injuries	22, 24, 33, 50, 79, 80	Minor Kitchen injuries, Incident reports involving injuries, workplace injuries and accidents, Workplace injuries and safety techniques, Workplace health incidents and response, Insect and fish bites/ stings	5%
Pulled back	3	Manual Work-related Back Injuries	3%

Injury to foot/ leg	28, 40, 58, 86	Workplace Injury: Knee and Ankle Trauma, Ankle injuries and accidents, Leg injuries and equipment, Ankle injuries on staircases	3%
Eye Irritation	6	Eye Irritation and Care	3%
Head Injury	14	Helmet-related head injuries	2%

7.4.3 Human, Environmental and Causal factors

Workplace safety violations were the most common contributing factor, with 98 incidents. Most of these incidents were due to unsafe working at height practices, where workers operated without using fall protection such as harnesses, lanyards, or lifelines. Many cases described personnel climbing structures, walking along beams, or working near unprotected edges at height over two meters without restraint systems in place. Others involved the incorrect anchoring of harnesses, use of damaged or expired personal fall protection equipment, or lack of secondary fall arrest systems. A number of these cases also highlighted training and competency issues, where workers were either not qualified for the task or were provided with incorrect or insufficient PPE during working-at height activities.

Workplace safety equipment use was the next most significant with 36 instances. These cases often concerned personnel not wearing the correct personal protective equipment for the task or location. Common examples included working without safety glasses, gloves, hard hats, or high-visibility clothing in operational areas, and entering restricted zones without mandatory PPE. There were several instances where workers used incorrect or expired PPE, such as out-of-date harnesses or non-compliant safety gear.

Work permit and contractor document failures were another significant category. In several cases, contractors were operating near high-risk environment such as live electric systems, overhead power lines, or confined space without procedural changes that left staff unaware of new requirements.

Instances were also recorded where safety documents, such as access certificates, were mishandled or incorrectly signed, potentially undermining procedural integrity.

Another notable cause category is workplace fatigue and stress management. Several incidents involved workers suffering from heat stress or dizziness after working in unshaded, hot environments without adequate breaks. These findings are consistent with previous research demonstrating that offshore wind personnel experience considerable occupational strain, including fatigue, stress, and physical health impairments associated with demanding working environments and high workloads (Mette et al., 2018). There were also incidents where employees lost consciousness, experienced chest pains, or required emergency medical attention due to work-related health issues. Long-term strain was also reported, with employees developing musculoskeletal pain, headaches, and fatigue from working in inadequate conditions over extended periods. Several cases highlighted high workloads, insufficient staffing, and poor work-life balance as significant stressors.

Poor housekeeping and cleanliness were also identified in 20 incidents, many involving untidy or unsafe work environments that posed potential hazards. These ranged from trip hazards caused by equipment, tools, and cleaning materials left lying around. Several welfare areas, including kitchens, changing rooms, and drying rooms, were reported in unsanitary or disorganised states, contributing to poor hygiene standards. Some reports specifically attributed poor housekeeping standards to insufficient cleaning resources or delayed cleaning activities. Details of these topics are shown in Table 28.

Table 28. Causal factors clusters with associated topics and their percentage of total classified reports.

Cluster	Representative topic numbers	GPT Topic Labels	% of total reports

Safety violations	4, 29	Workplace safety violations, workplace safety equipment use	5%
Weather-related incidents	25, 55	Ice and slip accidents, lightning strike risk monitoring	2%
Contractor Permits	32	Work permit and contractor documentation	1%
Fatigue and stress management	51	Workplace fatigue and stress management	1%
Cleaning and housekeeping issues	52	Poor housekeeping and cleanliness	1%

Environmental conditions were also found to contribute to several incidents, with the most common being ice and slip accidents. These accounted for 37 recorded incidents, ranging from minor slips in work areas to serious falls resulting in fractures. These occurred across diverse locations such as turbine platforms, quaysides, substations, walkways, and vessel decks. Other than icy surfaces, some cases also involved slippery conditions from wet weather.

Incidents involving lightning strike risk often involved failures or delays in detection and communication systems, such as Metrogroup or StormGeo, arriving late or not triggering at all. Several reports described lightning occurring in close proximity to working personnel, including a direct turbine strike with technicians on-site. In some cases, procedures for stripping work were unclear or inconsistently followed, with disputes between vessel crews and control rooms over when to halt operations.

Although not caused by an environmental factor, diesel refuelling can cause environmental impact. These incidents mainly involved leaks, spills, and exposure risks during vessel and generator refuelling operations. Common reasons were equipment failures such as burst hoses, damaged nozzles and

faulty pipe connections as well as procedural lapses. Several of these cases resulted in minor environmental releases, typically contained on deck, through some small quantities entered the water.

Collectively, these causal factors provide a structured means of identifying recurring weaknesses within offshore wind operations that extend beyond individual incidents. Rather than treating safety violations, fatigue, permit failures, or housekeeping deficiencies as isolated events, organisations can use these patterns to prioritise targeted interventions and investigate the underlying organisational conditions that give rise to them. For example, recurring working-at-height violations may prompt reviews of competency assurance, supervision, or permit-to-work systems, while repeated fatigue-related reports may justify further investigation into staffing levels, shift patterns, and work scheduling. By identifying these recurring themes across large incident datasets, organisations can prioritise safety improvements based on evidence rather than isolated events, supporting more proactive risk management.

There were also 12 cases of insect and marine animal bites/stings. These incidents occurred across both onshore and offshore work environments, including construction sites, crew transfer vessels, and offshore operational activities. They were predominantly caused by ticks, wasps, bees, spiders, jellyfish, and weaver fish. Most of these result in minor discomfort or swelling, though some involved individuals with known allergies had increasing severity.

7.5 Association Rule Mining Results

Association rule mining identified several strong links between incident topics and their reported consequences or work contexts; the highest confidence rules are shown in Table 5. For example, workplace safety violations were almost always linked with hazard-level outcomes (confidence=0.94). They also had a strong link to high-potential incidents (confidence=0.73). The very high lift with working-at-heights tasks (lift=14.03) suggests that procedural breaches in elevated environments disproportionately elevate risk.

Lighting deficiencies and PPE use failures, often overlooked in offshore safety discourse, also showed strong predictive links with hazard outcomes. Injury-specific topics, such as eye irritation, back strain, and head impacts, were consistently mapped to first-aid outcomes, reinforcing the practical accuracy of the extracted rules. Importantly, while ARM uncovers co-occurrence patterns, these should not be interpreted as causal mechanisms; rather, they highlight prevention priorities. In particular, rules with high lift values point to combinations of context and behaviour that warrant targeted interventions in offshore wind safety management.

While Table 29 highlights the strongest individual rules, Figure 42 shows the broader structure of associations across topics, consequences, and work contexts, including secondary relationships not captured in the table. Hand injuries from cutting tools and workplace safety violations were both linked with construction sites, suggesting that site-specific conditions may amplify risks in these categories. Manual work-related back injuries were closely connected with manual handling activities, reflecting the ergonomic demands of routine offshore tasks. Vehicle and road safety incidents were associated with business travel, while fire safety prevention clustered with oil and hydraulic leak issues. These relationships illustrate how ARM can uncover contextual links between seemingly distinct topics, providing a richer picture of the work processes, environments and consequences of incidents.

Table 29. ARM results with topic, linked attribute and metrics.

Antecedent (Topic)	Consequent (Linked Attribute)	Support	Confidence	Lift
Workplace Safety Violations	Consequence = Hazard	0.0170	0.94	2.91
Workplace Safety Violations	High Potential Incident = Yes	0.0133	0.73	3.29
Workplace Safety Violations	Work Process = Working at Heights	0.0131	0.72	14.03
Lighting & Illumination Issues	Consequence = Hazard	0.0061	0.89	2.77
Safety Equipment Use Issues	Consequence = Hazard	0.0056	0.83	2.59

Safety in Load Lifting Operations	Consequence = Hazard	0.0081	0.76	2.35
Safety in Load Lifting Operations	Dropped Object = Yes	0.0048	0.47	2.73
Mast Climbing & Harness Use	Dropped Object = Yes	0.0043	0.47	2.71
Eye Irritation & Care	Consequence = First-aid Injury	0.0109	0.78	3.55
Manual Work-related Back Injuries	Consequence = First-aid Injury	0.0104	0.57	2.61
Helmet-related Head Injuries	Consequence = First-aid Injury	0.0070	0.70	3.21
Hand Injuries from Cutting Tools	Consequence = First-aid Injury	0.0090	0.67	3.07

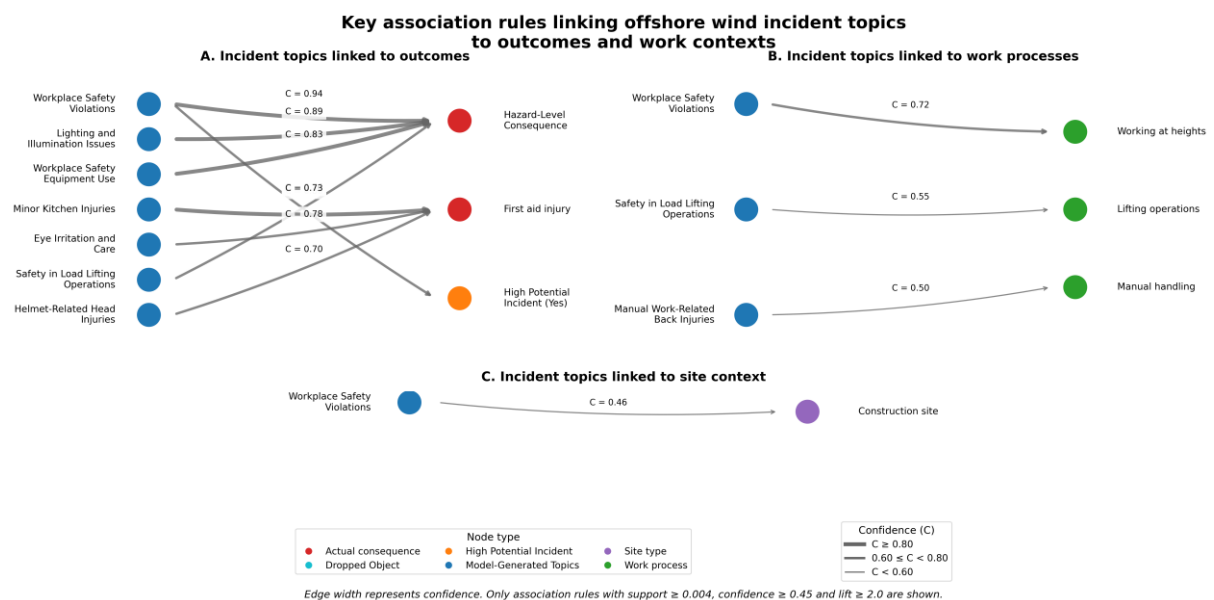


Figure 42. ARM network of offshore wind incidents.

7.6 Expert Validation

To assess the interpretability of the BERTopic clusters, four domain experts independently evaluated topic quality across three dimensions: keyword coherence, GPT-generated label fit, and document relevance. A common subset of 25 topics was evaluated by all four experts to enable assessment of inter-rater agreement and reliability. The remaining topics were distributed among the experts, allowing the complete set of identified topics to be evaluated. Agreement and reliability were assessed using multiple complementary statistical measures, while descriptive statistics were calculated across the full validation dataset.

Inter-rater analysis of the 25 commonly evaluated topics indicated variation in the extent to which experts converged across the three evaluation dimensions. Kendall's W was highest for GPT-generated label fit ($W = 0.607$), compared with keyword coherence ($W = 0.428$) and document relevance ($W = 0.424$). This suggests that the experts demonstrated greater convergence in their relative assessment of the suitability of the GPT-generated topic labels than in their assessment of keyword coherence or the relevance of the associated documents. Nevertheless, the results indicate that expert interpretation was not uniform across all dimensions. Given that incident narratives can contain overlapping operational, technical, and contextual elements, some variation in semantic interpretation was expected. This also reflects a broader challenge in incident analysis, whereby experienced practitioners may differ in how they interpret the boundaries and meaning of themes emerging from complex operational narratives.

Cronbach's α was calculated separately for each expert using all topics evaluated by that expert. Internal consistency was good for Expert 1 ($\alpha = 0.826$), acceptable for Expert 2 ($\alpha = 0.768$), and good for Expert 4 ($\alpha = 0.799$). Expert 3 demonstrated substantially lower internal consistency ($\alpha = 0.497$). Thus, three of the four experts demonstrated acceptable-to-good consistency across the three evaluation dimensions, while the ratings provided by Expert 3 were less internally consistent. This

variation indicates that the dimensions of keyword coherence, label fit, and document relevance were not necessarily interpreted or scored in the same way by all reviewers.

Intraclass correlation coefficients (Table 30) provided further evidence of differences between the three evaluation dimensions. GPT-generated label fit achieved moderate agreement ($ICC(2,k) = 0.701$, $p < 0.001$), representing the strongest inter-rater reliability of the three measures. Document relevance demonstrated poor agreement ($ICC(2,k) = 0.426$, $p = 0.003$), while keyword coherence also demonstrated poor agreement ($ICC(2,k) = 0.311$, $p = 0.029$). These results suggest that the experts were considerably more consistent in judging whether the GPT-generated labels appropriately represented the identified topics than in evaluating the coherence of the underlying keyword sets or the relevance of the associated documents.

Table 30. ICC scores per evaluation dimension.

Dimension	ICC(2,k)	Interpretation	p-value
Keyword Coherence	0.311	Poor agreement	0.02925
GPT Name Fit	0.701	Moderate Agreement	<0.001
Document Relevance	0.426	Poor Agreement	0.00328

Despite variation in inter-rater agreement, mean ratings across the complete validation dataset were positive for all three dimensions. Keyword coherence and GPT-generated label fit both achieved an overall mean score of 4.00, while document relevance achieved a slightly lower mean of 3.94. On a five-point evaluation scale, these results indicate that the topics were generally regarded positively by the expert reviewers. Importantly, however, the positive mean scores should be distinguished from the agreement statistics: high average ratings indicate that the experts generally considered the outputs meaningful, whereas the ICC and Kendall's W results indicate that they did not always agree on the relative quality of individual topics.

Table 31. Mean expert scores across topic validation dimensions.

Expert	Mean Keyword Coherence	Mean GPT Label fit	Mean label-document fit
Expert 1	4.610	4.439	4.220
Expert 2	4.171	3.512	3.805
Expert 3	3.2	3.5	3.125
Expert 4	4	4.55	4.6
Average	4.00	4.00	3.94

Differences between individual expert scores provide further context for the agreement results, shown in Table 31. Experts 1 and 4 generally provided higher ratings across the validation dimensions, whereas Expert 3 applied more conservative ratings, particularly for keyword coherence and document relevance. Expert 2 also rated GPT-generated label fit more critically than Experts 1 and 4. These differences in rating tendencies help explain why positive overall evaluations can coexist with comparatively weak agreement for some dimensions. In other words, the experts generally considered the identified topics to be meaningful, but differed in the extent to which individual topics satisfied each validation criterion.

Overall, the findings support the use of the extracted topics as a plausible and operationally useful representation of the underlying incident reports, while also demonstrating that topic interpretation remains partly dependent on expert judgement. The combination of positive overall ratings and variable inter-rater agreement is therefore not necessarily contradictory: the experts generally regarded the outputs favourably, but differed in their assessments of the relative quality of individual topics. This uncertainty should be recognised when interpreting individual clusters and reinforces the value of retaining expert judgement alongside automated topic modelling rather than treating the generated topic structure as an unambiguous classification of the underlying incidents.

7.7 Discussion of Topic Modelling Findings and ARM

Overall, the case study identified 85 distinct topics, where the majority were linked to incident types, and some were linked to injuries and causal factors. These topics revealed patterns of how offshore wind incidents occur and the conditions in which they arise. The most common incident type was found to be equipment, machinery, and lifting operations, making this the largest risk category overall. These incidents frequently arose from crane use, load handling, and failures of lifting gear, with common themes including dropped objects, sling malfunctions, and unsafe mooring or lifting practices. This trend is strongly echoed in the G+ reports with lifting and manual handling as some of the leading incident-causing work process categories (G+, 2022, G+, 2021, G+, 2020, G+, 2019). Similar patterns have been reported in offshore construction studies, where lifting activities are consistently flagged as among the highest-risk operations (Gurdal Ertek and Kailas, 2021). Inspection of lifting gear, reinforcement of safe-load-handling training, and the use of secondary securing systems are important to ensure a safe environment for all workers.

Importantly, these findings extend beyond the information available from the structured G+ classifications. While the statistical analysis identified lifting-related incidents as a common work process, BERTopic automatically separated these events into distinct operational themes including crane operations, load handling, sling failures, mooring practices, and dropped-object scenarios. This provides a considerably richer understanding of the specific activities contributing to risk, allowing safety interventions to be targeted at particular operational tasks rather than broad incident categories.

The structured statistical analysis undertaken earlier in this chapter identified broad patterns within the offshore wind dataset, such as the greater likelihood of high-potential incidents during construction and the strong association between dropped object incidents and high-potential outcomes. However, these analyses were necessarily limited to the predefined variables contained within the G+ database. While they identified where elevated risks existed, they provided little insight

into the operational activities, behaviours, or circumstances responsible for these patterns. The application of BERTopic therefore represents a substantial enhancement by automatically extracting these contextual details directly from the narrative reports.

Additionally, the harsh marine environment significantly increases the likelihood of equipment failure in offshore wind farms. Conditions such as high humidity, salt corrosion, and continuous exposure to turbulent sea states contribute to these failures. Consequently, the operational environment leads to higher maintenance costs, largely due to unexpected equipment breakdowns that necessitate corrective maintenance (Butler et al., 2013, Kou et al., 2022). This highlights the need for robust monitoring systems that can facilitate early detection and proactive response to potential failures.

The second most common incidents stemmed from maritime and vessel operations, particularly during personnel transfers, docking and subsea work. The G+ reports confirm this trend and in recent years, incidents on vessels have increased year on year from 2019 to 2022. This upward trend then stabilised to around one-third of all incidents involving CTVs (G+, 2024, G+, 2022, G+, 2021, G+, 2020). Prior research also highlights vessel transfers as one of the most hazardous phases of offshore work. Seyr and Muskulus (2016) found that incidents involving vessel mobilisation and transfers accounted for over 20% of all reported incidents in both 2013 and 2014. Safer landing platforms, improved vessel design, and clear stop-work protocols in poor weather are essential to reduce risk in this area.

Similarly, the topic modelling distinguished between multiple forms of maritime activity, including personnel transfers, docking operations, and subsea vessel work. These operational contexts are not explicitly represented within the structured incident database, demonstrating how analysis of the narrative reports provides substantially greater operational detail than conventional statistical summaries.

Failures in mechanical, hydraulic, and electrical systems were also frequent, involving oil leaks, cable faults, and turbine malfunctions. These failures not only create operational downtime but also pose safety and environmental risks. Again, G+ reports similarly highlight such failures, with electrical faults,

hydraulic leaks, and turbine component breakdowns repeatedly identified as contributing factors (G+, 2022, G+, 2021, G+, 2020, G+, 2019). Chen et al. (2020b) and Lu et al. (2021) note that the offshore environment magnifies these challenges due to restricted access and weather-related delays.

