DATA-DRIVEN APPROACHES FOR RISK ASSESSMENT IN THE CHEMICAL PROCESSING INDUSTRY: LEVERAGING TEXTUAL AND NUMERICAL DATA

by

© Mohammad Zaid Kamil

A Thesis Submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

October 2023

St. John's Newfoundland and Labrador.

Abstract

Chemical process industries are accident-prone due to handling hazardous materials and the complex interaction of process operations. Industries, including chemical processing industries, are transitioning to digitalization with higher productivity potential by better managing process operations. A continuous encouragement to adopt digitalization in process industries while ensuring operational safety has led to new opportunities and challenges. The former relates to underpinning digital changes that will open new data generation and collection avenues, whereas the latter deals with translating the data into meaningful information.

Two data types will play a key role in dealing with this evolving challenge of translating data into meaningful information. First, structured data (numerical data) determine the behavior of process systems. Second, unstructured data from accident investigation reports for learning lessons is utilized. Conventional risk analysis techniques are incapable of dealing with the evolving challenge. Risk evaluation for process operations during this transition requires advanced technologies. This thesis proposes new approaches for safety 4.0, which is the introduction of industry 4.0 technologies such as artificial intelligence and automation to monitor risk. The approaches integrate artificial intelligence with data-driven models. These advanced techniques address the widely recognized knowledge gap in the literature and serve as an important tool for safety 4.0.

The thesis looks at developing approaches to gain insights from operational (contemporary) and textual (historical) data. First, a framework is developed to introduce a learning-based likelihood model. Structured data are used to model the topology of the Bayesian network (BN) and learn parameters from the data. Learning from data makes the model unique and allows capturing changes in operational data that are reflected in model output. A novel methodology is introduced to utilize field data of microbiologically influenced corrosion (MIC) in the

likelihood model. Second, unstructured data in textual form is transformed into objective risk assessment by employing natural language processing (NLP). A novel methodology is developed to gain insights from corrosion investigation reports assessing the risk of MIC in pipelines. The methodology attempts to give a new dimension to risk assessment by developing a cause-effect scenario from the textual data. A named entity recognition (NER) model is trained to gain insights and, based on the findings, transformed into a risk estimation BN model and evaluated using a risk matrix. Third, unstructured data are used to develop a generalized causation model. A systematic approach comprised of NER, interpretive structural modeling (ISM), and BN is proposed to gain insights from unstructured data. The output is a generalized causation model for oil and refining accidents that lead to fire and explosion. A hierarchical BN model is developed for fire and explosion from the CSB database to identify commonalities among different incidents. Finally, this thesis looks into the integration of structured and unstructured data. The methodology of integrating both data types is proposed to provide a comprehensive picture. Insights from multiple sources are key for robust risk analysis. The methodology proposed gains insights from unstructured data using a co-occurrence network. These insights integrate with contemporary data and establish each factor dependence using ISM. The resulting digraph from the ISM is mapped into a generalized hybrid BN model. Industrial and simulated datasets are used to test and verify the effectiveness of the developed model in predicting adverse events. This thesis develops important tools for enhanced data-driven prediction of adverse events.

Acknowledgement

In the name of Allah, the Most Beneficent and the Most Merciful. All praise to Allah alone, and many salutations upon His noble prophet Muhammad (peace be upon him).

I could not have accomplished this work without the exceptional assistance and unparalleled encouragement of my supervisors, Dr. Faisal Khan, Dr. Paul Amyotte, and Dr. Salim Ahmed. I am deeply grateful to Dr. Khan for giving me the opportunity to pursue my master's and Ph.D. degrees and work on diverse process safety areas. His untiring support and exceptional guidance made this thesis possible, and his research ideas, prompt responses, and valuable feedback helped me to complete this journey. I am equally indebted to Dr. Amyotte for providing detailed feedback on my work that helped me improve the readability of the manuscripts. His able guidance, coupled with words of motivation, is the reason I upheld academic integrity with the highest research standards. I am also grateful to Dr. Ahmed for his critical remarks and valuable suggestions to improve the quality of the work. I would like to thank Dr. Khan for giving me the opportunity to work at the Mary Kay O'Connor Process Safety Centre (MKOPSC) at Texas A&M University.

I would like to acknowledge the financial support provided by Genome Canada and its supporting partners through the Large Scale Applied Research Project, the Canada Research Chair (CRC) Tier I Program in Offshore Safety and Risk Engineering, and the MKOPSC at Texas A&M University. This acknowledgement would be incomplete without thanking Dr. Guozheng Song, Dr. Taleb-Berrouane, Dr. S. Zohra Halim, and Dr. Tanjin Amin for providing me with insights when I needed them the most. I would also like to thank all members of C-RISE and MKOPSC who supported and encouraged me during my journey.

I would like to thank my friends for their unparalleled support and inspiration throughout my journey. Last but not least, my profound gratitude goes to my family members, especially my

father, mother, and siblings, for their unconditional love and affection. I dedicate this thesis to my family for all their countless efforts in raising me.