The most frequent injuries were cuts to hands and fingers, followed by pulled back and shoulder, and sprained or twisted ankles. This pattern aligns closely with the G+ (2024) report, which found feet/ankles, fingers/thumbs, and shoulders among the most frequently injured body parts, with sprains/strains the leading injury type. Additionally, Dethleff et al. (2016) found that most medical treatments involved traumatic injuries, primarily lacerations, cuts, and contusions, with 59% affecting the limbs. Similarly, Velasco Garrido et al. (2018) reported that strains from climbing and heavy lifting were highly prevalent among offshore workers, which may explain the frequent occurrence of back injuries and sprained ankles. These findings highlight the physical demands of offshore work and suggest a need for better acute medical responses such as heat stress, dehydration, and cardiovascular symptoms. ergonomic practices, enhanced protective equipment, and training that focuses on injury prevention during routine activities like climbing. Beyond traumatic injuries, a notable number of cases involved acute medical conditions such as heat stress, dehydration and cardiovascular symptoms. While not always caused directly by workplace hazards, these incidents highlight the importance of health monitoring, early symptom recognition, and robust emergency response protocols offshore. A significant finding was that workplace safety violations emerged as a major contributing factor in many incidents. The majority of these involved unsafe working-at-height practices, such as climbing structures or moving near unprotected edges without harnesses, fall arrest lanyards or safety lines. Such deliberate disregard frequently stems from deeper issues related to safety culture, such as insufficient training or a lack of understanding about the importance of these protocols. Lin and Lu (2023) highlight that weaknesses in safety culture and inadequate safety education are key risk factors in offshore wind operations, aligning with our findings. In addition, Shafiee and Adedipe (2022) suggest that fatigue and poorly designed safety protocols often exacerbate human errors. These errors are particularly problematic during maintenance and decommissioning activities, where they can lead to

significant operational risks. Addressing these underlying issues requires strengthening safety culture, enhancing worker education, and implementing fatigue management strategies to minimise the potential for error and improve overall safety. Additionally, environmental hazards such as ice formation and lightning risk also played a recurring role in incidents. These hazards, often outside human control, underline the need for reliable monitoring systems, clear stop-work protocols, and consistent communication between control rooms and vessel crews.

These findings also demonstrate the value of analysing unstructured narrative data. While structured incident databases may record the consequence or work process involved, they rarely capture behavioural factors such as unsafe shortcuts, failures to wear fall-protection equipment, or procedural non-compliance. Topic modelling automatically extracted these recurring behavioural themes across thousands of reports, enabling systematic identification of organisational and human-factor issues that would otherwise require extensive manual review.

While topic modelling identified the dominant operational themes present within the offshore wind incident narratives, association rule mining extended the analysis by quantifying how these themes were related to one another and to incident outcomes. This moved the analysis beyond identifying common incident types towards revealing statistically significant relationships between operational activities, injury mechanisms, and consequence severity. The association rule analysis highlights distinct clusters of incident-consequence relationships with clear safety implications. Workplace safety violations emerged as the strongest predictor of both hazard-level consequences and high-potential incidents, particularly when linked to working at heights tasks. This suggests that breaches of procedural safety in elevated-risk contexts may be a key driver of severe outcomes, underscoring the importance of fall-protection systems, permit-to-work enforcement and competency assurance.

Technical failures, including lighting, illumination issues, and safety equipment use lapses, were also strongly predictive of hazard outcomes. These findings point to the critical role of basic infrastructure reliability and PPE compliance in preventing severe accidents. Lifting operations displayed strong

associations both with hazard outcomes and dropped-object incidents, reinforcing industry awareness of lifting as a high-risk activity and highlighting the need for enhanced secondary retention systems and lifting protocols.

In contrast, injury-specific topics such as eye irritation, back strain, head impacts, and hand injuries from cutting tools were consistently linked with first-aid level consequences. While these incidents are typically of lower severity, their frequency suggests a significant cumulative impact on operational continuity and workforce health. Targeted training and ergonomic interventions could therefore yield substantial safety improvements.

Overall, the combined application of descriptive statistics, topic modelling, and association rule mining demonstrates the complementary value of each analytical stage. The structured statistical analysis provided an overview of when and where incidents occurred and their recorded outcomes, whereas BERTopic automatically extracted detailed operational themes relating to work activities, equipment, injury mechanisms, environmental conditions, and causal factors from the incident narratives. Association rule mining then quantified the relationships between these themes and incident consequences, identifying combinations of factors associated with higher-severity events. Together, these methods provide a substantially richer understanding of offshore wind safety than conventional statistical analysis alone and demonstrate how AI-assisted analysis can support more targeted, evidence-based safety interventions.

7.8 Chapter Summary

This chapter demonstrated the successful application of the low-information pathway of the proposed analytical framework to offshore wind incident reports, illustrating its transferability beyond the maritime domain. By combining descriptive statistical analysis with BERTopic, large language model-assisted topic interpretation, and association rule mining, the framework transformed unstructured incident narratives into structured safety intelligence. The analysis identified dominant operational activities, injury mechanisms, environmental conditions, and causal factors associated with offshore

wind incidents, while also revealing statistically significant relationships between incident themes and safety outcomes. These findings provided a considerably richer understanding of offshore wind safety than could be obtained from the structured incident database alone.

The results further demonstrated that the proposed framework is capable of analysing heterogeneous safety datasets across different safety-critical industries while maintaining consistent analytical outputs. By automatically extracting operational themes and contextual information from thousands of short narrative reports, the framework supports more scalable, evidence-based approaches to safety analysis and organisational learning. Together with the maritime case studies presented in Chapters 7 and 8, this chapter demonstrates the flexibility, scalability, and practical applicability of the integrated AI-based framework across multiple data types and operational domains.

Having established and evaluated both the high-information and low-information pathways of the proposed framework, the following chapter synthesises the findings from all three application chapters. It discusses the theoretical, methodological, and practical contributions of the research, evaluates the strengths and limitations of the proposed framework, and considers its implications for future maritime and offshore safety management.

8. Discussion

8.1 Chapter Overview

This chapter synthesises the findings of the three case studies and evaluates their contribution to the overall research aims and objectives. It first provides an integrated overview of the research contributions, with particular emphasis on the SHIELD-NLP case study. The chapter then discusses methodological and practical limitations, before outlining key directions for future research and application. The discussion draws on insights from all three case studies.

8.2 Research Overview

This research set out to investigate whether modern natural language processing and machine learning techniques can be used to automate the classification of human and causal factors from unstructured safety investigation text. Specifically, the work sought to establish whether a common analytical framework could be developed to evaluate both traditional long-form Safety-I accident investigation reports and the more emerging Safety-II forms of reporting, such as incidents, near misses, and prevention-focused narratives.

Unstructured safety text remains one of the most information-rich yet under-utilised sources of safety intelligence within maritime operations. This research, therefore, addresses a fundamental challenge in modern safety management: how to create a unified, scalable framework for extracting structured safety knowledge from unstructured narratives.

A central novelty of this research is the development of a unified AI-based analytical framework capable of operating across fundamentally different forms of safety reporting, from detailed accident investigation reports to short incident and near-miss narratives. Previous studies have typically focused either on supervised classification of accident reports or on unsupervised exploration of incident narratives in isolation. This research, to the author's knowledge, is the first to integrate supervised learning, transformer-based topic modelling, large-language model-assisted interpretation,

association rule mining, and human-in-the-loop validation within a single framework applicable across multiple safety-critical reporting contexts.

The first case study established the feasibility of applying NLP techniques to structured Safety-I accident investigation data. The SHIELD-NLP case study then provided the most comprehensive validation of the framework using 330 full maritime accident investigations, transforming a mature human factors taxonomy into a machine-learning-ready classification system. This demonstrated that detailed causal reasoning typically performed by expert analysts can be partially automated with high reliability for key human and organisational factors.

The later case studies extended this approach beyond traditional accident investigations into the Safety-II domain, where reporting documents on everyday operational disturbances, near misses, and successful recoveries. This demonstrated how advanced Natural Language Processing (NLP) and Machine Learning techniques can generate actionable insights from short, unstructured maritime and offshore incident narratives. Across the two case studies, CalMac ferry near-miss reports and G+ offshore wind incident summaries, the research showed that transformer-based topic modelling (BERTopic), when combined with GPT-assisted topic labelling, offers a scalable and interpretable way to extract incident themes, human and causal factors, and temporal safety patterns from free-text descriptions.

The ferry near-miss study produced over 75 incident themes and 87 prevention themes, revealing seasonal patterns (e.g. increases in passenger-related risks during summer and mechanical failures during winter) and identifying clear incident-prevention relationships through Association Rule Mining (ARM). This confirmed that near-miss data can provide proactive insights into operational risk, particularly when structured taxonomies are absent or inconsistently applied.

The offshore wind study demonstrated similar advantages at industry scale. From 5403 short incident descriptions, BERTopic identified a wide range of operational themes, for example, equipment and lifting operations, CTV collisions, hydraulic/electrical failures, and linked them to injury types and

human factors such as procedural violations and unsafe working-at-height practices. ARM further revealed strong patterns between hazards, outcomes, and high-risk operational contexts.

Together, the studies show that unsupervised learning with LLM interpretation and rule-based reasoning offers a novel and powerful analytical framework for large safety datasets. Importantly, the methods were applied successfully to two different sectors (ferry and offshore wind), demonstrating the generalisability and operational usefulness of the proposed methodology.

8.3 Research Novelty and Contributions

This thesis contributes to the fields of maritime and offshore wind safety through methodological, technical, and practical advances. While individual artificial intelligence techniques such as transformer-based language models, topic modelling, and association rule mining have previously been applied independently within safety research, this thesis demonstrates how these methods can be integrated into a unified analytical framework capable of extracting structured safety knowledge from multiple forms of unstructured incident reporting. The principal contributions of this research are discussed below.

1. Development of a unified AI framework for safety intelligence

This thesis introduces a unified AI-driven framework that integrates supervised learning, transformer-based topic modelling, large language model-assisted interpretation, and association rule mining within a single workflow. Unlike existing approaches that apply these techniques independently, the proposed framework enables scalable extraction of actionable safety knowledge from diverse forms of unstructured incident reporting.

2. Machine-learning adaptation of the SHIELD taxonomy

The research developed SHIELD-NLP, a machine-learning-ready adaptation of SHIELD taxonomy, overcoming one of the principal barriers preventing supervised learning using maritime human-factors taxonomies.

3. Development of a maritime domain-specific language model

This thesis introduces SeaSafeBERT, a maritime domain-adapted transformer model trained using accident investigation reports to improve understanding of technical maritime terminology and reporting language. Combined with a hierarchical human-in-the-loop classification framework, it enables reliable and explainable identification of contributory factors from unstructured accident narratives.

4. Automated extraction of operational safety knowledge from near-miss reports

This research demonstrated how transformer-based topic modelling, GPT-assisted interpretation, and association rule mining can be integrated to automatically identify operational themes, human factors, temporal trends, and preventative measures from short near-miss reports. This extends analysis beyond simple topic discovery towards actionable organisational learning from positive actions preventing accidents.

5. Demonstration of framework transferability across safety-critical domains

The proposed framework is successfully applied across three distinct reporting contexts; maritime accident investigations, ferry near misses, and offshore wind incident reports, demonstrating its adaptability to differing reports styles, data quality, and operational environments. This highlights the potential for wider application across safety-critical industries beyond traditional maritime accident analysis.

6. Practical decision-support from proactive safety management

Beyond methodological advances, the research demonstrates how AI can augment rather than replace expert judgement through human-in-the-loop decision support. The resulting framework provides a practical foundation for faster incident analysis, more consistent safety coding, improved identification of emerging risks, and evidence-based safety interventions.

The findings have several practical implications for maritime and offshore safety management. The principal value of the proposed framework is not the replacement of safety professionals, but the ability to increase the scale and consistency with which narrative safety information can be reviewed. For operators receiving hundreds or thousands of incidents and near-miss reports, the framework could provide an initial analytical layer that identifies recurring themes, human and organisational factors, emerging patterns and relationships between incidents and preventive actions for subsequent expert review. This could allow specialist investigation resources to be directed towards recurrent or potentially high-consequence patterns rather than relying exclusively on manual review of individual reports.

The case studies illustrate how these outputs could inform specific safety-management activities. SHIELD-NLP classifications could support more consistent coding and monitoring of recurring human and organisational factors across accident investigations. Topic and temporal analysis of near-miss reports could support periodic safety reviews by identifying emerging or seasonal operational risks, while associations between incident and prevention themes could help identify preventive actions repeatedly associated with successful recovery. In the offshore-wind case study, for example, the observed associations between workplace safety violations, high-potential incidents and working-at-height activities indicate areas in which organisations could prioritise reviews of working-at-height procedures, competency assurance, supervision and permit-to-work arrangements. These examples demonstrate how the analytical outputs can inform targeted safety assurance and intervention activities rather than functioning solely as descriptive statistics.

A further practical implication concerns the design of safety-reporting systems themselves. The short-text case studies demonstrated that analytical performance is constrained when reports contain only minimal descriptions or inconsistently record events, consequences, causal factors and preventive actions. The findings therefore support greater standardisation of incident and near-miss reporting at the point of capture. A practical reporting structure could combine concise free-text narratives with

structured prompts covering: (1) what happened; (2) operational context; (3) immediate and underlying contributory factors; (4) actual or potential consequences; (5) barriers that failed; (6) actions that prevented escalation; and (7) corrective or preventive actions. Retaining free text alongside these structured fields would preserve contextual information while producing more consistent data for both human review and automated analysis.

At policy and regulatory level, the findings suggest that the effectiveness of AI-enabled safety analysis depends not only on advances in modelling, but also on the consistency and quality of the underlying safety data. Regulators and industry bodies could therefore support future analytical capability by promoting greater harmonisation of incident and near-miss reporting structures, including consistent terminology for human factors, operational context and barrier performance. Common minimum reporting fields could facilitate comparison across operators and jurisdictions while improving the quality of data available for both conventional safety analysis and emerging AI-assisted methods.

Aggregated analytical outputs may also provide value beyond individual operators. Subject to appropriate governance, confidentiality and validation arrangements, recurring themes and emerging patterns identified across organisations could support industry-level benchmarking, identification of weak signals, and prioritisation of safety campaigns or guidance. However, the limitations identified in this research, including dataset representativeness, temporal bias and the interpretive nature of expert-generated labels, mean that automated outputs should not be treated as objective evidence of causality or used as substitutes for formal investigation. Their most appropriate role is as decision support, helping investigators, operators and regulatory bodies identify patterns requiring further expert examination.

Collectively, these contributions extend existing research beyond isolated applications of machine learning within maritime safety. Rather than proposing a single algorithm or analytical technique, this thesis demonstrates a unified framework that combines AI, expert knowledge, and statistical analysis to transform large-scale unstructured safety reports into actionable organisational learning. The

framework is shown to be transferable across multiple safety-reporting contexts, providing both methodological innovation and practical value for proactive safety management.

8.4 Achievement of Research Objectives

The research objectives defined in Chapter 4 have been successfully achieved through the methodological development and empirical analyses presented across the three case studies. Each objective is addressed in relation to the corresponding outcomes of the research.

The objective of reviewing existing NLP and machine learning techniques used in safety-critical domains was achieved through a focused methodological review of 188 relevant studies, which identified the most effective approaches for analysing unstructured safety text. For unsupervised analysis of short incident and near-miss reports, BERTopic emerged as the most suitable method due to its use of transformer-based embeddings and superior topic coherence compared to traditional models such as LDA. For supervised classification of labelled accident data, domain-adapted BERT models consistently outperform generic embeddings and traditional baseline, particularly when integrated into downstream classifiers. The review also identified the growing role of large language models (LLMs) in supporting automated interpretation and extraction, and the importance of human validation and decision-support system integration for operational deployment. These findings then directly informed the methodological framework and selection of approaches.

The objective of developing a unified AI-based methodological framework applicable to both short incident narratives and full accident investigation reports was fully achieved. A general framework integrating causal sentence extraction, supervised multi-label classification, unsupervised topic modelling, large language model-assisted interpretation, association rule mining, and human-in-the-loop validation was developed. This framework was demonstrated to operate effectively across fundamentally different reporting formats, including detailed maritime accident investigations, ferry near-miss narratives, and offshore wind incident descriptions. The successful application of the same analytical structure across these data sources confirms that a unified framework was achieved.

The objective of developing a human factors taxonomy for use within an NLP and ML framework was achieved through the systematic development of an NLP-ready version of SHIELD taxonomy (SHIELD-NLP). The original SHIELD taxonomy comprised 99 highly granular human factors codes, many of which exhibited extreme data sparsity and were therefore unsuitable for supervised machine learning. A structured consolidation and refinement process was undertaken to merge conceptually aligned low-frequency codes while preserving the theoretical intent of the original taxonomy. This resulted in a machine-learning-ready taxonomy of 53 SHIELD-NLO categories, with sufficient representation to support robust supervised modelling.

The objective of applying the supervised classification pipeline to the SHIELD dataset and evaluating its performance was fully achieved through the SHIELD-NLP case study. The SHIELD-NLP dataset of 330 incidents was used to train and evaluate a two-stage hierarchical classification pipeline. Strong classification performance was achieved for key causal factors such as decision-making, communication, supervision, fatigue, perception, and risk assessment. Furthermore, the trained models were deployed within a web-based human-in-the-loop decision-support tool, demonstrating that the method is not only effective in offline evaluation but also feasible for operational use.

The objective of applying an unsupervised analytical pipeline to ferry near-miss data to evaluate feasibility and to understand both the near-miss events and the preventions that helped stop incidents from developing was also fully achieved. A large dataset of 4,360 ferry near-miss reports were analysed using BERTopic, resulting in extraction of 75 coherent near-miss topics. In addition, 87 prevention topics were generated from prevention narrative fields. GPT-assisted topic labelling and expert validation supported the interpretability of the results at scale, while ARM enabled the identification of statistically significant relationships between near-miss event types and corresponding prevention measures. This case study demonstrated that AI-based methods can be used not only to analyse failures, but also to systematically extract insights into successful safety barrier functioning.

The objective of demonstrating the transferability of the proposed framework to an alternative safety-critical operational domain was achieved through the offshore wind case study. The same unsupervised analytical pipeline was successfully applied to a dataset of 5,403 offshore wind incident reports. Topic modelling, GPT-assisted interpretation, and ARM revealed dominant offshore operational risk patterns, common injury mechanisms, and recurring procedural and safety violations. Although offshore wind places greater emphasis on occupational safety than the preceding maritime case studies, it also incorporates vessel operations, marine logistics, and offshore working environments. The successful application of the framework within this hybrid domain demonstrates that the proposed methodology is robust to differing reporting styles and operational contexts, providing evidence of its transferability beyond traditional maritime accident investigations.

Together, the successful completion of all six objectives confirms that the overarching aim of the research, to develop and evaluate a unified AI-based framework for automated analysis of maritime and offshore safety reports, has been fully achieved. The framework was shown to support consistent analysis across the full spectrum of report types, from detailed accident investigation reports to short near-miss and prevention-focused narratives, thereby enabling continuous, data-driven safety learning across multiple safety-critical domains.

8.5 Limitations

Despite the strong results achieved across the three case studies, several important methodological limitations should be acknowledged.

A primary limitation of the SHIELD-NLP case study related to data size and class imbalance. Although the SHIELD dataset is large by human factors research standards, with 330 accident investigations and 3,543 annotations, it remains small for deep learning in a multi-label setting. Even after consolidation into the SHIELD-NLP taxonomy, several causal codes retained very low support in the test set. This resulted in unstable performance estimates for rare classes and, in some cases, zero F1-scores at the fixed decision threshold. While ROC-AUC values indicated latent discriminative potential for several of

these low-frequency categories, their practical learnability remains constrained by data scarcity rather than model capacity.

A related limitation concerns the representativeness of accident types within the SHIELD dataset. Although EMSA statistics identify loss-of-control events among the most frequently reported maritime casualty and incident categories, no accidents within the SHIELD dataset were classified within this accident type. Consequently, the SHIELD-NLP evaluation cannot establish whether the identified human-factor patterns or classification performance generalise to loss-of-control events. More broadly, the distribution of accident types within the training data may influence which causal relationships are learned by the models, meaning that the results should not be interpreted as representative of the full European maritime accident population. Future validation should therefore incorporate datasets with broader and more balanced coverage of accident types, including loss-of-control events.

A further limitation concerns the nature of the ground-truth labels used to train and evaluate SHIELD-NLP. The SHIELD annotations represent expert interpretations of accident investigation evidence rather than an objective causal truth. These interpretations are inevitably influenced by assessor judgement, the information available within the original investigation, organisational and regulatory context, and the causal assumptions embedded within the SHIELD taxonomy. Consequently, the supervised models learn to reproduce patterns of expert reasoning represented in the training data rather than independently determining the true causes of an accident. High agreement with the labelled data should therefore be interpreted as successful reproduction of the expert-coded taxonomy, rather than proof that the model has established causal truth. Human oversight remains necessary when model outputs are used to support safety investigation or decision-making. The dataset may additionally contain temporal, or period, bias due to the dataset spanning around three decades. Accident investigation practices, regulatory priorities, safety terminology, technology and prevailing understandings of human factors evolve over time. Models trained on historical investigation reports

may therefore learn relationships and linguistic patterns characteristic of the period represented in the training corpus that do not fully reflect contemporary operations or investigative practice. This is particularly relevant where emerging technologies, vessel systems or operational practices introduce hazards and terminology that are poorly represented in historical data. Periodic retraining and validation against more recent accident data would therefore be required before operational application.

The SHIELD-NLP models are also strongly dependent on the quality and consistency of accident investigation reports. Investigation reports vary substantially in depth, narrative style, and causal explicitness across authorities and jurisdictions. Some reports provide detailed human factors reasoning, while others emphasise regulatory compliance or technical failure. These differences affect both the reliability of ground truth labels and the linguistic learnability of several higher-level organisational factors.

Within short-text analysis of offshore wind data and CalMac Ferries near-miss data, both datasets consisted of extremely short incident descriptions, often only one or two sentences, which restricts the semantic richness captured by BERTopic. This leads to: increased outlier rates (e.g. 1657 outliers in ferry dataset and 2508 in offshore wind), limited ability to detect complex causal chains, and an inability to consistently extract deeper human or organisational factors.

The incident narratives were also written by different individuals with no standardised linguistic structure similar to the full maritime accident reports used in the SHIELD-NLP case study. As a result, some reports describe events, others only injuries or consequences, preventative actions were incidentally recorded, and terminology varied significantly across vessels, sites, and companies. This inconsistency affects cluster stability and interpretability. Beyond its impact on model performance, this variability also highlights a broader challenge for organisational safety management. Inconsistent reporting practices reduce the ability to compare incidents across operators, identify long-term trends, and systematically learn from operational experience. The findings therefore suggest that

improvements in reporting quality and consistency are likely to provide benefits that extend beyond AI performance, supporting more effective organisational learning and cross-industry safety benchmarking.

Neither dataset contained robust, expert-verified human and causal factor classification. Validation was therefore limited to mapping to company categories, mapping to structured metadata and expert judgement panels. While these helped verify topic quality, true supervised evaluation was not possible. The absence of consistent expert-labelled data also reflects current industry practice, where detailed human factors coding is rarely performed during routine investigations because of the expertise and time required. Consequently, many organisations are unable to systematically analyse human and organisational factors at scale. This limitation reinforces the importance of developing practical approaches that support, rather than replace, expert investigators while reducing the effort required for consistent safety analysis.

Although BERTopic outperformed traditional models, it has inherent constraints such as:

- Sensitivity to short texts
- High proportion of outliers
- Inability to model causal relationships on its own
- Difficulty capturing subtle human factors (e.g. fatigue, workload, decision-making)

Also, association rule mining helps identify co-occurrence patterns but does not imply causality, can produce sparse or weak rules when datasets contain many categories, and struggles when prevention labels are generic.

The studies were also based on only a single ferry operator and G+ dataset. The ferry operator only works within one small country, meaning data is skewed to issues recurring within that specific area and G+ data used, although international data does not include all offshore wind operators and so

does not represent the full industry. Generally, the data therefore does not represent the global maritime and offshore wind operations.

Finally, the effectiveness of AI-assisted safety analysis remains fundamentally dependent on the availability of high-quality operational data. While advances in machine learning can improve analytical capability, they cannot compensate for missing contextual information or inconsistent reporting practices. Consequently, improvements in reporting standards and investigator guidance should be viewed as complementary to advances in AI, rather than independent developments.

8.6 Recommendations for Future Research

While this thesis demonstrates the value of advanced NLP techniques for analysing maritime and offshore incident narratives, several avenues for future research have emerged from the findings the case studies. These opportunities fall into methodological, domain-specific, and operational categories, and collectively provide a roadmap for further development of AI-enabled safety analysis.

One of the contributions of this thesis is the development of SeaSafeBERT, a domain-specific transformed model trained on maritime accident narratives. Future work should continue expanding the underlying pre-training corpus to incorporate a wider range of global datasets, including additional flag-state and international investigation sources, with particular attention to accident types that are under-represented in the present dataset,. This will allow for improved linguistic coverage across vessel types, accident types, and regulatory environments. Additional evaluation should be conducted to benchmark SeaSafeBERT against general-purpose language models across multiple tasks (classification, clustering, question-answering), with a particular focus on its ability to capture subtle human and organisational factors.

The thesis also adapts the SHIELD taxonomy for machine learning through the creation of SHIELD-NLP. This provides a structured foundation for automating human and causal factor classification. Future development of SHIELD-NLP should focus specifically on the low-frequency factors that produce unstable or zero F1-scores in the present evaluation. Additional accident reports containing these

factors should be identified and independently coded by multiple human-factors experts to increase class representation and quantify inter-rater agreement. Alternative strategies for handling rare factors, including few-shot, class-weighted and data-augmentation approaches, could then be compared against the current classifier. Evaluation should determine whether performance improvements arise from increased training representation or changes in modelling strategy. This would establish the minimum data requirements needed for reliable automated classification of individual SHIELD-NLP factors.

A further validation study should examine the extent to which SHIELD-NLP reproduces individual expert judgement rather than stable expert consensus. A representative sample of previously unseen accident reports could be independently coded by multiple trained human-factors analysts and subsequently analysed by SHIELD-NLP. Agreement should be compared between individual experts, between the expert consensus and the model, and between the model and the original SHIELD labels. Such an experiment would distinguish model error from genuine disagreement in causal interpretation and provide a more defensible estimate of the level of automated performance that can reasonably be expected for subjective human-factors classifications.

All case studies highlight the difficulty of obtaining expert-labelled data for supervised learning. Future research should explore semi-supervised and weakly supervised approaches, including label propagation, pseudo-labelling, self-training, and reinforcement learning from human feedback (RLHF). These methods could significantly reduce the dependency on large volumes of expert-coded data and enable scalable deployment across organisations with limited human factors expertise.

Future research should empirically evaluate whether greater structure at the point of incident reporting improves both organisational learning and AI-assisted analysis. A semi-structured reporting template could be developed containing minimum fields for the event description, operational context, actual and potential consequences, contributory factors, failed or degraded barriers, actions that prevented escalation, and subsequent corrective actions. The template could be piloted with a

maritime or offshore operator and compared with the organisation's existing free-text reporting format. Reports collected before and after implementation should be compared for completeness, narrative length, missing information and consistency, while downstream topic-modelling quality, outlier rates and human-factor extraction performance should also be evaluated. This would provide empirical evidence of whether reporting standardisation improves analytical value without creating an unacceptable additional reporting burden. The current research also focused on text narratives, which are highly informative but limited in isolation. Future work could incorporate structured data such as vessel characteristics, equipment failure codes, weather conditions, operational phase, and sensor or telemetry data. Combining these data types would enable more comprehensive causal inference, support prediction of risk escalation, and allow for richer contextualisation of human factors within the broader operational environment.

An important extension would be to broaden the analytical framework beyond the identification of human and organisational factors to incorporate other categories of safety-critical contributors, particularly technological, equipment, environmental, and operational factors. While the present research primarily focuses on identifying and analysing human factors, maritime accidents typically emerge from interactions between multiple elements of the wider socio-technical system. Future research could therefore develop methods for automatically identifying non-human contributory factors, such as equipment failures, design limitations, automation or software-related issues, adverse environmental conditions, and infrastructure deficiencies, and linking these systematically with SHIELD-NLP human-factor classifications. Examining these relationships could reveal recurring combinations or pathways through which human, organisational, and technological factors interact in accident development, providing a more integrated representation of systemic maritime risk.

Furthermore, the cloud-based interface proposed in this thesis could form the basis of a real-time safety intelligence platform. Future work may develop automated ingestion pipelines, live clustering of newly reported incidents, real-time SHIELD-factor classification, and automated alerts when new

risk patterns or anomalous topic clusters emerge. Such a system would significantly enhance organisational capability to identify weak signals and respond proactively. Following technical validation, the framework should be evaluated within an operational Safety Management System (SMS). A pilot implementation could automatically process newly submitted incident and near-miss reports and provide safety managers with periodically updated human-factor distributions, emerging topic clusters, temporal trends and high-interest association rules. Predefined escalation criteria could be established for patterns requiring expert review, rather than allowing automated outputs to trigger safety decisions directly. The pilot should evaluate whether identified patterns subsequently influence risk-register updates, safety assurance activities, investigation priorities or targeted interventions. This would establish whether the analytical outputs produce measurable organisational learning rather than simply additional information. In addition, a controlled human-in-the-loop evaluation should also compare safety analysis performed with and without AI assistance. Investigators or safety practitioners could be randomly assigned accident and near-miss reports for analysis using either conventional manual review or the proposed decision-support interface. Outcomes should include analysis time, agreement with an expert reference assessment, number of causal factors identified, consistency between analysts, perceived workload and user trust in the outputs. A crossover design, in which participants use both approaches on different reports, would help control for differences in analyst experience. This would provide direct evidence of whether the framework improves the efficiency or consistency of safety analysis rather than demonstrating technical model performance alone. The transferability demonstrated in this thesis should also be tested across multiple independent operators rather than additional datasets from a single reporting environment. The short-narrative pipeline should be applied without major methodological modification to near-miss datasets from multiple ferry, shipping and offshore operators. Topic stability, outlier rates, expert coherence ratings and the recurrence of major incident and prevention themes should then be compared across organisations. This would distinguish genuinely transferable safety patterns from themes that primarily reflect the reporting culture, operating environment or terminology of individual organisations.

Finally, future research should also address several broader socio-technical and methodological challenges revealed through this thesis. First, NLP-enabled safety analysis must incorporate principles of Safety-II and resilience engineering by expanding beyond accident narratives and near-misses to normal operational conditions. Additionally, further work is required to examine issues of bias, fairness, and cross-lingual generalisation in maritime narratives, given the multinational composition of crews and the cultural differences in reporting practices.

9. Conclusions

9.1 Chapter Overview

This chapter presents the overall conclusions of this thesis. It summarises the key research outcomes, the main scientific and practical consequences, and the implications of applying AI and NLP techniques to the analysis of unstructured maritime safety texts across the full spectrum of safety reporting.

9.2 Conclusions

This thesis demonstrates that modern natural language processing (NLP) and machine learning techniques can be successfully applied to transform large volumes of unstructured maritime and offshore safety texts into structured, actionable safety intelligence. Through the development and validation of a unified AI-based evaluation framework, the research shows how accident investigations, incident reports, near misses and prevention-focused narratives can be analysed automatically to identify human and organisational factors, operational themes, causal relationships, and emerging safety trends. The framework was evaluated across three complementary investigations using the SHIELD taxonomy, unsupervised analysis of ferry near-miss reports, and application to offshore wind incident data, demonstrating its effectiveness across multiple reporting contexts and safety-critical domains.

This research has demonstrated that it is technically feasible to extract, classify, and structure human and organisational safety information from free-text safety reports with a high level of reliability for key causal factors. Furthermore, the work has shown that a single analytical framework can be applied consistently across different levels of event severity and reporting detail, enabling continuous safety learning from severe accidents through to everyday safety-critical events.

The main concluding contributions of this thesis are summarised below:

- The critical review identified a gap between established human-factors approaches, which provide structured but predominantly expert-led analysis, and AI-based methods capable of

analysing safety narratives at scale, together with a need for analytical approaches that account for differences in the level of detail, structure and causal information available across safety-reporting data.

- This gap was addressed through the development of a unified, data-dependent framework comprising two complementary pathways: supervised taxonomy-aligned human-factors classification for detailed accident narratives and unsupervised semantic analysis for shorter narratives where detailed causal classification is not supported by the available information.
- The high-information pathway was demonstrated through SHIELD-NLP, which transformed the SHIELD taxonomy into 53 machine-learning-ready categories and combined hierarchical multi-label classification, lexical and domain-adapted contextual representations, supervised learning and human-in-the-loop review to classify human and organisational factors within detailed maritime accident reports.
- The low-information pathway was demonstrated using 4,360 ferry near-miss reports, where BERTopic identified recurring incident and prevention themes and subsequent temporal and Association Rule Mining analyses examined changes in reported themes and relationships between near-miss and prevention information without imposing unsupported causal classifications.
- Application of the low-information pathway to 5,403 offshore wind occupational incident reports demonstrated its transferability to a different safety-critical reporting context, with topic discovery extended through temporal and association analysis of structured incident information and supervised multi-label modelling using existing G+ classifications.

Collectively, these findings demonstrate that the overall research aim and associated objectives have been achieved through the development and evaluation of an integrated framework that matches AI-based analytical methods to the characteristics and evidential content of heterogeneous maritime and offshore safety data.

Beyond the methodological advances presented, this research provides a practical foundation for the next generation of AI-enabled safety management systems. The framework has the potential to support investigators through more consistent causal analysis, assist safety managers in identifying emerging hazards and recurring organisational weaknesses, and provide regulators and industry with richer evidence to guide policy development and risk reduction. As reporting volumes continue to increase across maritime and offshore industries, the ability to automatically transform unstructured narratives into actionable safety intelligence will become increasingly important.

Ultimately, this thesis demonstrates that artificial intelligence is not simply a tool for automating safety analysis, but a means of strengthening organisational learning. By integrating explainable AI with established human factors methodologies and maintaining human oversight throughout the analytical process, the research provides a scalable, transparent, and transferable framework capable of supporting safer maritime and offshore operations.

References

- (IOGP), I. A. O. O. A. G. P. 2025. Safety performance indicators - 2024 data.
- ABE, T., SATO, H. & NAKAMURA, K. 2022. Extracting Safety-II Factors From an Incident Reporting System by Text Analysis. *CUREUS JOURNAL OF MEDICAL SCIENCE*, 14.
- ADUMENE, S., AFENYO, M., SALEHI, V. & WILLIAM, P. 2022. An adaptive model for human factors assessment in maritime operations. *International journal of industrial ergonomics*, 89, 103293.
- AHADH, A., BINISH, G. V. & SRINIVASAN, R. 2021. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process Safety and Environmental Protection*, 155, 455-465.
- AKHTAR, M. J. & UTNE, I. B. 2014. Human fatigue's effect on the risk of maritime groundings – A Bayesian Network modeling approach. *Safety Science*, 62, 427-440.
- AKIYAMA, M. & FUJITA, K. How to improve patient safety by text mining with medical incident reports: Innovative technologies using e-health and health technology assessment. 2013 2013. 2543-2550.
- AKYILDIZ, H. & MENTES, A. 2023. Risk Assessment of Launching Airbags Using Functional Resonance Accident Model. *Transactions on Maritime Science*, 12.
- AKYUZ, E. 2015. A hybrid accident analysis method to assess potential navigational contingencies: The case of ship grounding. *Safety Science*, 79, 268-276.
- AKYUZ, E. 2017. A marine accident analysing model to evaluate potential operational causes in cargo ships. *Safety Science*, 92, 17-25.
- AKYUZ, E. & CELIK, M. 2014. Utilisation of cognitive map in modelling human error in marine accident analysis and prevention. *Safety Science*, 70, 19-28.
- AKYUZ, E., CELIK, M. & CEBI, S. 2016. A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique. *Safety Science*, 87, 63-75.
- ALAN, B. & SOGUT, O. S. 2020. HUMAN FACTORS IMPLICATIONS ON SHIP BRIDGE DESIGN. *SIGMA JOURNAL OF ENGINEERING AND NATURAL SCIENCES-SIGMA MUHENDISLIK VE FEN BILIMLERI DERGISI*, 38, 1567-1579.
- ALEXANDER, D. E. 2012. The 'Titanic Syndrome': Risk and Crisis Management on the Costa Concordia. *Journal of Homeland Security and Emergency Management*, 9.
- ANDRADE, S. R. & WALSH, H. S. 2023. Machine learning framework for Hazard Extraction and Analysis of Trends (HEAT) in wildfire response. *Safety Science*, 167, 106252.
- ANTÃO, P. & GUEDES SOARES, C. 2008. Causal factors in accidents of high-speed craft and conventional ocean-going vessels. *Reliability Engineering & System Safety*, 93, 1292-1304.
- ASHRAF, L. 2014. Chapter 12 - Titanic, the Unsinkable Ship That Sank. In: ASHRAF, L. (ed.) *Learning from Failures*. Oxford: Butterworth-Heinemann.
- AYDIN, M., UĞURLU, Ö. & BORAN, M. 2022. Assessment of human error contribution to maritime pilot transfer operation under HFACS-PV and SLIM approach. *Ocean Engineering*, 266, 112830-112830.
- BAIRAMI-KHANKANDI, S., BOLBOT, V., BAHOOTOROODY, A. & GOERLANDT, F. 2025. A systems-theoretic approach using association rule mining and predictive Bayesian trend analysis to identify patterns in maritime accident causes. *Reliability Engineering & System Safety*, 258, 110911.
- BAKER, H., HALLOWELL, M. R. & TIXIER, A. J. P. 2020a. AI-based prediction of independent construction safety outcomes from universal attributes. *Automation in Construction*, 118, 103146-103146.
- BAKER, H., HALLOWELL, M. R. & TIXIER, A. J. P. 2020b. Automatically learning construction injury precursors from text. *Automation in Construction*, 118, 103145-103145.