Table of Contents

Abstract ..... *i*

Acknowledgement ..... *iii*

List of Figures ..... *xi*

List of Tables ..... *xiii*

List of Abbreviations ..... *xv*

1 Introduction *1*

1.1 Background and Motivation **1**

1.2 Objectives **4**

1.3 Outline **5**

1.4 Co-authorship Statement **8**

1.5 References **10**

2 Literature Review *12*

2.1 What is Microbiologically influenced corrosion (MIC)? **12**

2.2 MIC risk-based models **13**

2.3 NLP risk-based models **15**

2.4 Identified Knowledge Gaps **18**

2.5 References **20**

3 Data-Driven Operational Failure Likelihood Model for Microbiologically Influenced Corrosion *27*

Preface **27**

Abstract **27**

3.1 Introduction **28**

3.2 The Concept of Bayesian Learning **37**

3.2.1 Bayesian Network and its Structural Learning 38

3.2.2 Bayesian Parameter Learning 42

3.3 The LBN Model **43**

3.3.1 System Identification 44

3.3.2 System based Operational and Microbiological Data Collection 44

3.3.3 Data Preparation 46

3.3.4 Bayesian Learning 46

3.4 Application of LBN Model **47**

3.4.1 System Identification 48

3.4.2 System based Operational Data and Microbiological Data Collection 48

3.4.3 Data Preparation 49

3.4.4 Bayesian Learning 50

3.4.5 Application of the LBN Model with Missing Values 55

3.4.6 LBN Model Stability 60

3.4.7 Testing of Model on Data Set – Clean and Corrupt Data 61

3.5 Validation of the LBN Model **62**

3.6 Conclusions **65**

3.7 Acknowledgements **65**

3.8 References **65**

4 Textual Data Transformations Using Natural Language Processing for Risk Assessment *75*

Preface **75**

Abstract **75**

4.1 Introduction **76**

4.2 Proposed methodology **81**

4.2.1 Preprocessing of Corpus 84

4.2.2 Text Processing using Machine Learning 85

4.2.3 Transform Qualitative Features into Numerical Reasoning 86

4.2.4 Risk Evaluation 88

4.3 Application of Methodology **89**

4.3.1 Data Preparation for Custom NER Model 89

4.3.2 Automated Feature Extraction and Causation Construction 90

4.3.3 Transforming Qualitative Features to Quantitative Reasoning 98

4.4 Results and Discussion **108**

4.5 Verification of NER Model **114**

4.5.1 Purpose of Verification 114

4.5.2 Verification Results and Discussion 114

4.6 Conclusion **120**

4.7 Acknowledgements **121**

4.8 References **122**

5 A methodical approach for knowledge-based fire and explosion accident likelihood analysis *133*

Preface **133**

Abstract **133**

5.1 Introduction **134**

5.2 Methodology to Develop Knowledge-based Accident Causation Model **140**

5.2.1 Application of Natural Language Processing (NLP) 141

5.2.2 Interpretative Structural Model (ISM) 144

5.2.3 Quantitative reasoning using Bayesian Network (BN) 147

5.3 Application to CSB database (oil and refining - downstream) **149**

5.3.1 Development of NER model 150

5.3.2 Establishing hierarchy and interrelationships among factors 156

5.3.3 Generalized causation likelihood model 160

5.4 Results and Discussion **162**

5.4.1 Model testing and verification 162

5.4.2 Sensitivity Analysis 167

5.5 Conclusions **170**

5.6 Acknowledgements **171**

5.7 References **172**

6 Multi-source heterogeneous data integration for incident likelihood analysis in the processing systems ***194***

Preface **194**

Abstract **194**

6.1 Introduction **195**

6.2 Research Methodology 200

6.2.1 Employing Natural Language Processing (NLP) 200

6.2.2 Numerical data 203

6.2.3 Interpretive Structure Modelling (ISM) 204

6.2.4 Quantitative reasoning 207

6.2.5 Generalized Hybrid Causation Model 212

6.2.6 Updated Hybrid Causation Model 212

6.3 Application to CSB Database 212

6.3.1 Heterogeneous Data Sources 212

6.3.2 Establishing Interrelationship among Textual and Numerical Data 216

6.3.3 Developing Hybrid Causation Model 220

6.4 Results and discussion 225

6.4.1 Scenario-based verification 226

6.4.2 Sensitivity Analysis 230

6.5 Conclusions 232

6.6 Acknowledgments 233

6.7 References 234

7 Summary, Conclusions and Recommendations 249

7.1 Summary 249

7.2 Conclusions 250

7.2.1 Development of a learning-based likelihood model 250

7.2.2 Risk estimation and evaluation from textual data 251

7.2.3 Generalized causation likelihood analysis 251

7.2.4 Multi-source data integration for generalized causation analysis 252

7.3 Recommendations 252

7.3.1 Data requirements 253

7.3.2 Automated causation extraction 254

7.3.3 Uncertainty handling 254

List of Figures

Figure 1-1 Transformation of data into actions 3

Figure 1-2: Overview of research 5

Figure 1-3: Summary of the work presented in the thesis 8

Figure 3-1 The evolution of MIC risk assessment/modelling 31

Figure 3-2 Overview of the study conducted 37

Figure 3-3 Two possible structures of three node BN denoted as (a) and (b) 41

Figure 3-4 The proposed data-driven MIC model 45

Figure 3-5 A flow diagram of FPSO platform topside view adapted from Nicoletti (Nicoletti, 2020) 48