- BAO, J., CHEN, Y., YIN, J., CHEN, X. & ZHU, D. 2023. Exploring topics and trends in Chinese ATC incident reports using a domain-knowledge driven topic model. *Journal of Air Transport Management*, 108, 102374.
- BATALDEN, B. M. & SYDNES, A. K. 2014. Maritime safety and the ISM code: A study of investigated casualties and incidents. *WMU Journal of Maritime Affairs*, 13, 3-25.
- BATBAATAR, E., LI, M. & RYU, K. H. 2019. Semantic-Emotion Neural Network for Emotion Recognition From Text. *IEEE ACCESS*, 7, 111866-111878.
- BAYAZIT, O. & KAPTAN, M. 2024. Dynamic risk analysis of allision in port areas using DBN based on HFACS-PV. *Ocean Engineering*, 298, 117183.
- BICEN, S. & CELIK, M. 2022. A hybrid approach to near-miss report investigation towards next-generation safety solutions on-board ships. *Ocean Engineering*, 112768-112768.
- BIN, H. & SHIQI, F. 2025. Analysis of marine oil spill accidents using a human-centered approach. *Transportation Research Part D: Transport and Environment*, 147, 104963.
- BINDER, S., PAUL, J., BEN, P. & BOB, W. 2010. 20 Years on lessons learned from Piper Alpha. The evolution of concurrent and inherently safe design. *Journal of Loss Prevention in the Process Industries*, 23, 936-953.
- BIRD, F. E. 1969. Accident Investigation and Analysis in Industry. *Van Nostrand Reinhold*.
- BLACK, H., LOUIS, YASIN BURAK, K., YASER, F., DWITYA, W., RAFET, K. & OSMAN, T. 2026a. Uncovering patterns in ferry near misses: A topic modelling approach with GPT-assisted labelling. *Ocean Engineering*, 363, 126862.
- BLACK, H., RAFET, K. & OSMAN, T. 2026b. A domain-adapted NLP pipeline for human factors classification of maritime accidents. *Ocean Engineering*, 357, 125625.
- BORJALI, A., MAGNÉLI, M., SHIN, D., MALCHAU, H., MURATOGLU, O. K. & VARADARAJAN, K. M. 2021. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. *Computers in Biology and Medicine*, 129, 104140-104140.
- BOWO, L. P., PRILANA, R. E. & FURUSHO, M. A hybrid methodology for maritime accident analysis: The case of ship collision. *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE*, 2019. American Society of Mechanical Engineers (ASME).
- BUTLER, S., RINGWOOD, J. V. & O'CONNOR, F. 2013. Exploiting scada system data for wind turbine performance monitoring. *2013 Conference on Control and Fault-Tolerant Systems (SysTol)*.
- BYON, H. D., HARRIS, C., CRANDALL, M., SONG, J. & TOPAZ, M. 2023. Identifying Type II workplace violence from clinical notes using natural language processing. *WORKPLACE HEALTH & SAFETY*, 71, 484-490.
- CAO, K., CHEN, S., YANG, C., LI, Z., LUO, L. & REN, Z. 2024a. Revealing the coupled evolution process of construction risks in mega hydropower engineering through textual semantics. *Advanced Engineering Informatics*, 62, 102713.
- CAO, K., CHEN, S., ZHANG, X., CHEN, Y., LI, Z. & WANG, D. 2024b. Identification of causative factors for fatal accidents in the electric power industry using text categorization and catastrophe association analysis techniques. *Alexandria Engineering Journal*, 102, 290-308.
- CAO, Y., WANG, X., YANG, Z., WANG, J., WANG, H. & LIU, Z. 2023. Research in marine accidents: A bibliometric analysis, systematic review and future directions. *Ocean Engineering*, 284, 115048.
- CELIK, M. & CEBI, S. 2009. Analytical HFACS for investigating human errors in shipping accidents. *Accident Analysis & Prevention*, 41, 66-75.
- CERQUEIRA, I., SILVA, A. J. N., ÁVILA FILHO, S. & ÁVILA, J. Application of Tool to Evaluate the Causes of Accident According to the Social-technical Model of Rasmussen: Case of Wreck in Mar Grande-Brazil. *In: TRZCIELINSKI, S., MRUGALSKA, B., KARWOWSKI, W., DI NICOLANTONIO, M. & ROSSI, E., eds. Advances in Intelligent Systems and Computing*, 2019. Springer Verlag, 439-448.

- CEYLAN, B. O., AKYUZ, E. & ARSLAN, O. 2021. Systems-Theoretic Accident Model and Processes (STAMP) approach to analyse socio-technical systems of ship allision in narrow waters. *Ocean Engineering*, 239, 109804.
- CEYLAN, B. O., AKYUZ, E. & ARSLANOĞLU, Y. 2022. Modified quantitative systems theoretic accident model and processes (STAMP) analysis: A catastrophic ship engine failure case. *Ocean Engineering*, 253, 111187.
- CHANG, J. & DE JESUS MARTINEZ, J. 2024. Application of natural language processing for spill reduction in an exploration and production company. *Process Safety and Environmental Protection*, 185, 21-24.
- CHAUVIN, C., LARDJANE, S., MOREL, G., CLOSTERMANN, J.-P. & LANGARD, B. 2013. Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. *Accident Analysis & Prevention*, 59, 26-37.
- CHEN, D., PEI, Y. & XIA, Q. 2020a. Research on human factors cause chain of ship accidents based on multidimensional association rules. *Ocean Engineering*, 218, 107717.
- CHEN, H., COHEN, E., WILSON, D. & ALFRED, M. 2024. A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study. *JMIR Human Factors*, 11.
- CHEN, J., ZHUANG, C., SHI, J., JIANG, H., XU, J. & LIU, J. 2025a. Risk factors extraction and analysis of Chinese ship collision accidents based on knowledge graph. *Ocean Engineering*, 322, 120536.
- CHEN, P., MOU, J. & VAN GELDER, P. H. A. J. M. 2019. Integration of individual encounter information into causation probability modelling of ship collision accidents. *Safety Science*, 120, 636-651.
- CHEN, P., ZHANG, A., ZHANG, S., DONG, T., ZENG, X., CHEN, S., SHI, P., WONG, Y. D. & ZHOU, Q. 2025b. Maritime Near-Miss prediction framework and model interpretation analysis method based on Transformer neural network model with multi-task classification variables. *Reliability Engineering & System Safety*, 257, 110845.
- CHEN, S.-T., WALL, A., DAVIES, P., YANG, Z., WANG, J. & CHOU, Y.-H. 2013. A Human and Organisational Factors (HOFs) analysis method for marine casualties using HFACS-Maritime Accidents (HFACS-MA). *Safety Science*, 60, 105-114.
- CHEN, S. T. 2020. An approach of identifying the common human and organisational factors (HOFs) among a group of marine accidents using GRA and HFACS-MA. *Journal of Transportation Safety and Security*, 12, 1252-1294.
- CHEN, W., WANG, X., ZHANG, F., LIU, H. & LIN, Y. 2020b. Review of the application of hydraulic technology in wind turbine. *Wind Energy*, 23, 1495-1522.
- CHEN, Z., HUANG, K., WU, L., ZHONG, Z. & JIAO, Z. 2022. Relational Graph Convolutional Network for Text-Mining-Based Accident Causal Classification. *APPLIED SCIENCES-BASEL*, 12.
- CHENG, M.-Y., KUSOEMO, D. & GOSNO, R. A. 2020. Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, 118, 103265-103265.
- COSMA, G., SINGH, M. K., WATERSON, P., JUN, G. T. & BACK, J. Intelligent Multi-document Summarisation for Extracting Insights on Racial Inequalities from Maternity Incident Investigation Reports. In: XIE, X., STYLES, I., POWATHIL, G. & CECCARELLI, M., eds. ARTIFICIAL INTELLIGENCE IN HEALTHCARE, PT II, AIIH 2024, 2024. 316-329.
- COSTANTINO, F. & SABETTA, N. 2025. Large Language Models for topic analysis: insights from robotic accident narratives. *Procedia Computer Science*, 253, 1462-1472.
- DANIAL, S. N., SMITH, D. & VEITCH, B. 2021. A Method to Detect Anomalies in Complex Socio-Technical Operations Using Structural Similarity. *Journal of Marine Science and Engineering* [Online], 9.
- DAVID, R., DAVID, M. & JAMES, C. 2024. Offshore wind H&S: A review and analysis. *Renewable and Sustainable Energy Reviews*, 189, 113928.
- DAVIDSON, J. E., YE, G., PARRA, M. C., CHOFLET, A., LEE, K., BARNES, A., HARKAVY-FRIEDMAN, J. & ZISOOK, S. 2021. Job-Related Problems Prior to Nurse Suicide, 2003-2017: A Mixed Methods

- Analysis Using Natural Language Processing and Thematic Analysis. *Journal of Nursing Regulation*, 12, 28-39.
- DEMIRCI, S. M. E., CANIMOĞLU, R. & ELÇİÇEK, H. 2023. Analysis of causal relations of marine accidents during ship navigation under pilotage: A DEMATEL approach. *Proceedings of the Institution of Mechanical Engineers Part M: Journal of Engineering for the Maritime Environment*, 237, 308-321.
- DETHLEFF, D., WEINRICH, N., KOWALD, B., HORY, D., FRANZ, R., NIELSEN, M. V., SEIDE, K., JÜRGENS, C. & STUHR, M. 2016. Air Medical Evacuations From the German North Sea Wind Farm Bard Offshore 1: Traumatic Injuries, Acute Diseases, and Rescue Process Times (2011-2013). *Air Med J*, 35, 216-226.
- DINGDING, Y., XIAOLAN, H., BAOJING, X., ZHEN, D., JUAN, Z. & JING, W. 2025. Multidimensional causation chain analysis of marine oil spill accidents: based on grounded theory and bow-tie model. *Ocean Engineering*, 341, 122824.
- EBRAHIMI, H., SATTARI, F., LEFSRUD, L. & MACCIOTTA, R. 2023. A machine learning and data analytics approach for predicting evacuation and identifying contributing factors during hazardous materials incidents on railways. *Safety Science*, 164, 106180.
- ELIZABETH, R. M. C., SATTARI, F., LEFSRUD, L. & GUE, B. 2025. Visualizing what's missing: Using deep learning and Bow-Tie diagrams to identify and visualize missing leading indicators in industrial construction. *Journal of Safety Research*, 93, 1-11.
- EMSA 2024. Annual overview of marine casualties and incidents 2024.
- EMSA 2025. Annual overview of marine casualties and incidents 2025.
- EVANS, H. P., ANASTASIOU, A., EDWARDS, A., HIBBERT, P., MAKEHAM, M., LUZ, S., SHEIKH, A., DONALDSON, L. & CARSON-STEVENSON, A. 2020. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. *HEALTH INFORMATICS JOURNAL*, 26, 3123-3139.
- FAHREZA, M. I. & HIRATA, E. 2024. Maritime piracy and armed robbery analysis in the Straits of Malacca and Singapore through the utilization of natural language processing. *Maritime Policy and Management*.
- FENG, X., DAI, Y., JI, X., ZHOU, L. & DANG, Y. 2021. Application of natural language processing in HAZOP reports. *Process Safety and Environmental Protection*, 155, 41-48.
- FILHO, A. P. G., JUN, G. T. & WATERSON, P. 2019. Four studies, two methods, one accident – An examination of the reliability and validity of Accimap and STAMP for accident analysis. *Safety Science*, 113, 310-317.
- FORSTER, A. J., ANDRADE, J. & VAN WALRAVEN, C. 2005. Validation of a Discharge Summary Term Search Method to Detect Adverse Events. *Journal of the American Medical Informatics Association*, 12, 200-206.
- FUJITA, K., AKIYAMA, M., PARK, K., YAMAGUCHI, E. & FURUKAWA, H. 2012. Linguistic analysis of large-scale medical incident reports for patient safety. *Studies in Health Technology and Informatics*, 180, 250-254.
- G+ 2019. G+ Global Offshore Wind Health & Safety Organisation 2018 incident data report. ISBN: N 978 1 78725 095 6.
- G+ 2020. G+ Global Offshore Wind Health & Safety Organisation 2019 incident data report. ISBN: 978 1 78725 203 5.
- G+ 2021. G+ Global Offshore Wind Health & Safety Organisation 2020 incident data report. ISBN: N 978 1 78725 260 8.
- G+ 2022. G+ Global Offshore Wind Health & Safety Organisation 2021 incident data report. ISBN: 978 1 78725 327 8.
- G+ 2023a. G+ Global Health and
- G+ 2023b. *G+ Global Offshore Wind Health and Safety Organisation 2022 incident data report* ISBN: 978 1 78725 377 3, Energy Institute.

- G+ 2024. G+ Global Offshore Wind Health & Safety Organisation 2023 incident data report. . Energy Institute.
- GALIERIKOVÁ, A. 2019. The human factor and maritime safety. *Transportation Research Procedia*, 40, 1319-1326.
- GAN, L., YE, B., HUANG, Z., XU, Y., CHEN, Q. & SHU, Y. 2023. Knowledge graph construction based on ship collision accident reports to improve maritime traffic safety. *Ocean & coastal management*, 240, 106660.
- GANGADHARI, R. K., RABIEE, M., KHANZODE, V., MURTHY, S. & KUMAR TAREI, P. 2024. From unstructured accident reports to a hybrid decision support system for occupational risk management: The consensus converging approach. *Journal of Safety Research*, 89, 91-104.
- GOLDEN, A. S. & WEISBROD, R. E. 2016. Trends, Causal Analysis, and Recommendations from 14 Years of Ferry Accidents. *Journal of Public Transportation*, 19, 17-27.
- GOMMOSANI, M., TURAN, O. & KURT, R. 2021. Analysis of Maritime Accidents due to Poor Situational Awareness. *International Conference on the Stability and Safety of Ships and Ocean Vehicles*. Glasgow, Scotland.
- GRAZIANO, A., TEIXEIRA, A. P. & GUEDES SOARES, C. 2016. Classification of human errors in grounding and collision accidents using the TRACER taxonomy. *SAFETY SCIENCE*, 86, 245-257.
- GROENEVELD, M., LEURS, W. L. M., BOUWMAN, A. R. A., SCHENK, J., LAMMERS, L., DIERICK, A., KORSTEN, E. & VAN DER LINDEN, C. C. M. J. 2025. Text-based fall prediction in hospital: Development and internal validation of a model to predict in-hospital falls in older patients using free text from daily nursing records. *Applied Nursing Research*, 82, 151923.
- GROUIN, C., GRABAR, N., HAMON, T., ROSSET, S., TANNIER, X. & ZWEIGENBAUM, P. 2013. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association*, 20, 820-827.
- GUAN, Y., SUN, Y., WANG, Z. & ZHENG, Q. 2017. Natural Language Process: A New Kind of Nuclear Quality Assurance Management Tool. *Energy Procedia*, 127, 201-219.
- GUARD), U. S. C. G. U. C. 1990. *Oil Pollution Act of 1990 (OPA)* [Online]. Available: [https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About NPFC/OPA/](https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About_NPFC/OPA/) [Accessed].
- GUCMA, L., ANDROJNA, A., ŁAZUGA, K., VIDMAR, P. & PERKOVIČ, M. 2021. Reconstructing maritime incidents and accidents using causal models for safety improvement: Based on a case study. *Journal of Marine Science and Engineering*, 9.
- GURDAL ERTEK & KAILAS, L. 2021. Analyzing a decade of wind turbine accident news with topic modeling. *Sustainability*, 13.
- GURDAL ERTEK, XU CHI, ALLAN N. ZHANG & ASIAN, S. 2017. Text mining analysis of wind turbine accidents: An ontology-based framework. *IEEE*.
- HÄNNINEN, H. 2003. Normalization of risks: a stream of bow visor incidents and the estonia ferry accident. *Technological Change and Organizational Action*, 117-134.
- HASSEL, M., ASBJØRNSLETT, B. E. & HOLE, L. P. 2011. Underreporting of maritime accidents to vessel accident databases. *ACCIDENT ANAL PREV*, 43, 2053-2063.
- HAUGEN, S. & KRISTIANSEN, S. 2022. *Maritime Transportation: Safety Management and Risk Analysis (2nd ed.)*. Routledge.
- HAWKINS, F. 1993. *Human Factors in Flight*, London, Routledge.
- HE, C. 2020. *A Practical Guide to Human Factors Analysis and Classification System (HFACS) 7.0*.
- HE, L., XIAOXUE, M., WEILIANG, Q. & YANG, L. 2022. A methodology to assess the causation relationship of seafarers' unsafe acts for ship grounding accidents based on Bayesian SEM. *Ocean & Coastal Management*, 225, 106189-106189.
- HEBBAR, A., A., M., JR, M. Q. & HANZU-PAZARA, R. 2023. Safety of domestic ferries: A scoping study of seven high-risk countries.
- HEINRICH, H. W. 1941. *Industrial Accident Prevention*, McGraw-hill Book Company Inc., New York and London.

- HETHERINGTON, C., FLIN, R. & MEARNS, K. 2006. Safety in shipping: The human element. *Journal of Safety Research*, 37, 401-411.
- HINRICHSSEN, V. L., KRUSKAL, B., O'BRIEN, M. A., LIEU, T. A. & PLATT, R. 2007. Using Electronic Medical Records to Enhance Detection and Reporting of Vaccine Adverse Events. *Journal of the American Medical Informatics Association*, 14, 731-735.
- HOLLNAGEL, E. 1998. Cognitive reliability and error analysis method [internet resource] : CREAM. CREAM.
- HOLLNAGEL, E. 2014. *Safety-I and Safety-II: The Past and Future of Safety Management (1st ed.)*, CRC Press. .
- HOLLNAGEL, E. 2017. Resilience Engineering and the Future of Safety Management. *Handbook of Safety Principles*.
- HONG, W.-T., CLIFTON, G. & NELSON, J. D. 2024. A data-driven conceptual framework for understanding the nature of hazards in railway accidents. *Transport Policy*, 152, 102-117.
- HOU, Z., WANG, H., YUE, Y., XIONG, M. & CHE, C. 2025. A novel method for cause portrait of aviation unsafe events based on hierarchical multi-task convolutional neural network. *Expert Systems with Applications*, 270, 126466.
- HUANG, X., WEN, Y., ZHANG, F., LI, H., SUI, Z. & CHENG, X. 2025a. Accident analysis of waterway dangerous goods transport: Building an evolution network with text knowledge extraction. *Ocean Engineering*, 318, 120176.
- HUANG, Y., ZHANG, Z. & HU, H. 2025b. Risk propagation mechanisms in railway systems under extreme weather: A knowledge graph-based unsupervised causation chain approach. *Reliability Engineering & System Safety*, 260, 110976.
- HUANG, Y., ZHANG, Z., TAO, Y. & HU, H. 2022. Quantitative risk assessment of railway intrusions with text mining and fuzzy Rule-Based Bow-Tie model. *Advanced Engineering Informatics*, 54, 101726-101726.
- HUGHES, P., ROBINSON, R., FIGUERES-ESTEBAN, M. & VAN GULIJK, C. 2019. Extracting safety information from multi-lingual accident reports using an ontology-based approach. *Safety Science*, 118, 288-297.
- HUGHES, P., SHIPP, D., FIGUERES-ESTEBAN, M. & VAN GULIJK, C. 2018. From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram. *Safety Science*, 110, 11-19.
- HULME, A., STANTON, N. A., WALKER, G. H., WATERSON, P. & SALMON, P. M. 2019. What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. *Safety Science*, 117, 164-183.
- IAEA, I. A. E. A. 2013. Human and Organizational Factors in Nuclear Safety in the Light of the Accident at the Fukushima Daiichi Nuclear Power Plant.
- IDAKWO, P. O., ADEKANMBI, O., SORONNADI, A. & DAVID, A. 2025. Geo-parsing and analysis of road traffic crash incidents for data-driven emergency response planning. *Heliyon*, 11, e41067.
- IMO. 1974. *International Convention for the Safety of Life at Sea (SOLAS), 1974* [Online]. Available: [https://www.imo.org/en/About/Conventions/Pages/International-Convention-for-the-Safety-of-Life-at-Sea-\(SOLAS\),-1974.aspx](https://www.imo.org/en/About/Conventions/Pages/International-Convention-for-the-Safety-of-Life-at-Sea-(SOLAS),-1974.aspx) [Accessed].
- IMO. 2010. *Casualties* [Online]. Available: <https://www.imo.org/en/OurWork/MSAS/Pages/Casualties.aspx> [Accessed].
- IRSHAD, L. & WALSH, H. 2024. IDENTIFYING HUMAN ERRORS AND ERROR MECHANISMS FROM ACCIDENT REPORTS USING LARGE LANGUAGE MODELS.
- IVANA, J., MAJA, P. & NIKOLA, V. 2025. Assessment of human contribution to cargo ship accidents using Fault Tree Analysis and Bayesian Network Analysis. *Ocean Engineering*, 323, 120628.
- IVERGÅRD, T., LUNDBERG, M., NORBERG, L. & SVEDUNG, I. 1978. Arbetsmiljö inom sjöfarten—en kartläggning (Work environment in the shipping industry—a survey). *Stockholm: Sjöfartens arbetarskyddsnämnd, SAN*.

- JIA, Q., FU, G., XIE, X., XUE, Y. & HU, S. 2024. Enhancing accident cause analysis through text classification and accident causation theory: A case study of coal mine gas explosion accidents. *Process Safety and Environmental Protection*, 185, 989-1002.
- JIANG, B. & WANG, K. 2024. Railway accident causation prediction with improved transformer model based on lexical information and contextual relationships. *Knowledge-Based Systems*, 296, 111897.
- JIANG, W., ZHANG, M., LI, M. & XU, Y. 2025a. Study on the occurrence path and prediction of unsafe behaviours of hazardous chemical storage personnel. *Journal of Loss Prevention in the Process Industries*, 96, 105653.
- JIANG, Z., XIAO, Z., FENG, Y. & ZHOU, D. 2025b. Human factor analysis of ship pilotage accidents based on HFACS-BN model. *Dalian Haishi Daxue Xuebao/Journal of Dalian Maritime University*, 51, 63-70 and 91.
- JIAO, Z., ZHANG, Z., JUNG, S. & WANG, Q. 2023. Machine learning based quantitative consequence prediction models for toxic dispersion casualty. *Journal of Loss Prevention in the Process Industries*, 81, 104952.
- JIAXIN, W., HANWEN, F., ZHENG, C. & JING, L. 2025. Unleashing data power: Driving maritime risk analysis with Bayesian networks. *Reliability Engineering & System Safety*, 264, 111310.
- JIN, L., LI, P., WANG, Y. & YANG, Z. 2025a. Risk analysis of Arctic navigation using text mining (TM) and improved association rule mining (ARM) methods. *Regional Studies in Marine Science*, 81, 103990.
- JIN, Y., TONG, X., ZHENG, X., LI, Y. & DONG, B. 2025b. Risk assessment of coal and gas outbursts driven by the theory of three types of hazards coupled with 80 accident cases. *Safety Science*, 184, 106771.
- JINGPENG, L., JING, L. & HANWEN, F. 2025. Uncovering maritime accident causation: A two-phase framework based on multi-layer data-driven Bayesian network. *Ocean Engineering*, 340, 122444.
- JOSE, S., NGUYEN, K. T. P., MEDJAHHER, K., ZEMOURI, R., LÉVESQUE, M. & TAHAN, A. 2024. Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models. *Expert Systems with Applications*, 255, 124603.
- JULIET WANGUI, W., GOKHAN, C., CENK, S., SANGWON, P., YOUNGSOO, P., MUHAMMET, A. & DAEWON, K. 2025. Probabilistic modeling of domestic ferry accident causes in Kenya's Likoni ferry route using fuzzy Bayesian network. *Ocean Engineering*, 340, 122388.
- KAMIL, M. Z., KHAN, F., AMYOTTE, P. & AHMED, S. 2024. Multi-source heterogeneous data integration for incident likelihood analysis. *Computers & Chemical Engineering*, 185, 108677.
- KAMIL, M. Z., KHAN, F., HALIM, S. Z., AMYOTTE, P. & AHMED, S. 2023a. A methodical approach for knowledge-based fire and explosion accident likelihood analysis. *Process Safety and Environmental Protection*, 170, 339-355.
- KAMIL, M. Z., TALEB-BERROUANE, M., KHAN, F., AMYOTTE, P. & AHMED, S. 2023b. Textual data transformations using natural language processing for risk assessment. *Risk Analysis*, 43, 2033 – 2052.
- KANDEMIR, C. & CELIK, M. 2021. Determining the error producing conditions in marine engineering maintenance and operations through HFACS-MMO. *Reliability Engineering & System Safety*, 206, 107308.
- KAPTAN, M., SARIALIOĞLU, S., UĞURLU, Ö. & WANG, J. 2021a. The evolution of the HFACS method used in analysis of marine accidents: A review. *International Journal of Industrial Ergonomics*, 86, 103225.
- KAPTAN, M., UĞURLU, Ö. & WANG, J. 2021b. The effect of nonconformities encountered in the use of technology on the occurrence of collision, contact and grounding accidents. *Reliability Engineering & System Safety*, 215, 107886.
- KARAMANOLAKIS, G., HSU, D. & GRAVANO, L. 2024. Interactive Machine Teaching by Labeling Rules and Instances. *Transactions of the Association for Computational Linguistics*, 12, 1441-1459.

- KECECI, T. & ARSLAN, O. 2017. SHARE technique: A novel approach to root cause analysis of ship accidents. *Safety Science*, 96, 1-21.
- KHAN, N., NADEAU, S., PHAM, X.-T. & BOTON, C. 2024. Exploring associations between accident types and activities in construction using natural language processing. *Automation in Construction*, 164, 105457.
- KIM, H., CHOI, J., NAM, Y. & YOUN, J.-H. 2022. Characteristic Analysis of the Built Environment of Ferry Terminals: A Case Study of Mokpo, South Korea. *Sustainability*, 14, 2115.
- KIM, J., KIM, J., LEE, A., KIM, J. & DIESNER, J. 2024. LERCause: Deep learning approaches for causal sentence identification from nuclear safety reports. *PLoS ONE*, 19.
- KIM, T. E., NAZIR, S. & ØVERGÅRD, K. I. 2016. A STAMP-based causal analysis of the Korean Sewol ferry accident. *Safety Science*, 83, 93-101.
- KIZAKI, H., SATOH, H., EBARA, S., WATABE, S., SAWADA, Y., IMAI, S. & HORI, S. 2024. Construction of a Multi-Label Classifier for Extracting Multiple Incident Factors From Medication Incident Reports in Residential Care Facilities: Natural Language Processing Approach. *JMIR Medical Informatics*, 12.
- KOU, L., LI, Y., ZHANG, F., GONG, X., HU, Y., YUAN, Q. & KE, W. 2022. Review on monitoring, operation and maintenance of smart offshore wind farms. *Sensors*, 22, 2822.
- KRIPPENDORFF, K. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed., Thousand Oaks, CA, Sage.
- KRISTIANSEN, S. 2005. Maritime transportation : safety management and risk analysis.
- KU, C.-H. & LEROY, G. 2014. A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly*, 31, 534-544.
- KUHN, K. D. 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87, 105-122.
- KUMI, L., JEONG, J. & JEONG, J. 2024. Data-driven automatic classification model for construction accident cases using natural language processing with hyperparameter tuning. *Automation in Construction*, 164, 105458.
- KUMI, L., JEONG, J. & JEONG, J. 2025. Proactive approach to enhancing safety management using deep learning classifiers for construction safety documentation. *Engineering Applications of Artificial Intelligence*, 153, 110889.
- KURIAN, D., SATTARI, F., LEFSRUD, L. & MA, Y. 2020. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. *Safety Science*, 130, 104873-104873.
- KWAYU, K. M., KWIGIZILE, V., LEE, K. & OH, J.-S. 2021. Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. *Accident Analysis & Prevention*, 150, 105899-105899.
- LAIHAO, M., XIAOXUE, M., RUIWEN, Z. & QIAOLING, D. 2025. Investigation of distinct and joint contributions of human factors and operational conditions to different types of maritime accidents. *Ocean & Coastal Management*, 267, 107750.
- LAN, H., MA, X., QIAO, W. & MA, L. 2022. On the causation of seafarers' unsafe acts using grounded theory and association rule. *Reliability Engineering & System Safety*, 223, 108498.
- LÁZARO, F. L., MADEIRA, T., MELICIO, R., VALÉRIO, D. & SANTOS, L. F. F. M. 2025. Identifying Human Factors in Aviation Accidents with Natural Language Processing and Machine Learning Models. *Aerospace*, 12.
- LE, D. V., MONTGOMERY, J., KIRKBY, K. C. & SCANLAN, J. 2018. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *Journal of Biomedical Informatics*, 86, 49-58.
- LEE, J. & CHUNG, H. 2018. A new methodology for accident analysis with human and system interaction based on FRAM: Case studies in maritime domain. *Safety Science*, 109, 57-66.
- LEE, J., YOON, W. C. & CHUNG, H. 2020. Formal or informal human collaboration approach to maritime safety using FRAM. *Cognition, Technology and Work*, 22, 861-875.

- LEE, S., MOH, Y. B., TABIBZADEH, M. & MESHKATI, N. 2017. Applying the AcciMap methodology to investigate the tragic Sewol Ferry accident in South Korea. *Applied Ergonomics*, 59, 517-525.
- LEURS, W. L. M., LAMMERS, L. A. S., COMPAGNER, W. N., GROENEVELD, M., KORSTEN, E. H. H. M. & VAN DER LINDEN, C. M. J. 2022. Text mining in nursing notes for text characteristics associated with in-hospital falls in older adults: A case-control pilot study. *Aging and Health Research*, 2, 100078-100078.
- LEVESON, N. 2004. A new accident model for engineering safer systems. *Safety Science*, 42, 237-270.
- LI, P., CHEN, S., YUE, L., XU, Y. & NOYCE, D. A. 2024a. Analyzing relationships between latent topics in autonomous vehicle crash narratives and crash severity using natural language processing techniques and explainable XGBoost. *Accident Analysis & Prevention*, 203, 107605.
- LI, S., YOU, M., LI, D. & LIU, J. 2022a. Identifying coal mine safety production risk factors by employing text mining and Bayesian network techniques. *Process Safety and Environmental Protection*, 162, 1067-1081.
- LI, W., WU, P., HUANG, J. & XU, Y. 2024b. A new paradigm for construction safety management in China: Introducing knowledge graph and accident database into the early-stage of BIM. *Journal of Cleaner Production*, 470, 143367.
- LI, Y., CHENG, Z., YIP, T. L., FAN, X. & WU, B. 2022b. Use of HFACS and Bayesian network for human and organizational factors analysis of ship collision accidents in the Yangtze River. *Maritime Policy and Management*, 49, 1169-1183.
- LI, Y., SHYR, C., BORYCKI, E. M. & KUSHNIRUK, A. W. 2021. Automated thematic analysis of health information technology (HIT) related incident reports. *Knowledge Management and E-Learning*, 13, 408-420.
- LI, Z. & ZHAO, J. 2025. HAZOPCT: A HAZOP analysis completeness tool based on knowledge graph reasoning. *Process Safety and Environmental Protection*, 197, 107025.
- LIANG, C., ZHOU, S., YAO, B., HOOD, D. & GONG, Y. 2020. Toward systems-centered analysis of patient safety events: Improving root cause analysis by optimized incident classification and information presentation. *International Journal of Medical Informatics*, 135, 104054-104054.
- LIN, M. S.-M. & LU, B.-S. L. 2023. Risk assessment and management in the offshore wind power industry: A focus on component handling operations in ports. *Safety Science*, 167.
- LIN, S., LI, Y., HU, Y., SHI, J., WANG, S., WANG, P., MEI, Q., LI, X. & WEI, H. 2025. Research on ship dynamic feature extraction and prediction method based on visual data. *Ocean Engineering*, 327, 120938.
- LIU, C. & YANG, S. 2022. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Systems with Applications*, 207, 117991-117991.
- LIU, C. & YANG, S. 2023. A text mining-based approach for understanding Chinese railway incidents caused by electromagnetic interference. *Engineering Applications of Artificial Intelligence*, 117, 105598.
- LIU, D. & CHENG, L. 2024. MAKG: A maritime accident knowledge graph for intelligent accident analysis and management. *Ocean Engineering*, 312, 119280.
- LIU, G., BOYD, M., YU, M., HALIM, S. Z. & QUDDUS, N. 2021. Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Safety and Environmental Protection*, 152, 37-46.
- LIU, H., HU, M. & YANG, L. 2024. A new risk level identification model for aviation safety. *Engineering Applications of Artificial Intelligence*, 136, 108901.
- LIU, J., JI, L., CHEN, K., LI, C. & DUAN, H. 2025a. Railway operational hazard prediction and control based on knowledge graph embedding and topological analysis. *Reliability Engineering & System Safety*, 258, 110917.
- LIU, J., WONG, Z. S. Y., TSUI, K. L., SO, H. Y. & KWOK, A. 2019. Exploring hidden in-hospital fall clusters from incident reports using text analytics. *Studies in Health Technology and Informatics*, 264, 1526-1527.

- LIU, Q., LI, F., NG, K. K. H., HAN, J. & FENG, S. 2025b. Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety. *Expert Systems with Applications*, 269, 126422.
- LIU, S., WANG, F., XU, Y. & YIN, Z. 2022. *Analysis of human factors in marine collision accidents based on HFACS*, SPIE.
- LOUGHNEY, S., NGWOKE, K., WANG, J., YILDIZ, S. & UĞURLU, Ö. INVESTIGATION AND EVALUATION OF MARINE ACCIDENTS IN TERMS OF GROUNDING AND CONTACTS/COLLISIONS IN THE ENGLISH CHANNEL UTILISING THE HFACS-PV APPROACH. *In*: LEVA, M. C., PATELLI, E., PODOFILLINI, L. & WILSON, S., eds. *Proceedings of the 32nd European Safety and Reliability Conference, ESREL 2022 - Understanding and Managing Risk and Reliability for a Sustainable Future*, 2022. Research Publishing, 544-551.
- LU, Y., SUN, L. & YANG, X. 2021. Research on a comprehensive maintenance optimization strategy for an offshore wind farm. *Energies*, 14, 965.
- LUO, X., LI, X., GOH, Y. M., SONG, X. & LIU, Q. 2023. Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents. *Safety Science*, 163, 106138.
- MA, L., CHEN, L., MA, X., WANG, T. & ZHANG, J. 2024a. Incorporating human and organizational failures into the formation pattern for different Arctic maritime accidents using a data-driven Bayesian network. *Ocean Engineering*, 312, 119125.
- MA, L., MA, X., LAN, H., LIU, Y. & DENG, W. 2022. A data-driven method for modeling human factors in maritime accidents by integrating DEMATEL and FCM based on HFACS: A case of ship collisions. *Ocean engineering*, 266, 112699.
- MA, L., MA, X., WANG, T., CHEN, L. & LAN, H. 2024b. On the development and measurement of human factors complex network for maritime accidents: A case of ship groundings. *Ocean & coastal management*, 248, 106954.
- MA, L., MA, X., WANG, T., ZHAO, Y. & LAN, H. 2024c. A data-driven approach to determine the distinct contribution of human factors to different types of maritime accidents. *Ocean Engineering*, 295, 116874.
- MA, Y., LIU, Q. & YANG, L. 2024d. A seafarers' cognitive error mining model based on IDA and TRACER and mitigation measures for ship collision accidents. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*.
- MA, Z. & CHEN, Z.-S. 2024. Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Automation in Construction*, 161, 105343.
- MACÊDO, J. B., DAS CHAGAS MOURA, M., AICHELE, D. & LINS, I. D. 2022. Identification of risk features using text mining and BERT-based models: Application to an oil refinery. *Process Safety and Environmental Protection*, 158, 382-399.
- MACEDO, J. B., RAMOS, P. M. S., MAIOR, C. B. S., MOURA, M. J. C., LINS, I. D. & VILELA, R. F. T. 2022. Identifying low-quality patterns in accident reports from textual data. *INTERNATIONAL JOURNAL OF OCCUPATIONAL SAFETY AND ERGONOMICS*.
- MACEDO, J. B., RAMOS, P. M. S., MAIOR, C. B. S., MOURA, M. J. C., LINS, I. D. & VILELA, R. F. T. 2023. Identifying low-quality patterns in accident reports from textual data. *INTERNATIONAL JOURNAL OF OCCUPATIONAL SAFETY AND ERGONOMICS*, 29, 1088-1100.
- MACKENZIE, A., TESKE, A., ABIELMONA, R. & PETRIU, E. Maritime Incident Information Extraction using Machine and Deep Learning Techniques. 2021 IEEE SYMPOSIUM SERIES ON COMPUTATIONAL INTELLIGENCE (IEEE SSCI 2021), 2021 2021.
- MADEIRA, T., MELÍCIO, R., VALÉRIO, D. & SANTOS, L. 2021. Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace*, 8, 1-18.
- MAIB, M. A. I. B. 2024. 2023 Marine Accident Recommendations and Statistics.
- MALIK, N., ALTAf, S., TARIQ, M. U., AHMED, A. & BABAR, M. 2023. A Deep Learning Based Sentiment Analytic Model for the Prediction of Traffic Accidents. *Computers, Materials and Continua*, 77, 1599-1615.