Figure 3-6 BN network learned from data set for SC 1013,1032 and 1037 locations 50

Figure 3-7 BN network learned for SC 1035 location 51

Figure 3-8 BN model for likelihood of pitting rate 52

Figure 3-9 Missing data pattern adapted from Imtiaz et al. (Imtiaz & Shah, 2008) 57

Figure 3-10 Bayesian learning with respect to missing values in data set 58

Figure 3-11 LBN progress with respect to data points 61

Figure 4-1 The proposed methodology for risk assessment from textual data 83

Figure 4-2 A cause-effect relationship from identified entities of case 1 95

Figure 4-3 A cause-effect relationship from identified entities of case 2 96

Figure 4-4 A cause-effect relationship from identified entities of case 3 96

Figure 4-5 A cause-effect relationship from identified entities of case 4 97

Figure 4-6 A cause-effect relationship from identified entities of case 5 98

Figure 4-7 Conversion of a linguistic variable into the likelihood of an event 100

Figure 4-8 Mapping algorithm from NER to BN 103

Figure 4-9 BN structure for case 1 105

Figure 4-10 BN structure for case 2 106
Figure 4-11 BN structure for case 3 106
Figure 4-12 BN structure for case 4 107
Figure 4-13 BN structure for case 5 108
Figure 4-14 Highlighted entities from constructed induced causation by (G. Liu et al., 2021) 117
Figure 5-1 A three-step systematically integrated approach to develop a generalized causation likelihood model 140
Figure 5-2 The proposed methodology for learning lessons from past experiences and predicting adverse events 142
Figure 5-3 Mapping algorithm from ISM into BN 148
Figure 5-4 Highlighted entities from NER model 151
Figure 5-5 Digraph developed from the ISM method 160
Figure 5-6 BN mapped from the ISM digraph 161
Figure 5-7 Tornado chart developed for sensitivity analysis of each factor 167
Figure 6-1 The methodology of creating a hybrid causation model from multi-source heterogeneous data 202
Figure 6-2 Steps of ISM process 204
Figure 6-3 Estimation scale of a linguistic variable into fuzzy likelihood 210
Figure 6-4 Co-occurrence network of release incidents from the CSB database of LOC incidents 214
Figure 6-5 Developed digraph from heterogeneous data sources 220
Figure 6-6 Simulated sensor data for temperature 224
Figure 6-7 Mapped BN from ISM digraph 225
Figure 6-8 Tornado chart to analyze the sensitivity of fuzzy and monitored nodes 231

List of Tables

Table 1-1 An overview of technical chapters 6

Table 3-1 Summary of common machine learning models used in risk assessment 35

Table 3-2 Major differences in Bayesian learning methods 38

Table 3-3 A data set to test the likelihood of BN structure's 41

Table 3-4 MIC likelihood of FPSO platform equipment 51

Table 3-5 Expected risk of pitting in FPSO platform 53

Table 3-6 Expected risk of pitting in FPSO platform, RA (prior probability), RB (microorganisms activity detected) and RC (corrective measure applied) 54

Table 3-7 Expected risk of pitting with respect to missing values in data set 59

Table 3-8 Confusion matrix with clean testing data 62

Table 3-9 Confusion matrix with 80% clean and 20% corrupt data 62

Table 3-10 Scenarios 1-6 evidence 62

Table 3-11 Comparison of MIC scenario-based likelihood results (rounded percentage) compared to Kannan et al.(Kannan et al., 2020) 64

Table 4-1 Selected description of cases narratives taken from the PHMSA database available in the public domain (Pipeline and Hazardous Materials Safety Administration, 2022) with NER model output 91

Table 4-2 Basic event probability for root nodes 103

Table 4-3 Events along with their symbols 104

Table 4-4 Likelihood of failure and consequences for cases shown in Table 1 108

Table 4-5 Categorization of likelihood and severity of consequences 109

Table 4-6 Proposed risk assessment matrix 110

Table 4-7 Risk level of PHMSA incidents cases 111

Table 4-8 The incidents narrative reported in (G. Liu et al., 2021) with NER model result 114