- MARINE, C. 2003. Safety Pyramid based on a study.
- MATEROVÁ, A. & MATERNA, M. 2023. Research of maritime accidents based on HFACS framework. *Transportation Research Procedia*, 74, 1224-1231.
- MATHEW, F., WANG, H., MONTGOMERY, L. & KILDEA, J. 2021. Natural language processing and machine learning to assist radiation oncology incident learning. *JOURNAL OF APPLIED CLINICAL MEDICAL PHYSICS*, 22, 172-184.
- MAZAHARI, A., MONTEWKA, J., NISULA, J. & KUJALA, P. 2015. Usability of accident and incident reports for evidence-based risk modeling - A case study on ship grounding reports. *SAFETY SCIENCE*, 76, 202-214.
- MERTES, P. M., MORGAND, C., BARACH, P., JURKOLOW, G., ASSMANN, K. E., DUFETELLE, E., SUSPLUGAS, V., ALAUDDIN, B., YAVORDIOS, P. G., TOURRES, J., DUMEIX, J.-M. & CAPDEVILA, X. 2024. Validation of a natural language processing algorithm using national reporting data to improve identification of anesthesia-related ADverse evENTs: The "ADVENTURE" study. *Anaesthesia Critical Care & Pain Medicine*, 43, 101390.
- METTE, J., VELASCO GARRIDO, M., HARTH, V., PREISSER, A. M. & MACHE, S. 2018. Healthy offshore workforce? A qualitative study on offshore wind employees' occupational strain, health, and coping. *BMC Public Health*, 18, 172.
- MINGYANG, G., MIAO, C., LIHAO, Y., ZHIHUI, Z., JIA, L. & ZHIYONG, C. 2025. Investigation of ship collision accident risk factors using BP-DEMATEL method based on HFACS-SCA. *Reliability Engineering & System Safety*, 257, 110875.
- MITOMO, N., HIKIDA, K., YOSHIMURA, K., NISHIZAKI, C. & TAKEMOTO, T. Common performance condition for marine accident - Experimental approach. International Conference on Emerging Trends in Engineering and Technology, ICETET, 2012. 100-104.
- MIYAMOTO, A., BENDARKAR, M. V. & MAVRIS, D. N. 2022. Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns. *Aerospace*, 9.
- MORAIS, C., YUNG, K. L., JOHNSON, K., MOURA, R., BEER, M. & PATELLI, E. 2022. Identification of human errors and influencing factors: A machine learning approach. *Safety Science*, 146, 105528.
- MOREL, G. & CHAUVIN, C. 2006. A socio-technical approach of risk management applied to collisions involving fishing vessels. *Safety Science*, 44, 599-619.
- MUGURO, J., NJERI, W., MATSUSHITA, K. & SASAKI, M. 2022. Road traffic conditions in Kenya: Exploring the policies and traffic cultures from unstructured user-generated data using NLP. *IATSS Research*, 46, 329-344.
- MUJTABA, G., SHUIB, L., RAJ, R. G., RAJANDRAM, R. & SHAIKH, K. 2018. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *Journal of Forensic and Legal Medicine*, 57, 41-50.
- NA, X. U., LING, M. A., LIU, Q., LI, W. A. N. G. & DENG, Y. 2021. An improved text mining approach to extract safety risk factors from construction accident reports. *Safety Science*, 138, 105216-105216.
- NAKAMURA, M., HAYAMIZU, S., MASANORI, H., FUSEYA, T., IWAMATSU, H. & TERADA, K. 2024. Causal Reasoning of Occupational Incident Texts Using Large Language Models. *Procedia Computer Science*, 246, 820-829.
- NAKATA, T. Text-mining on incident reports to find knowledge on industrial safety. 2017 2017.
- NAKATA, T. & SOHRAB, M. 2018. Detection of typical progress patterns of industrial incidents by text mining technique. *Advances in Intelligent Systems and Computing*, 589, 221-229.
- NASUR, J., BOGUSŁAWSKI, K., WOLSKA, P., GIL, M. & WRÓBEL, K. 2025. Toward modeling emergency unmooring of manned and autonomous ships – A combined FRAM+HFACS-MA approach. *Safety Science*, 181, 106676.
- NAVAS DE MAYA, B., KHALID, H. & KURT, R. E. 2021. Application of card-sorting approach to classify human factors of past maritime accidents. *Maritime Policy and Management*, 48, 75-90.

- NGUYEN, M., BEIDLER, P., LYBARGER, K., ANDERSON, A., HOLMBERG, O., KANG, J. & FORD, E. C. 2022. Automatic Prediction of Severity Score of Incident Learning Reports in Radiation Oncology Using Natural Language Processing. *International Journal of Radiation Oncology*Biological*Physics*, 114, S93-S94.
- NI, Z.-J. & ZHANG, Q. 2023. Mining the Critical Factors of Hazardous Material Road Transportation Accidents From Chinese Investigation Reports. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 24, 11121-11131.
- NIELSEN, D., CLARKE, S. S. B. & KALYANAM, K. M. 2024. Towards an Aviation Large Language Model by Fine-tuning and Evaluating Transformers.
- NURDUHAN, M. & KULEYIN, B. 2024. Cluster-based Visualization of human element interactions in marine accidents. *Ocean engineering*, 298.
- OLIAEE, A. H., DAS, S., LIU, J. & RAHMAN, M. A. 2023. Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. *Natural Language Processing Journal*, 3, 100007.
- ONG, M. S., MAGRABI, F. & COIERA, E. 2010. Automated categorisation of clinical incident reports using statistical text classification. *Quality and Safety in Health Care*, 19.
- ONO, M. & NAKANISHI, M. 2023. Analysis of Human Factors and Resilience Competences in ASRS Data Using Natural Language Processing. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14029 LNCS, 548 – 561.
- ORGANISATION, I. M. 2000. GUIDELINES ON ERGONOMIC CRITERIA FOR BRIDGE EQUIPMENT AND LAYOUT. London.
- PAN, X., ZHONG, B., WANG, Y. & SHEN, L. 2022. Identification of accident-injury type and bodypart factors from construction accident reports: A graph-based deep learning framework. *Advanced Engineering Informatics*, 54, 101752-101752.
- PAOLO, F., GIANFRANCO, F., LUCA, F., MARCO, M., ANDREA, M., FRANCESCO, M., VITTORIO, P., MATTIA, P. & PATRIZIA, S. 2021. Investigating the Role of the Human Element in Maritime Accidents using Semi-Supervised Hierarchical Methods. *Transportation Research Procedia*, 52, 252-259.
- PARIKH, P., PENFIELD, J. & JUAIRE, M. 2024. Automatic identification of incidents involving potential serious injuries and fatalities (PSIF). *Scientific Reports*, 14.
- PENCE, J., FARSHADMANESH, P., KIM, J., BLAKE, C. & MOHAGHEGH, Z. 2020. Data-theoretic approach for socio-technical risk analysis: Text mining licensee event reports of U.S. nuclear power plants. *Safety Science*, 124, 104574-104574.
- PENG, C., ZHEN, X. & HUANG, Y. 2023. Human-automation interaction centered approach based on FRAM for systemic safety analysis of dynamic positioning operations for offshore tandem offloading. *Ocean Engineering*, 267, 113249.
- PERBOLI, G., GAJETTI, M., FEDOROV, S. & GIUDICE, S. L. 2021. Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology. *Expert systems with applications*, 186, 115694.
- PEREIRA, F. C., RODRIGUES, F. & BEN-AKIVA, M. 2013. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37, 177-192.
- QI, Y., ZHAO, X., KHASTGIR, S. & HUANG, X. 2025. Safety analysis in the era of large language models: A case study of STPA using ChatGPT. *Machine Learning with Applications*, 19, 100622.
- QIAO, W., LIU, Y., MA, X. & LIU, Y. 2020. A methodology to evaluate human factors contributed to maritime accident by mapping fuzzy FT into ANN based on HFACS. *Ocean Engineering*, 197, 106892.
- QIAO, W., YANG, J., ZHAO, Y., DENG, W. & MA, X. 2024. On the determination of the maritime-specific EPC values in reducing human factors based on maritime foundering accidents in China. *Ocean Engineering*, 307, 118192.

- RAMOS, P. M. S., MACEDO, J. B., MAIOR, C. B. S., MOURA, M. C. & LINS, I. D. 2024. Combining BERT with numerical variables to classify injury leave based on accident description. *PROCEEDINGS OF THE INSTITUTION OF MECHANICAL ENGINEERS PART O-JOURNAL OF RISK AND RELIABILITY*, 238, 903+.
- REASON, J. 1990. Human error
- REN, T., ZHANG, Z., JIA, B. & ZHANG, S. 2025. Retrieval-Augmented Generation-aided causal identification of aviation accidents: A large language model methodology. *Expert Systems with Applications*, 278, 127306.
- ROSE, R. L., PURANIK, T. G., MAVRIS, D. N. & RAO, A. H. 2022. Application of structural topic modeling to aviation safety data. *Reliability Engineering & System Safety*, 224, 108522-108522.
- SABETTA, N., COSTANTINO, F. & STABILE, S. 2025. A comparative analysis for automated information extraction from OSHA Lockout/Tagout accident narratives with Large Language Model. *Procedia Computer Science*, 253, 1362-1372.
- SAHOO, S., MAITI, J. & TEWARI, V. K. 2024a. A framework to model contractors' hazard and risk exposure at process plants using unsupervised text mining. *Process Safety and Environmental Protection*, 183, 24-45.
- SAHOO, S., MUKANE, P., MAITI, J. & TEWARI, V. K. 2024b. A framework for process risk assessment incorporating prior hazard information in text mining models using chunking. *Process Safety and Environmental Protection*, 189, 486-504.
- SALIHOGU, E. & BAL BEŞİKÇİ, E. 2021. The use of Functional Resonance Analysis Method (FRAM) in a maritime accident: A case study of Prestige. *Ocean Engineering*, 219, 108223.
- SANDHÅLAND, H., OLTEDAL, H. & EID, J. 2015. Situation awareness in bridge operations – A study of collisions between attendant vessels and offshore facilities in the North Sea. *Safety Science*, 79, 277-285.
- SARIALIOĞLU, S., UĞURLU, Ö., AYDIN, M., VARDAR, B. & WANG, J. 2020. A hybrid model for human-factor analysis of engine-room fires on ships: HFACS-PV&FFTA. *Ocean Engineering*, 217, 107992-107992.
- SARKAR, S., PRAMANIK, A. & MAITI, J. 2023. An integrated approach using rough set theory, ANFIS, and Z-number in occupational risk prediction. *Engineering Applications of Artificial Intelligence*, 117, 105515.
- SCHRÖDER-HINRICHS, J. U., BALDAUF, M. & GHIRXI, K. T. 2011. Accident investigation reporting deficiencies related to organizational factors in machinery space fires and explosions. *Accident Analysis & Prevention*, 43, 1187-1196.
- SEYR, H. & MUSKULUS, M. 2016. Safety Indicators for the Maritime Operations in the Installation and Operating Phase of an Offshore Wind Farm. *Energy Procedia*, 94, 72-81.
- SHAFIEE, M. & ADEDIPE, T. 2022. A bayesian network model for the probabilistic safety assessment of offshore wind decommissioning. *Wind Engineering. Wind engineering*, 47, 104-125.
- SHAPPELL, S. & WIEGMANN, D. 2000. The Human Factors Analysis and Classification System–HFACS
- SHEN, L., LEVIE, A., SINGH, H., MURRAY, K. & DESAI, S. 2022. Harnessing Event Report Data to Identify Diagnostic Error During the COVID-19 Pandemic. *The Joint Commission Journal on Quality and Patient Safety*, 48, 71-80.
- SHI, X., ZHUANG, H. & XU, D. 2021. Structured survey of human factor-related maritime accident research. *Ocean Engineering*, 237, 109561-109561.
- SHIQI, F. & ZAILI, Y. 2024. Accident data-driven human fatigue analysis in maritime transport using machine learning. *Reliability Engineering & System Safety*, 241, 109675.
- SHOME, A., ALAM, M. M., JANNATI, S. & BAIRAGI, A. K. 2024. Monitoring human behaviour during pandemic — Attacks on healthcare personnel scenario. *Telematics and Informatics Reports*, 15, 100149.
- SHORROCK, S. T. & KIRWAN, B. 2002. Development and application of a human error identification tool for air traffic control. *Applied Ergonomics*, 319-336.

- SHRESTHA, S., MORSHED, S. A., PRADHANANGA, N. & LV, X. Leveraging Accident Investigation Reports as Leading Indicators of Construction Safety Using Text Classification. *In: ELASMAR, M., GRAU, D. & TANG, P. B., eds. CONSTRUCTION RESEARCH CONGRESS 2020: SAFETY, WORKFORCE, AND EDUCATION, 2020 2020. 490-498.*
- SIDDESHWAR, V., AZIM, A., ALWIDIAN, S. & MAKREHCHI, M. 2024. Towards Enhancing Aviation Safety Through Advanced Incident Analysis Using Large Language Models.
- SINGLE, J. I., SCHMIDT, J. & DENECKE, J. 2020. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. *Safety Science*, 129, 104747-104747.
- SOBHAN ASIAN, GURDAL ERTEK, CAGRI HAKSOZ, SENA PARKER & ULUN, S. 2017. Wind Turbine Accidents: A data mining study. *IEEE Systems Journal* 11.
- SOOMER, H., RANTA, H. & PENTTILA, A. 2001. Identification of victims from the m/s estonia. *International Journal of Legal Medicine*, 114, 259-262.
- SOTIRALIS, P., VENTIKOS, N. P., HAMANN, R., GOLYSHEV, P. & TEIXEIRA, A. P. 2016. Incorporation of human factors into ship collision risk models focusing on human centred design aspects. *Reliability Engineering & System Safety*, 156, 210-227.
- STEVENS, S. & PARSONS, M. 2002. Effects of Motion at Sea on Crew Performance: A Survey. *Marine Technology*, 39, 29-47.
- STROEVE, S., KIRWAN, B., TURAN, O., KURT, R. E., VAN DOORN, B., SAVE, L., JONK, P., NAVAS DE MAYA, B., KILNER, A., VERHOEVEN, R., FARAG, Y. B. A., DEMIRAL, A., BETTIGNIES-THIEBAUX, B., DE WOLFF, L., DE VRIES, V., AHN, S. I. & POZZI, S. 2023. SHIELD Human Factors Taxonomy and Database for Learning from Aviation and Maritime Safety Occurrences. *Safety (Basel)*, 9, 14.
- SUFI, F. 2024. An innovative GPT-based open-source intelligence using historical cyber incident reports. *Natural Language Processing Journal*, 7, 100074.
- SUH, Y. 2021. Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. *Safety Science*, 142, 105363-105363.
- SVEDUNG, I. & RAMSMUSSEN, J. 2002. Graphic representation of accident scenarios: mapping system structure and the causation of accidents. *Safety Science*, 40, 397-417.
- SYED, K., SLEEMAN, W., HAGAN, M., PALTA, J., KAPOOR, R. & GHOSH, P. 2020. Automatic Incident Triage in Radiation Oncology Incident Learning System. *HEALTHCARE*, 8.
- TABAIE, A., TRAN, A., CALABRIA, T., BENNETT, S. S., MILICIA, A., WEINTRAUB, W., GALLAGHER, W. J., YOSAITIS, J., SCHUBEL, L. C., HILL, M. A., SMITH, K. M. & MILLER, K. 2024. Evaluation of a Natural Language Processing Approach to Identify Diagnostic Errors and Analysis of Safety Learning System Case Review Data: Retrospective Cohort Study. *Journal of Medical Internet Research*, 26.
- TANG, Z., PAN, X. & GU, Z. 2024. Analyzing public demands on China's online government inquiry platform: A BERTopic-Based topic modeling study. *PLOS ONE*, 19, e0296855.
- TANGUY, L., TULECHKI, N., URIELI, A., HERMANN, E. & RAYNAL, C. 2016. Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in industry*, 78, 80-95.
- TAWFIK, N. S. & SPRUIT, M. R. 2020. Evaluating sentence representations for biomedical text: Methods and experimental results. *JOURNAL OF BIOMEDICAL INFORMATICS*, 104.
- TINGJIANG, T., ENYUAN, W., KE, Z. & CHANGFANG, G. 2023. Research on assisting coal mine hazard investigation for accident prevention through text mining and deep learning. *Resources Policy*, 85, 103802.
- TIRUNAGARI, S. 2015. Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports. *arXiv*.
- TIXIER, A. J. P., HALLOWELL, M. R., RAJAGOPALAN, B. & BOWMAN, D. 2016. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62, 45-56.

- TOYABE, S. I. 2012. Detecting inpatient falls by using natural language processing of electronic medical records. *BMC Health Services Research*, 12.
- UDDIN, M. I. & AWAL, Z. I. 2020. Systems-theoretic approach to safety of inland passenger ship operation in Bangladesh. *Safety Science*, 126.
- UĞURLU, Ö., YILDIZ, S., LOUGHNEY, S. & WANG, J. 2018. Modified human factor analysis and classification system for passenger vessel accidents (HFACS-PV). *Ocean Engineering*, 161, 47-61.
- UĞURLU, Ö., YILDIZ, S., LOUGHNEY, S., WANG, J., KUNTCHULIA, S. & SHARABIDZE, I. 2020. Analyzing Collision, Grounding, and Sinking Accidents Occurring in the Black Sea Utilizing HFACS and Bayesian Networks. *Risk Analysis*, 40, 2610-2638.
- UNCTAD. 2025. *Shipping data: UNCTAD releases new seaborne trade statistics* [Online]. Available: <https://unctad.org/news/shipping-data-unctad-releases-new-seaborne-trade-statistics> [Accessed].
- UNG, S. T. 2018. Human error assessment of oil tanker grounding. *Safety Science*, 104, 16-28.
- VALCAMONICO, D., BARALDI, P., ZIO, E., DECARLI, L., CRIVELLARI, A. & ROSA, L. L. 2024. Combining natural language processing and bayesian networks for the probabilistic estimation of the severity of process safety events in hydrocarbon production assets. *Reliability Engineering & System Safety*, 241, 109638.
- VELASCO GARRIDO, M., METTE, J., MACHE, S., HARTH, V. & PREISSER, A. M. 2018. A cross-sectional survey of physical strains among offshore wind farm workers in the German exclusive economic zone. *BMJ OPEN*, 8, e020157-e020157.
- WALI, B., KHATTAK, A. J. & AHMAD, N. 2021. Injury severity analysis of pedestrian and bicyclist trespassing crashes at non-crossings: A hybrid predictive text analytics and heterogeneity-based statistical modeling approach. *Accident Analysis & Prevention*, 150, 105835-105835.
- WANG, B. & ZHAO, J. 2022. Automatic frequency estimation of contributory factors for confined space accidents. *Process Safety and Environmental Protection*, 157, 193-207.
- WANG, F. & GU, W. 2022. Intelligent HAZOP analysis method based on data mining. *Journal of Loss Prevention in the Process Industries*, 80, 104911-104911.
- WANG, H., CHEN, N., WU, B. & GUEDES SOARES, C. 2024a. Human and organizational factors analysis of collision accidents between merchant ships and fishing vessels based on HFACS-BN model. *Reliability Engineering & System Safety*, 249, 110201.
- WANG, W., ZHAO, J., CHEN, Y., SHAO, P. & JIA, P. 2025. Process safety enhancement in maritime operations: A Bayesian network-based risk assessment framework for the RCEP region. *Process Safety and Environmental Protection*, 194, 1235-1256.
- WANG, Y., COIERA, E. & MAGRABI, F. 2020. Can Unified Medical Language System-based semantic representation improve automated identification of patient safety incident reports by type and severity? *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, 27, 1502-1509.
- WANG, Y., SHENG, K., NIU, P., CHU, C., LI, M. & JIA, L. 2024b. A comprehensive analysis method of urban rail transit operation accidents and safety management strategies based on text big data. *Safety Science*, 172, 106400.
- WELLS, P. 2017. The iconic Torrey Canyon oil spill of 1967 - Marking its legacy. *Marine Pollution Bulletin*, 115, 1-2.
- WICAKSANA, I. M. R., FUCHI, M., KONISHI, T. & ADHITA, I. G. M. S. 2025. Resilience analysis of ship near-miss accident using FRAM. *WMU Journal of Maritime Affairs*.
- WILLIAMS, T. & BETAK, J. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. *Procedia Computer Science*, 130, 98-102.
- WONG, Z. S. Y., SO, H. Y., KWOK, B. S. C., LAI, M. W. S. & SUN, D. T. F. 2020. Medication-rights detection using incident reports: A natural language processing and deep neural network approach. *Health Informatics Journal*, 26, 1777-1794.