Table 5-1 Relevant features from each incident using NER model.....151
Table 5-2 List of factors from NER output and their probabilities.....157
Table 5-3 Model testing results in estimating fire and explosion likelihood.....163
Table 5-4 Accidents related to oil and gas (downstream) from ("Learning Lessons from Major Incidents," 2022).....164
Table 5-5 Model verification through unseen incident evidence.....166
Table 5-6 Structural self-interaction matrix (SSIM) developed by performing pair-wise comparison.....180
Table 5-7 Final reachability matrix (FRM).....182
Table 5-8 Partitioning of FRM.....186
Table 5-9 Conical matrix.....189
Table 6-1 Linguistic variables and associated fuzzy numbers to describe fuzzy event, adopted from (Chen Shu-Jen and Hwang, 1992; Zarei et al., 2019).....208
Table 6-2 Identified factors for ISM process.....217
Table 6-3 Fuzzy probability estimation.....222
Table 6-4 Scenario-based hard and soft evidence for verification.....227
Table 6-5 Structural self-interaction matrix (SSIM) Created using pair-wise comparison of each factor.....241
Table 6-6 Final reachability matrix (FRM).....242
Table 6-7 Level Partitioning.....244
Table 6-8 Conical matrix.....245

List of Abbreviations

Symbol	Definition
ABT	Adaptive Bow-Tie
AI	Artificial intelligence
ALARP	As low as reasonably practicable
ANN	Artificial neural network
APB	Acid-producing bacteria
BERT	Bidirectional Encoder Representations from Transformers
BN	Bayesian network
CCOHS	Canadian Centre for Occupational Health and Safety
CF	Causal/contributing factor
CMIC	Chemical microbiologically influenced corrosion
CNN	Convolutional neural network
CPI	Chemical processing industries
CPT	Conditional probability table
CSB	Chemical Safety and Hazard Investigation Board
EM	Expectation-Maximization
EMIC	Electrical microbiologically influenced corrosion
EPA	Environmental protection agency
ER	Emergency response
F&E	Fire & explosion
FPr	Fuzzy probability
FPs	Fuzzy possibility
FPSO	Floating, Production, Storage and Offloading

FRM	Final reachability matrix
HS	Hydrogen sulfide
IoT	Internet of things
IOB	Iron-oxidizing bacteria
IRB	Iron reducing bacteria
ISM	Interpretive Structural Model
LBN	Learning-based Bayesian network
LOC	Loss of containment
LPG	Liquified petroleum gas
MIC	Microbiologically influenced corrosion
MMM	Molecular Microbiological Methods
MoC	Management of Change
MPN	Most Probable Number
NER	Named entity recognition
NLP	Natural language processing
NRM	Nitrate-reducing microorganisms
OOBN	Object-Oriented Bayesian Network
PHMSA	Pipeline and Hazardous Material Safety Administration
QRA	Quantitative Risk Assessment
RM	reachability matrix
SCC	Stress Corrosion Cracking
SRB	sulphate-reducing bacteria
SSIM	Structural self-interaction matrix
SVM	support vector machine
TFZ	Triangular or trapezoidal fuzzy numbers

TRB	Thiosulfate-reducing bacteria

1 Introduction

1.1 Background and Motivation

The chemical processing industries (CPI) frequently witness accidents worldwide. In the last two years, 162 accidents have been reported in the U.S.A. alone (U.S. Chemical Safety and Hazard Investigation Board, 2022). These accidents include chemical manufacturing, distribution, combustible dust explosion, vapor cloud explosion, oil & refining, loss of containment (LOC), and fire & explosion (F&E). Accidents in the CPI are due to the hazardous nature of the materials involved. A near mishap in CPI can escalate into catastrophe, as seen in the past century (Khan & Abbasi, 1999), and continue to occur (Amyotte et al., 2016). Thus, accidents show the need for safety technologies to continually evolve with advancement in process operations. Chemical accidents are not only limited to chemical process industries but are also seen in other industries with hazardous chemicals. For example, a chemical explosion in an electronics facility took place in Hapur, India. The incident caused 10 fatalities and 22 severe injuries. According to the initial report, regulatory oversights and open disregard for safety norms were the reasons behind the incident (Reuters, 2022).

The occurrence of process accidents clearly shows that lessons are not learned from past accidents. The most important question remains: why do adverse accidents keep happening? The accidents happen due to a lack of database knowledge implementation, insufficient procedure and training, and process digitalization adoption in process operations (Amyotte et al., 2016). As Trevor Kletz observed, *"Accidents are not due to lack of knowledge, but failure to use the knowledge we have"*(IChemE Safety and Loss Prevention, 2022).