- WONG, Z. S. Y., WATERS, N., LIU, J. & USHIRO, S. 2024. A large dataset of annotated incident reports on medication errors. *Scientific Data*, 11.
- WU, B., YIP, T. L., YAN, X. & GUEDES SOARES, C. 2022. Review of techniques and challenges of human and organizational factors analysis in maritime transportation. *Reliability Engineering & System Safety*, 219, 108249.
- XING, Y., WU, Y., ZHANG, S., WANG, L., CUI, H., JIA, B. & WANG, H. 2024. Discovering latent themes in aviation safety reports using text mining and network analytics. *International Journal of Transportation Science and Technology*, 16, 292-316.
- XINGRONG, G., QIAN QIAN, Z. & YIMING, G. 2025. Maritime accident causation: A spatiotemporal and HFACS-Based approach. *Ocean Engineering*, 340, 122329.
- XIONG, M., WANG, H., CHE, C. & SUN, M. 2024a. Application of text mining and coupling theory to depth cognition of aviation safety risk. *Reliability Engineering & System Safety*, 245, 110032.
- XIONG, M., WANG, H., WONG, Y. D. & HOU, Z. 2024b. Enhancing aviation safety and mitigating accidents: A study on aviation safety hazard identification. *Advanced Engineering Informatics*, 62, 102732.
- XU, H., LIU, Y., SHU, C.-M., BAI, M., MOTALIFU, M., HE, Z., WU, S., ZHOU, P. & LI, B. 2022. Cause analysis of hot work accidents based on text mining and deep learning. *Journal of Loss Prevention in the Process Industries*, 76, 104747-104747.
- XU, L., MAO, Z. & FENG, Z. 2023. Research on the Causal Chain of Marine Engine Accidents Based on HFACS.
- XU, P., WANG, Q., YE, Y., WONG, S. C. & ZHOU, H. 2025. Text as data: Narrative mining of non-collision injury incidents on public buses by structural topic modeling. *Travel Behaviour and Society*, 39, 100981.
- XUE, N., ZHANG, W., ZHONG, H., LIAO, W. & ZHAO, T. 2025. How to realize the knowledge reuse and sharing from accident reports? A knowledge-driven modeling method combining ontology and deep learning. *Journal of Loss Prevention in the Process Industries*, 94, 105525.
- YAN, K., WANG, Y., JIA, L., WANG, W., LIU, S. & GENG, Y. 2023. A content-aware corpus-based model for analysis of marine accidents. *Accid Anal Prev*, 184, 106991-106991.
- YANG, S. & DEMICHELA, M. 2023. A data-driven narratives skeleton pattern recognition from accident reports dataset for human-and-organizational-factors analysis. *Journal of Loss Prevention in the Process Industries*, 83, 105047.
- YANG, Z. L., BONSALL, S., WALL, A., WANG, J. & USMAN, M. 2013. A modified CREAM to human reliability quantification in marine engineering. *Ocean engineering*, 58, 293-303.
- YILDIZ, S., UĞURLU, Ö., WANG, J. & LOUGHNEY, S. 2021. Application of the HFACS-PV approach for identification of human and organizational factors (HOFs) influencing marine accidents. *Reliability Engineering & System Safety*, 208, 107395-107395.
- YOKOTA, S., SHINOHARA, E. & OHE, K. 2018. Can staff distinguish falls: Experimental hypothesis verification using Japanese incident reports and natural language processing. *Studies in Health Technology and Informatics*, 250, 159-163.
- YOON, N. K., QUINN, T. D., FUREK, A., PAYNE, N. Y. & HAAS, E. J. 2025. Improving the usability of large emergency 911 data reporting systems: A machine learning case study using emergency incident descriptions. *Journal of Safety Research*, 93, 335-341.
- YOSHIMURA, K., TAKEMOTO, T. & MITOMO, N. The support for using the cognitive reliability and error analysis method (CREAM) for marine accident investigation. 2015 4th International Conference on Informatics, Electronics and Vision, ICIEV 2015, 2015. Institute of Electrical and Electronics Engineers Inc.
- YOSHIMURA, K., TAKEMOTO, T., MURATA, S. & MITOMO, N. A study for adapting a human reliability analysis technique to marine accidents. PSAM 2014 - Probabilistic Safety Assessment and Management, 2014. Techno-Info Comprehensive Solutions (TICS).

- YUAN, C., FU, G., ZHANG, J., LI, Y., HAN, M., LU, Y., ZHAO, J. & WU, Z. 2025. Causal relationship analysis and risk assessment of contributing factors in coal mine ordinary accidents: A case study (2002–2023). *Process Safety and Environmental Protection*, 197, 107065.
- YUHONG, W., PENGCHANG, L., CHENG, H. & ZAILI, Y. 2025. Causation analysis of ship collisions using a TM-FRAM model. *Reliability Engineering & System Safety*, 260, 111035.
- YUXIN, W., GUI, F., QIAN, L., JINGRU, W., YALI, W., MENG, H., YUXUAN, L. & XUECAI, X. 2024. Accident case-driven study on the causal modeling and prevention strategies of coal-mine gas-explosion accidents: A systematic analysis of coal-mine accidents in China. *Resources Policy*, 88, 104425.
- ZHANG, F., FLEYEH, H., WANG, X. & LU, M. 2019a. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238-248.
- ZHANG, H., CHEN, B., ZHAO, Q., YU, J. & FANG, Z. 2024. Identification of risk key factors and prevention strategies for collision accidents between merchant and fishing vessels in China waters based on complex network. *Ocean Engineering*, 307, 118148.
- ZHANG, M., ZHANG, D., GOERLANDT, F., YAN, X. & KUJALA, P. 2019b. Use of HFACS and fault tree model for collision risk factors analysis of icebreaker assistance in ice-covered waters. *Safety Science*, 111, 128-143.
- ZHANG, X., CHEN, W., XI, Y., HU, S. & TANG, L. 2020. Dynamics simulation of the risk coupling effect between maritime pilotage human factors under the hfacs framework. *Journal of Marine Science and Engineering*, 8.
- ZHANG, X., SRINIVASAN, P. & MAHADEVAN, S. 2021. Sequential deep learning from NTSB reports for aviation safety prognosis. *SAFETY SCIENCE*, 142.
- ZHANG, Z., YANG, X., SUN, L., SUN, Y. & KANG, J. 2025. Research on constructing and reasoning the collision knowledge graph of autonomous navigation ship based on enhanced BERT model. *Expert Systems with Applications*, 278, 127429.
- ZHOU, Z., YU, X., MAGOUA, J. J., CUI, J., LUAN, H. & LIN, D. 2025. Integrating machine learning and a large language model to construct a domain knowledge graph for reducing the risk of fall-from-height accidents. *Accident Analysis & Prevention*, 215, 108009.
- ZHU, Y., LIAO, H. & HUANG, D. 2023. Using text mining and multilevel association rules to process and analyze incident reports in China. *Accident Analysis & Prevention*, 191, 107224.
- ZOU, Y., KIVINIEMI, A. & JONES, S. W. 2017. Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in Construction*, 80, 66-76.

Appendix A: Preliminary Study 1: Development of NLP Models for Offshore Wind Incident Classification

Appendix A.1 Purpose, dataset and pipeline

Prior to the development of the SHIELD-NLP framework presented in this thesis, an initial proof-of-concept study was undertaken to investigate the feasibility of applying Natural Language Processing (NLP) and supervised machine learning to offshore wind safety incident reports. The objective of the study was to determine whether incident narratives could be automatically classified into predefined safety categories using traditional machine learning techniques.

This preliminary investigation provided valuable experience in working with large collections of industrial safety reports and established the methodological foundations that informed the subsequent development of SHIELD-NLP and SeaSafeBERT. In particular, the study explored text preprocessing, feature engineering, model selection, and evaluation procedures that were later refined and extended within the main body of this thesis.

The study utilised safety incident reports supplied by G+, an international organisation representing offshore wind operators and contractors. The dataset contained 5,403 incident reports recorded between 2017 and 2021, covering construction, commissioning, operations and maintenance activities across multiple offshore wind projects.

Each incident consisted of a free-text narrative describing the event together with associated metadata and binary labels representing three independent safety classifications:

- High Potential (HiPo)
- Dropped Object
- Emergency Response / Medical Evaluation (ERME)

The objective of the study was to develop binary classification models capable of automatically predicting these safety attributes from the incident descriptions.

The preliminary study followed a structured NLP and supervised machine learning workflow (shown in Figure 43). The pipeline consisted of five stages: data preparation, text preprocessing, feature extraction, supervised classification, and performance evaluation. The workflow provided a systematic approach for investigating the suitability of conventional machine learning algorithms for safety incident classification before more advanced transformer-based methods were explored.

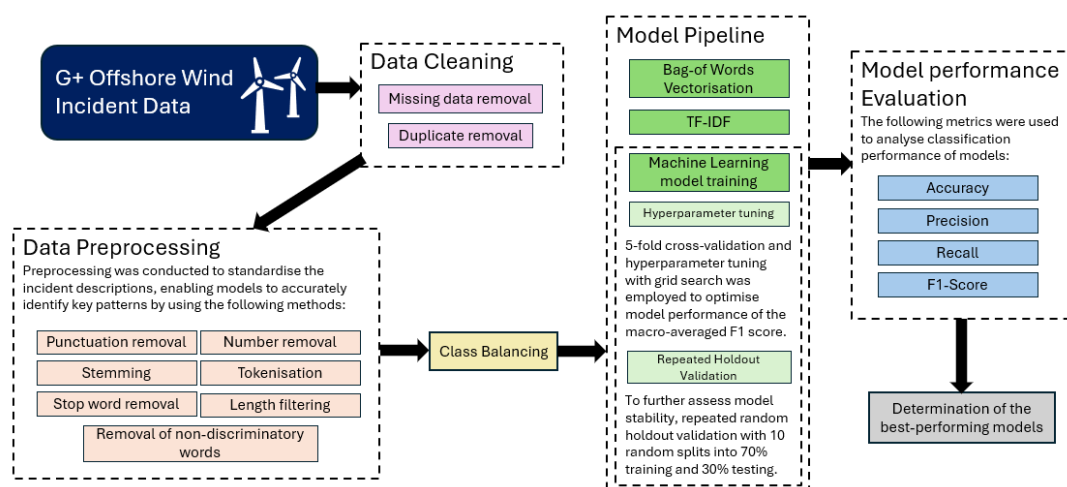


Figure 43. Flow chart of preliminary study methodology.

Prior to model development, the dataset underwent a series of cleaning and preprocessing procedures to prepare the textual data for analysis. Duplicate records and incomplete reports were removed before the incident narratives were standardised through lowercasing, punctuation removal, stop-word removal, tokenisation and stemming.

In addition to standard preprocessing techniques, common domain-specific terms that appeared frequently across all incident categories but provided limited discriminatory information were removed. Examples included frequently occurring operational terms such as *vessel*, *technician*, *CTV*, and *work*. This step reduced noise within the textual representations while retaining words that were more informative for classification.

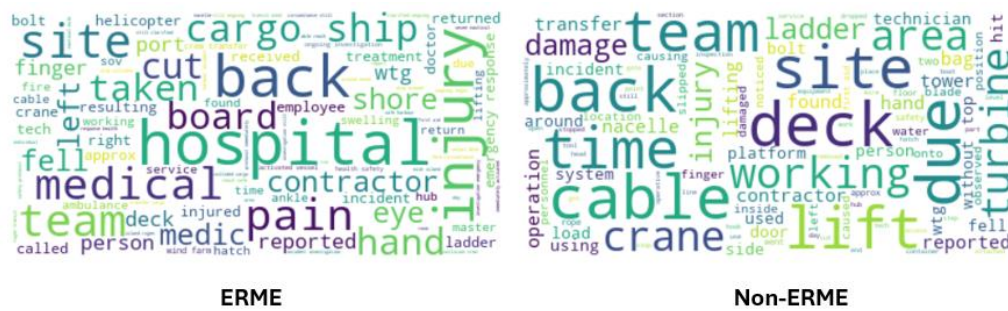


Figure 46. Word cloud of ERME and non-ERME incidents.

Following preprocessing, the textual data were transformed into numerical feature vectors using a Bag-of-Words representation combined with Term Frequency–Inverse Document Frequency (TF-IDF) weighting.

These feature vectors formed the input to the supervised machine learning algorithms evaluated during the study.

- Naïve Bayes
- Support Vector Machine (SVM)
- Logistic Regression
- Random Forest
- Decision Tree
- K-Nearest Neighbours (KNN)

Hyperparameter optimisation was performed using Grid Search combined with five-fold cross-validation. The optimised models were subsequently evaluated using an independent test dataset.

The objective of this comparison was not to identify a universally superior algorithm, but rather to establish baseline classification performance and investigate the suitability of conventional NLP approaches for safety incident classification.

Appendix A.2 Experimental results

The preliminary study demonstrated that machine learning techniques could successfully classify offshore wind incident narratives using textual information alone.

Across the three classification tasks, Support Vector Machine consistently achieved the strongest overall performance, particularly for the High Potential and Dropped Object classifications. Logistic Regression and Random Forest also produced competitive results, while Naïve Bayes and Decision Tree generally achieved lower classification performance.

The results confirmed that meaningful safety information could be extracted automatically from free-text incident reports using traditional NLP methods and established an important baseline for subsequent methodological development.

Tables 30-32 present the complete performance results for all evaluated algorithms.

Table 32. Model performance metrics for predicting high-potential incidents.

ML Algorithms	Performance Metrics			
	Accuracy	F1 Score	Precision	Recall
Naïve Bayes	0.7230	0.7396	0.7051	0.7782
SVM	0.7084	0.7218	0.6931	0.7538
KNN	0.6772	0.6900	0.6667	0.7164
Logistic Regression	0.7046	0.7202	0.6868	0.7576
Random Forest	0.6740	0.6622	0.6915	0.6402
Decision Tree	0.6289	0.6363	0.6263	0.6469

Table 33. Model performance metrics for predicting dropped object incidents.

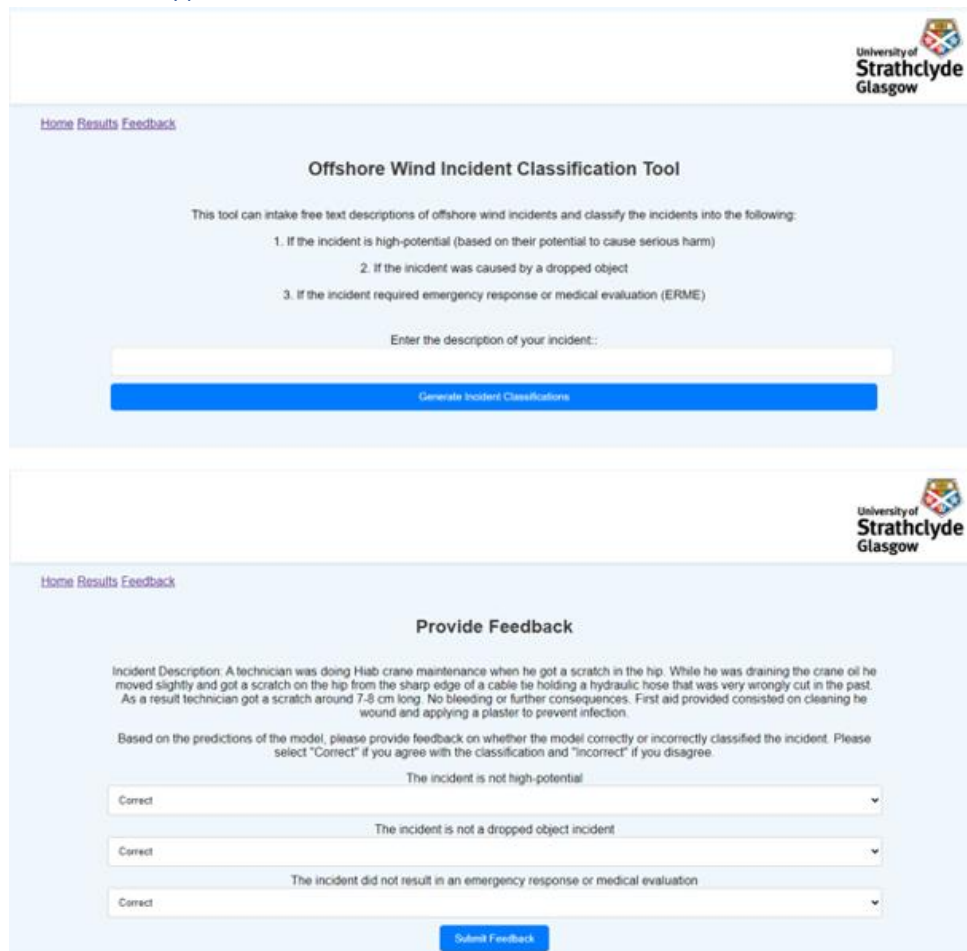
ML Algorithms	Performance Metrics			
	Accuracy	F1 Score	Precision	Recall
Naïve Bayes	0.7772	0.8008	0.7359	0.8809

SVM	0.8211	0.8211	0.8245	0.8188
KNN	0.7438	0.7683	0.7041	0.8462
Logistic Regression	0.8146	0.8175	0.8084	0.8278
Random Forest	0.8195	0.8211	0.8178	0.8255
Decision Tree	0.7633	0.7634	0.7681	0.7601

Table 34. Model performance metrics for predicting ERME incidents.

ML Algorithms	Performance Metrics			
	Accuracy	F1 Score	Precision	Recall
Naïve Bayes	0.7081	0.7369	0.6620	0.8472
SVM	0.7712	0.7461	0.8358	0.6767
KNN	0.6982	0.7243	0.6673	0.7957
Logistic Regression	0.7676	0.7574	0.7906	0.7299
Random Forest	0.7676	0.7453	0.8219	0.6834
Decision Tree	0.6928	0.6854	0.6740	0.6740

Appendix A.3 Prototype Offshore Wind Classification Tool



University of Strathclyde Glasgow

[Home](#) [Results](#) [Feedback](#)

Offshore Wind Incident Classification Tool

This tool can intake free text descriptions of offshore wind incidents and classify the incidents into the following:

1. If the incident is high-potential (based on their potential to cause serious harm)
2. If the incident was caused by a dropped object
3. If the incident required emergency response or medical evaluation (ERME)

Enter the description of your incident:

[Generate Incident Classifications](#)

University of Strathclyde Glasgow

[Home](#) [Results](#) [Feedback](#)

Provide Feedback

Incident Description: A technician was doing Hiab crane maintenance when he got a scratch in the hip. While he was draining the crane oil he moved slightly and got a scratch on the hip from the sharp edge of a cable tie holding a hydraulic hose that was very wrongly cut in the past. As a result technician got a scratch around 7-8 cm long. No bleeding or further consequences. First aid provided consisted on cleaning the wound and applying a plaster to prevent infection.

Based on the predictions of the model, please provide feedback on whether the model correctly or incorrectly classified the incident. Please select "Correct" if you agree with the classification and "Incorrect" if you disagree.

The incident is not high-potential

Correct

The incident is not a dropped object incident

Correct

The incident did not result in an emergency response or medical evaluation

Correct

[Submit Feedback](#)

Figure 47. Prototype tool interface for offshore wind classification.

Appendix A.3 Lessons Learned

The preliminary study demonstrated the feasibility of applying Natural Language Processing and supervised machine learning to industrial safety incident reports and provided several important insights that directly influenced the subsequent research presented in this thesis.

First, the study confirmed that textual incident narratives contain sufficient information to support automated safety classification. Second, it highlighted the importance of careful preprocessing and feature engineering when working with operational safety data. Third, although TF-IDF representations and traditional machine learning algorithms produced encouraging classification performance, they were limited in their ability to capture contextual meaning within complex safety narratives.

These observations motivated the development of the more advanced methodologies presented in the main body of the thesis. In particular, the limitations identified during this proof-of-concept study informed the transition from traditional feature engineering towards contextual transformer-based language models, ultimately leading to the development of SeaSafeBERT and the SHIELD-NLP framework. The experience gained through this preliminary investigation therefore represents an important stage in the evolution of the overall research.