In the 21st century, industries are taking notable technological advancements, incorporating devices with Internet of Things capabilities, cloud computing, artificial intelligence (AI) and big data analytics. This smart technological advancement leads to the current era of Industry

4.0, characterized by automation and digitalization (Reis & Kenett, 2018). Industry 4.0 refers to the industrial revolution concerned with integrating o technological advancement, automation, and data exchange in various sectors. However, introducing Industry 4.0 brings new challenges that pertain to a leadership style that can deal with the challenges of Industry 4.0. Leadership is termed leadership 4.0. The advantage of Industry 4.0 requires leaders to understand and navigate the rapidly changing environment. Leadership 4.0 also requires leaders to strive for innovation and growth of an organization by adopting changes to meet with Industry 4.0 (Behie et al., 2023). Safety 4.0 requires integrating AI with data-driven approaches to proactively identify precursors before accidents happen. In the era of safety 4.0, safety science is going through a paradigm change in the age of big data (large and complex datasets that cannot be processed or managed using conventional tools), AI, and industry 4.0 called computational safety science (Wang, 2021). Safety 4.0 brings evolution to process safety due to AI and data-driven approaches (Qian et al., 2023). New approaches are needed to cope with the changes by re-engineering safety and handling evolving technological risks (Pasman & Fabiano, 2021).

Industry stakeholders acknowledge the important role of data. Virginia Rometty (CEO of IBM) said, "*What steam was to the 18th century, electricity to the 19th and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource*." (Reis & Kenett, 2018). The data are available in numerical and textual forms but lacks models that can leverage the available resources. Novel data-driven approaches with the implementation of AI are needed to learn from data in different forms. Data can be available in numerical or textual forms. The former is encountered by process data from sensors for monitoring purposes. The latter is commonly found in accident investigation reports where operators can make observations in free text as naturally spoken text. Process data and knowledge should drive risk assessment approaches in the era of safety 4.0 to analyze technological risks. The fundamental

challenge is to develop methods that can learn from data in various forms for robust risk assessment models. Figure 1-1 illustrates data transformation into meaningful information that serves as knowledge over time. This knowledge governs decision-making in the era of safety 4.0. There is a need for models capable of learning from process and textual data to leverage data in different forms and assess risks.

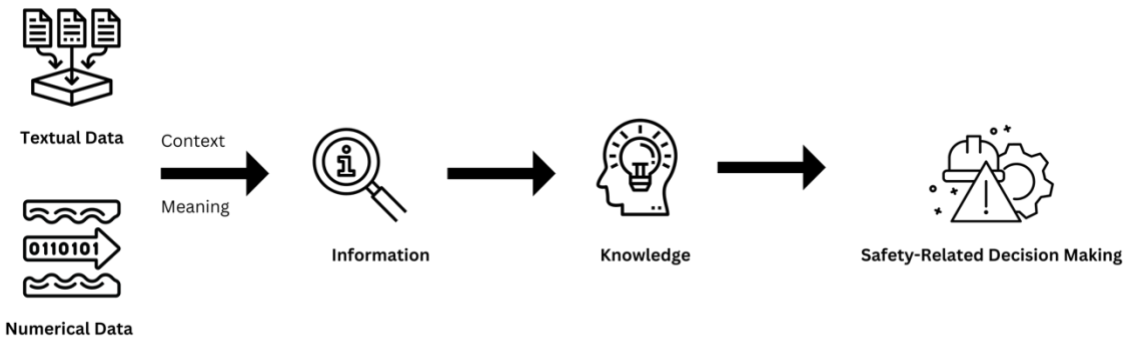


Figure 1-1 Transformation of data into actions

Continuous monitoring and preventive actions are key to avoiding abnormal situations (Khan et al., 2016). Implementing AI with data-driven approaches provides information for continuous monitoring and assisting in taking preventive actions based on precursors. Structured and unstructured data should govern predictive analysis for safety 4.0. The data-driven approaches to assess risk based on different data types are as follows:

Corrosion is a ubiquitous concern for CPI, especially microbiologically influenced corrosion (MIC) due to the complex behavior of microorganisms. Corrosion results in the loss of containment of hazardous materials that can easily develop into a catastrophe. Learning from MIC data are key to monitoring MIC threats and developing mitigation strategies. The advantage of learning is to capture causal factors interactions among each other that are missed in high-level heuristic observations (Kannan et al., 2020). Therefore, there is a need for a learning-based MIC model that can directly use operational and laboratory data to drive meaningful information for MIC (Skovhus et al., 2017).

Unstructured data in the form of textual data are another important resource for learning. The accident databases are available resources to gain insights into what went wrong. The database must be used to extract specific information for learning lessons (Mannan & Waldram, 2014). Based on database insights, a new dimension of risk analysis can be developed to assess risk based on past events.

The popularity of natural language processing (NLP) tools shows technological advancement. Therefore, real-time risk monitoring must be considered to develop a knowledge base/intelligent system (Khan et al., 2016). Hence, the motivation of this research is to bridge technological gaps between existing methods and requirements of safety 4.0.

The focus is assessing MIC likelihood and developing models for LOC and F&E to prevent accidents. The work presented here aims to contribute to risk assessment using field data and accident investigation reports. The research activities aim to introduce risk modeling approaches to predict adverse events based on past experiences. These approaches are applied on databases like Pipeline and Hazardous Material Safety Administration (PHMSA), and Chemical Safety and Hazard Investigation Board (CSB).

1.2 Objectives

The thesis aims to introduce data-driven tools to assess a current situation based on available knowledge to predict adverse events in process industries. The knowledge is obtained from contemporary or historical data. Therefore, a tool must be able to process textual and numerical data. The scope is restricted to risk estimation and evaluation for decision-making purposes. It does not incorporate the procedure of reducing risk through risk management strategies. With the overall objective and scope in mind, the thesis aims to answer the following questions.

i. How can field and laboratory data be used for Bayesian learning to assess risk?
ii. How can textual data be used to assess objective risk?
iii. How can database knowledge be used systematically to assess risk?

iv. How can heterogeneous data (i.e., numerical and textual data) be fused to assess accident likelihood?

The research activities performed aim to answer these research questions. The objectives of the thesis, as shown in Figure 1-2, are to:

1. Develop a data-driven methodology for assessing MIC risk from field data. Relying on field data can encounter incomplete datasets. Hence, the methodology should be capable of handling complete and incomplete datasets.
2. Transform textual data using NLP to evaluate objective risk
3. Develop a generalized causation model to predict adverse events based on past experiences.
4. Integrate textual and numerical data for robust accident likelihood analysis.

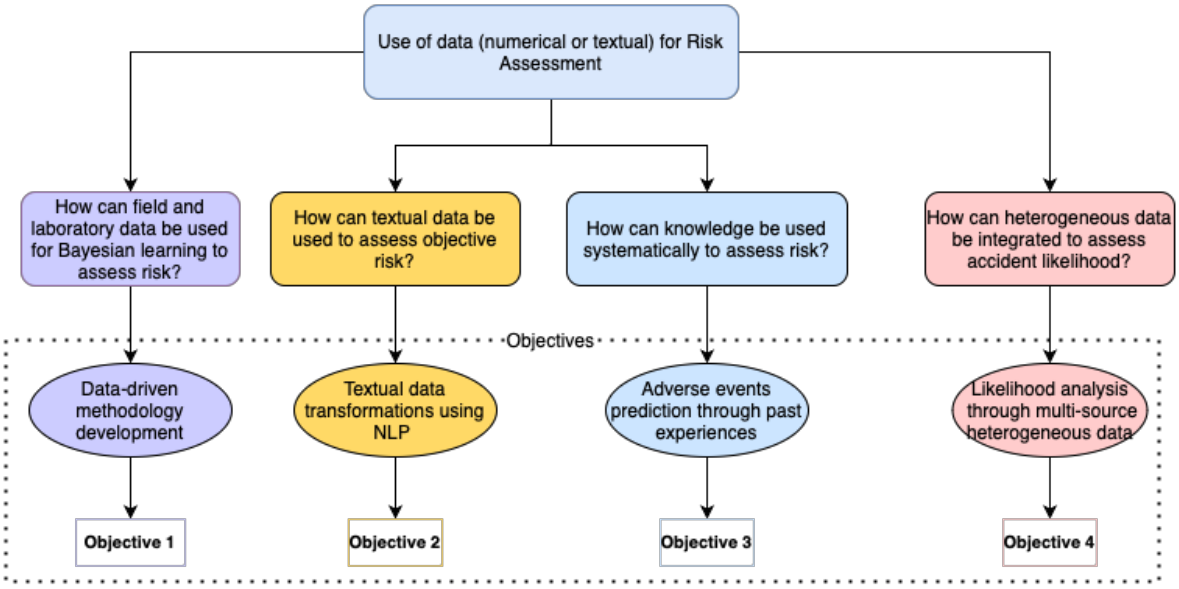


Figure 1-2: Overview of research

1.3 Outline

The thesis presented here is organized into seven chapters and is a manuscript-type thesis. The first introductory chapter briefly overviews the research activities, objectives, motivations and contributions. The second chapter briefly overviews the available literature and identifies

knowledge gaps. The last chapter states the conclusion derived from the thesis. Chapters three to five are based on peer-reviewed journal papers published and chapter six is a paper submitted to Computers & Chemical Engineering for peer-review publication. A short description of the technical chapters is presented in Table 1-1.

Table 1-1 An overview of technical chapters

Chapters	Research Objective	Tool(s) used	Title	Case Study
3	Data-Driven MIC Likelihood	Bayesian score-based method EM algorithm Bayesian network (BN)	Data-Driven Operational Failure Likelihood Model for Microbiologically Influenced Corrosion	Industrial partner data: FPSO facility located in North America
4	Objective Risk Assessment from Textual Data	Named entity recognition Fuzzy logic BN Risk matrix	Textual Data Transformations using Natural Language Processing for Risk Assessment	MIC related incidents from PHMSA database
5	Knowledge-Based Accident Causation Model	Named entity recognition Interpretive structural model	A methodical approach for knowledge-based fire and explosion	CSB database (oil and refining

		BN	accident likelihood analysis	- downstream)
6	Data Fusion of Textual and Numerical Data	Co-occurrence network Interpretive structural model BN	Multi-source heterogeneous data fusion for likelihood analysis	CSB database (loss of containment)

Chapter three proposes a framework of data-driven Bayesian learning that can model structure and parameters from data. The K2 algorithm is used to learn a topology of BN, whereas the EM algorithm is used for parameter estimation to assess MIC risk. The chapter has been published in Process Safety and Environmental Protection.