Appendix B: Preliminary Study 2: Development of SHIELD-NLP

Appendix B.1 Purpose and preliminary classification pipeline

Following the successful proof-of-concept study on offshore wind incident classification (Appendix A), a second preliminary investigation was undertaken to explore the automated classification of human and causal factors using the SHIELD taxonomy.

Unlike the offshore wind study, which focused on predefined incident attributes, this investigation aimed to determine whether Natural Language Processing could support the classification of accident reports according to an established human factors framework. The study represented the first stage in the development of SHIELD-NLP and investigated taxonomy restructuring, preprocessing strategies, traditional machine learning algorithms, and prototype deployment before the development of the final framework presented in Chapter 5.

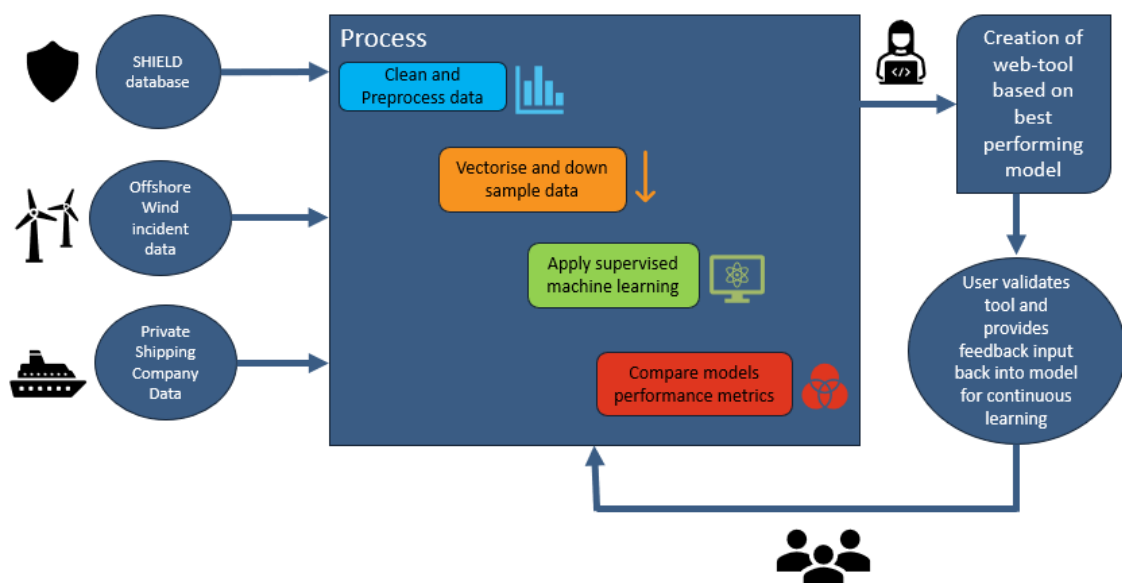


Figure 48. Methodology flow diagram of SHIELD preliminary study.

Figure 55 presents the classification pipeline developed during the preliminary SHIELD-NLP investigation. The workflow followed a structured Natural Language Processing (NLP) and supervised machine learning process, beginning with data preparation and concluding with automated SHIELD

classification. The pipeline comprised five main stages: data cleaning, text preprocessing, feature extraction, model development, and performance evaluation.

Initially, accident reports were cleaned to remove duplicate and incomplete records before the textual narratives underwent preprocessing. This included standardisation of the text through lowercasing, punctuation removal, tokenisation, stop-word removal and lemmatisation. Following preprocessing, the incident descriptions were transformed into numerical feature representations using Term Frequency–Inverse Document Frequency (TF-IDF), providing input features for model development.

Several supervised machine learning algorithms were subsequently trained and evaluated to classify incident reports according to the SHIELD taxonomy. Hyperparameter optimisation and cross-validation were performed during model development to identify the most suitable model configurations before evaluation using an independent test dataset. Model performance was assessed using standard classification metrics, including accuracy, precision, recall and F1-score.

This preliminary pipeline established the feasibility of automated SHIELD classification using conventional NLP techniques and formed the basis for the subsequent methodological developments presented in Chapter 5. In particular, the limitations identified during this investigation informed the transition towards contextual language models and the development of the final SHIELD-NLP framework described in the main body of the thesis.

Appendix B.2 Results of SHIELD-NLP preliminary study

One of the principal outcomes of the preliminary SHIELD study was the development of a revised SHIELD-NLP taxonomy to support automated human factor classification. The original SHIELD framework was adapted to improve its suitability for Natural Language Processing by refining the Acts layer into six clearly defined behavioural categories. As illustrated in Figure 56, the revised taxonomy comprised Inadequate Decision or Plan, Late/No/Wrong Detection, Timing Error, Poor Quality Work, Routine Violation (Workaround), and Exceptional Violation (Workaround). This refined structure

formed the basis for the machine learning models evaluated during the study and was subsequently adopted as the foundation for the SHIELD-NLP framework presented in the main body of this thesis.

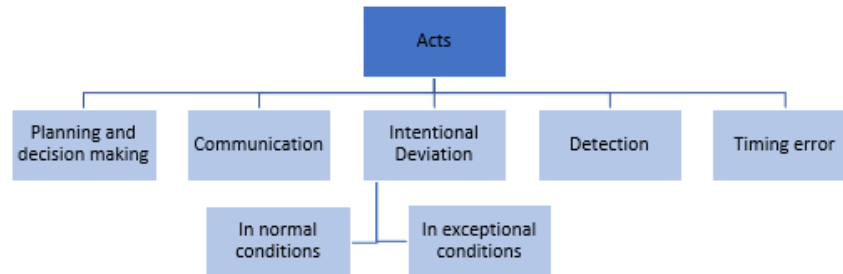


Figure 49. SHIELD-NLP Initial Acts Layer from preliminary study.

Following the development of the revised taxonomy, exploratory analysis of the incident narratives was undertaken using word cloud visualisations to examine the language associated with each SHIELD layer. The resulting word clouds highlighted distinctive terminology across the *Acts*, *Preconditions*, *Operational Leadership*, and *Organisational* classifications, demonstrating that each layer contained characteristic language patterns that could be exploited during model development. These visualisations also provided an initial assessment of the dataset and supported the refinement of the preprocessing pipeline.

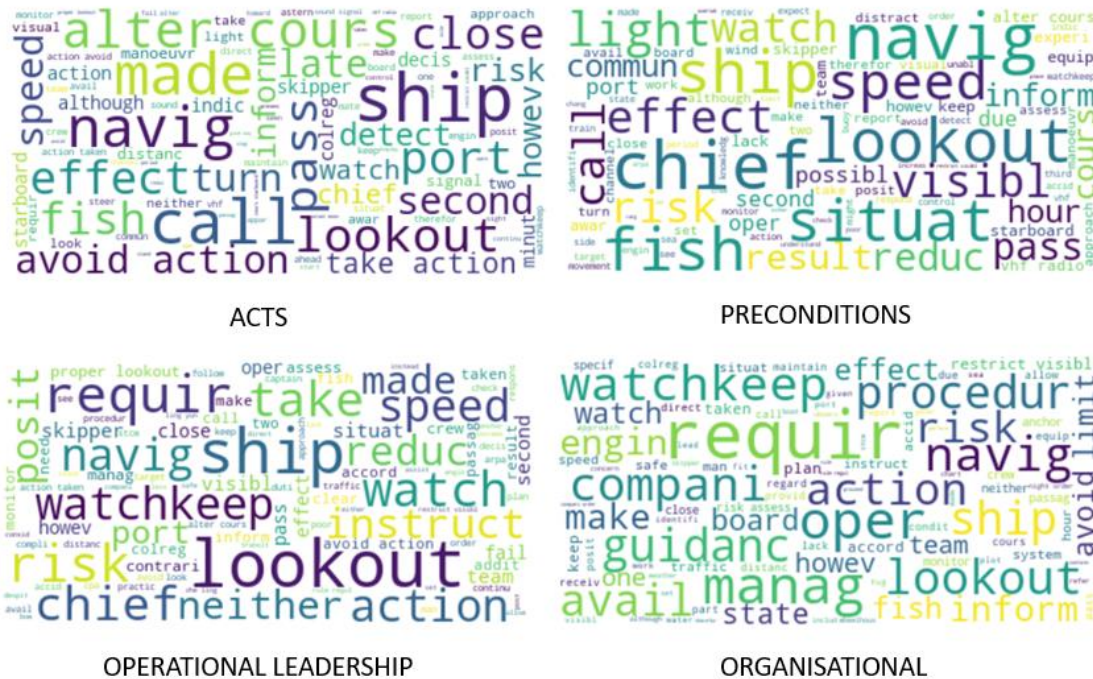


Figure 50. Word clouds, showing the most frequent words in each of the following layers: acts, preconditions, operational leadership and organisational.

The processed incident reports were subsequently used to evaluate six supervised machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, and K-Nearest Neighbours (KNN). Each model was trained and evaluated using the revised SHIELD-NLP taxonomy, with performance assessed using accuracy, precision, recall, and F1-score. The complete classification results for all evaluated algorithms are presented in Tables 33-38.

Across the evaluated models, Support Vector Machine and Naïve Bayes generally produced the strongest and most consistent classification performance, while Logistic Regression and Random Forest also demonstrated competitive results for several of the SHIELD categories. Decision Tree and K-Nearest Neighbours produced comparatively lower performance overall, although individual categories exhibited variation depending on the characteristics of the training data. Collectively, the results demonstrated that conventional NLP techniques could successfully automate elements of SHIELD classification and provided encouraging evidence for the application of machine learning to maritime human factor analysis.

Although the preliminary study demonstrated the feasibility of automated SHIELD classification, it also identified several limitations associated with conventional feature engineering and traditional machine learning approaches. In particular, the reliance on TF-IDF representations limited the ability of the models to capture contextual meaning within accident narratives, while performance varied across categories due to differences in data availability and linguistic complexity. These findings directly informed the subsequent development of the contextual language models and explainable SHIELD-NLP framework presented in Chapter 5.

Table 35. SVM Model performance metrics for all Acts.

SVM	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6803	0.6723	0.6780	0.6667
Inadequate transmission of information	0.76	0.775	0.775	0.775
Timing error	0.55	0.5263	0.625	0.4545
Late/no/wrong detection	0.6667	0.6761	0.6667	0.6857
Workaround in normal conditions	0.625	0.64	0.8	0.5333
Workaround in exceptional conditions	0.9286	0.9412	0.8889	1.0

Table 36. Naive Bayes Model performance metrics for all ACTS.

Naïve Bayes	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6885	0.7164	0.6486	0.8
Inadequate transmission of information	0.7733	0.7901	0.7805	0.8
Timing error	0.55	0.5263	0.625	0.4545
Late/no/wrong detection	0.6812	0.7027	0.6667	0.7429
Workaround in normal conditions	0.5417	0.5217	0.75	0.4
Workaround in exceptional conditions	0.7857	0.8	0.8571	0.75

Table 37. K-nearest neighbour model performance metrics for all Acts.

K-nearest neighbour	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6065	0.6065	0.5968	0.6167
Inadequate transmission of information	0.68	0.7333	0.66	0.825
Timing error	0.6	0.6	0.6667	0.5454
Late/no/wrong detection	0.5507	0.6075	0.5454	0.6857
Workaround in normal conditions	0.5	0.625	0.5882	0.6667
Workaround in exceptional conditions	0.9286	0.9412	0.8889	1.0

Table 38. Logistic regression model performance metrics for all Acts.

Logistic regression	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6639	0.6555	0.6610	0.65
Inadequate transmission of information	0.7333	0.7368	0.7778	0.7
Timing error	0.6	0.5556	0.7142	0.4545
Late/no/wrong detection	0.6522	0.6522	0.6522	0.6522
Workaround in normal conditions	0.5	0.3337	1.0	0.2
Workaround in exceptional conditions	0.7143	0.7143	0.8333	0.625

Table 39. Random forest model performance metrics for all Acts.

Random forest	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6475	0.5981	0.6809	0.5333
Inadequate transmission of information	0.76	0.7632	0.8056	0.725
Timing error	0.8	0.7778	1.0	0.6363
Late/no/wrong detection	0.6522	0.6363	0.6774	0.6
Workaround in normal conditions	0.5	0.3333	1.0	0.2
Workaround in exceptional conditions	0.7143	0.6667	1.0	0.5

Table 40. Decision tree model performance metrics for all Acts.

Decision Trees	Accuracy	F1 Score	Precision	Recall
Inadequate decision or plan	0.6066	0.6	0.6	0.6
Inadequate transmission of information	0.68	0.7	0.7	0.7
Timing error	0.8	0.7778	1.0	0.6363
Late/no/wrong detection	0.6667	0.6230	0.7308	0.5429
Workaround in normal conditions	0.5	0.4	0.8	0.2667
Workaround in exceptional conditions	0.8571	0.8571	1.0	0.75

Appendix B.3 Prototype Interface

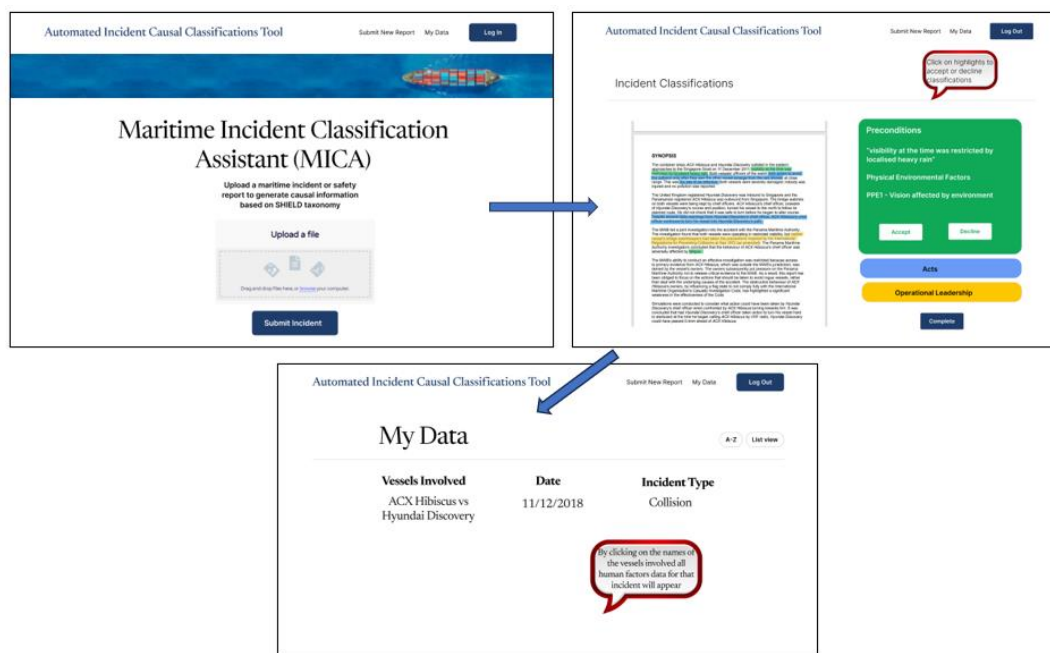


Figure 51. SHIELD Prototype interface preliminary study.

Appendix B.4 Lessons Learned

The preliminary SHIELD investigation demonstrated that traditional Natural Language Processing techniques could successfully automate elements of the SHIELD classification process. However, several limitations were identified.

The reliance on TF-IDF representations limited the ability of the models to capture contextual relationships within complex accident narratives.

Performance also varied across SHIELD layers due to differences in class balance and the complexity of human factor descriptions.

These observations motivated the transition towards contextual transformer-based language models and ultimately informed the development of SeaSafeBERT, and the complete SHIELD-NLP framework presented in Chapter 5.

Appendix C: SHIELD-NLP Development Tables

Table 41. Mapping of Original SHIELD Codes to SHIELD-NLP Codes

Original Code(s)	SHIELD-NLP Code
AP2 – Auditory detection, AP3 – Kinaesthetic detection, AP4 – Other senses	AP2 – No/wrong/late auditory or other detection
AR1 – Timing error, AR2 – Sequence error	AR1 – Timing or sequence error
AR3 – Right action/wrong object, AR4 – Wrong action/right object	AR3 – Action and Object Mismatch
OC1- Safety Culture, OC2 – Multi-cultural factors	OC1 – Safety culture and multi-cultural factors
OE1 – Contractors, OR1 – Personnel, OR4 – Inadequate training programme	OR1 – Personnel, contractors or inadequate training
OE3 – Economic pressure, OR2 – Budgets, OR3 – Materials availability	OR3 – Equipment/materials availability and/or budgets/economic pressures
PCO2 – Rank-based communication issues, PCO3 – Language difficulties, PCO4- Complex communication	PCO2 – Inadequate or complex communication and language difficulties
PPE2 – Movement affected, PPE3 – Hearing, PPE4 – Mental process, PPE5 – Heat/cold stress	PPE2 – Environmental stressors on human performance
PPF1 – Emotional state, PPF2 – Personality style, PPF3 – Confidence level, PPF5 – Motivation	PPF1 – Personal Factors
PME1 – Forget action/intentions, PME2 – No/inaccurate recall, PME – Negative habit	PME1- Memory
PER2 – Visual illusion, PER3 – Misperception of changing environment, PER4 – Misinterpreted or misread instrument	PER2 – Misperception and/or visual illusion
LO3 – Inappropriate pressure to perform a task, LO4 – Directed task with inadequate qualification, LO5 – Directed task with limited experience, LO6 – Directed task with inadequate equipment	LO3 – Unsafe supervisory direction/pressure to perform

LP1 – No personnel measures against regular risky behaviour, LP2 – Inappropriate behaviour affects learning, LP4 – Lack of feedback on safety reporting	LP1 - Inadequate Supervision
PEW2 – Technology creates an unsafe situation, PEW5 – Communication equipment inadequate, PEW6 – Fuels or materials	PEW2 - Technology and equipment creates an unsafe situation
PPE6 - Operation more difficult due to weather and environment, PPE5 - Heat or cold stress	PPE6 - Operation more difficult due to weather and environment
PTG1 - Team working towards different goals, PTG4 - No adaptation of team performance in demanding situation	PTG1 - Team working issues
PTG2 - No cross-check and speaking-up by team member, PTG3 - No monitoring and speaking-up of team status and functioning	PTG2 - No cross-check and speaking-up by team member

Table 42. Codes removed from analysis and reasons

Code	Count	Reason Removed
AI4 – Sabotage	1	Conceptually distinct, no plausible merging category
AR5 - Lack of physical coordination		
PAW6 – unsuitable mental model	1	Unique cognitive mechanism
PAW5 – Geographically lost	2	No semantic neighbour, extremely rare
PCS4 – Body/coordination limitations	1	Physiological factor; too isolated
PDN1 – Alcohol/drugs	10	Not conceptually aligned with other cognitive/perceptual codes
PMW3 – Information processing overload	1	Unique mechanism not represented elsewhere
LP3 – Personality conflict	3	Interpersonal issue; no meaningful merge
LT4 – Allowing unwritten policies to become standard	2	Rare organisational factor; too isolated
OE4 – Tempo of operations	5	Operational factor; no stable merge category

PAW7 – Pre-conceived notion or expectancy	11	Cognitive bias; insufficient support
PEW4 – Personal protective equipment inference	1	Highly specific; no merge option
PPC1 – Injury or illness existed during operation	4	Medical condition; conceptually isolated
PPC3 – Mentally exhausted (burnout)	6	Long-term psychological state; too rare
PPF4 – Performance/peer pressure	10	Social pressure; no clear merge