Chapter four proposes a novel framework to assess objective risk assessment from textual data. The PHMSA database is used to develop a risk model by utilizing NLP. Named entity recognition (NER) method is used for feature extraction that serves as a basis for the risk model. The chapter has been published in Risk Analysis.

Chapter five proposes a unique approach to developing a generalized causation model from past experiences. Domain expertise with lessons learned from accidents is used to assess similarities among different accidents. The chapter has been published in Process Safety and Environmental Protection.

Chapter six introduces a framework for multi-source heterogeneous data fusion. Historical and contemporary data are used to develop an accident likelihood model. Hierarchical BN is developed to model complex interactions among textual and numerical data. The chapter has been submitted to Computers and Chemical Engineering.

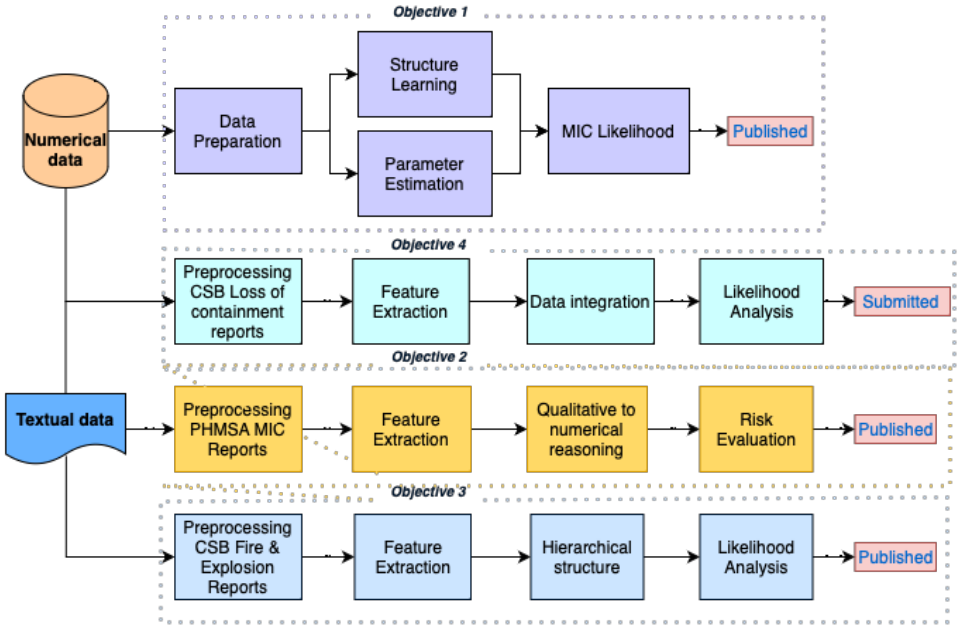


Figure 1-3: Summary of the work presented in the thesis

Chapter seven draws a summary of the work and states the conclusion drawn from the research studies. Future work recommendations are given at the end of this thesis.

Figure 1-3 shows an overview of the work presented in this thesis and their respective status. Further detail on each chapter is presented in designated chapters.

1.4 Co-authorship Statement

I am the sole author of this manuscript-type thesis and the primary author of technical chapters that are either published or submitted for peer-review publication. With the help of co-authors, Drs. Faisal Khan, Paul Amyotte and Salim Ahmed, I developed the first draft of the manuscripts presented in the chapters. I carried out the analysis, model development, testing, and verification. Dr. Khan helped me with conceptualization, model development and reviewing the work. He also assisted in the revision of the manuscripts. Drs. Amyotte and Ahmed helped review, revise and improve the manuscript's readability. They also assisted me in incorporating peer-reviewed feedback and checking the models testing and verification

exercise. Drs. Taleb-Berrouane and S. Zohra Halim also assisted me as co-authors of selected manuscripts in reviewing, revising and incorporating peer-reviewed feedback.

1.5 References

1. Amyotte, P. R., Berger, S., Edwards, D. W., Gupta, J. P., Hendershot, D. C., Khan, F. I., Mannan, M. S., & Willey, R. J. (2016). Why major accidents are still occurring. *Current Opinion in Chemical Engineering*, *14*, 1–8.
2. Behie, S. W., Pasman, H. J., Khan, F. I., Shell, K., Alarfaj, A., El-Kady, A. H., & Hernandez, M. (2023). Leadership 4.0: The changing landscape of industry management in the smart digital era. *Process Safety and Environmental Protection*, *172*, 317–328.
3. IChemE *Safety and Loss Prevention*. (2022). https://www.icheme.org/membership/communities/special-interest-groups/safety-and-loss-prevention/resources/lessons-learned-database/
4. Kannan, P., Kotu, S. P., Pasman, H., Vaddiraju, S., Jayaraman, A., & Mannan, M. S. (2020). A systems-based approach for modeling of microbiologically influenced corrosion implemented using static and dynamic Bayesian networks. *Journal of Loss Prevention in the Process Industries*.
5. Khan, F. I., & Abbasi, S. A. (1999). Major accidents in process industries and an analysis of causes and consequences. *Journal of Loss Prevention in the Process Industries*, *12*(5), 361–378.
6. Khan, F., Thodi, P., Imtiaz, S., & Abbassi, R. (2016). Real-time monitoring and management of offshore process system integrity. *Current Opinion in Chemical Engineering*, *14*, 61–71.
7. Mannan, M. S., & Waldram, S. P. (2014). Learning lessons from incidents: A paradigm shift is overdue. *Process Safety and Environmental Protection*, *92*(6), 760–765.
8. Pasman, H. J., & Fabiano, B. (2021). The Delft 1974 and 2019 European Loss Prevention Symposia: Highlights and an impression of process safety evolutionary

changes from the 1st to the 16th LPS. *Process Safety and Environmental Protection*, *147*, 80–91.

9. Qian, Y., Vaddiraju, S., & Khan, F. (2023). Safety education 4.0 – A critical review and a response to the process industry 4.0 need in chemical engineering curriculum. *Safety Science*, *161*, 106069.
10. Reis, M. S., & Kenett, R. (2018). Assessing the value of information of data-centric activities in the chemical processing industry 4.0. *AIChE Journal*, *64*(11), 3868–3881.
11. Reuters. (2022). At Least 10 People Killed in India Factory Explosion. https://www.reuters.com/world/india/least-six-killed-india-chemical-factory-explosion-2022-06-04/
12. Skovhus, T. L., Enning, D., & Lee, J. S. (2017). Microbiologically influenced corrosion in the upstream oil and gas industry. In *Microbiologically Influenced Corrosion in the Upstream Oil and Gas Industry* (Vol. 1).
13. U.S. Chemical Safety and Hazard Investigation Board. (2022). Incident Reporting Rule Submission Information and Data. https://www.csb.gov/news/incident-report-rule-form-/
14. Wang, B. (2021). Safety intelligence as an essential perspective for safety management in the era of Safety 4.0: From a theoretical to a practical framework. *Process Safety and Environmental Protection*, *148*, 189–199.https://doi.org/https://doi.org/10.1016/j.psep.2020.10.008

2 Literature Review

The chapter presents a literature review on specific aspects to identify the knowledge gaps that can be addressed in the thesis. The detailed literature reviews are included in the subsequent chapters of the thesis. The literature review in the present chapter covers risk assessment methods, focusing on

i) MIC risk-based models

ii) NLP risk-based models

2.1 What is Microbiologically influenced corrosion (MIC)?

Corrosion is a challenging and major concern for industries such as CPI and oil and gas. Due to corrosion, industries have suffered significant losses, posing an economic challenge. There are many forms of corrosion in which MIC is considered a complex phenomenon that includes microorganisms responsible for creating a corrosive environment due to their presence or activity at the metal surface (Little & Lee, 2014). Microorganisms tend to alter electrochemical conditions at the metal surface (Salgar-Chaparro et al., 2020; Videla & Herrera, 2005). This type of corrosion involving microorganisms is found in pipelines and storage vessels. Loss of containment (LOC) due to MIC releases hydrocarbons, leading to fatalities, property damage, business interruption, reputation loss and environmental damage (Kannan et al., 2020). MIC poses a risk to process operations that leads to catastrophic failures. There have been many process accidents attributed to MIC, such as propane tank failure leading to an explosion in Umm Said NGL Plant in Qatar (Salgar-Chaparro et al., 2020), a crude oil spill on Alaska's North Slope, gas leakage and an explosion in New Mexico(A. Abdullah et al., 2014; Salgar-Chaparro et al., 2020; Sooknah et al., 2008) and the Abkatun standing platform fire in the Gulf of Mexico which killed four workers and injured 16 others (Kannan et al., 2018). Additional MIC-attributed cases were reported by Skovhus et al. (Skovhus et al., 2017), such as failure in high-pressure production at the Gas Oil separation plant due to the growth of microorganisms,

corrosion of the tube in a heat exchanger caused by acid-producing bacteria resulting in leakage, failure of fire hydrants due to the presence of sulfate-reducing bacteria and archaea and failure of a diesel pipeline due to the presence of pitting because of a corrosive deposit.

MIC is a complex and diverse process; it includes various species, such as sulfate reducers, acid-producing bacteria and iron reducers, that develop biofilm attached to the metal surface (Geissler et al., 2014; Kannan et al., 2020). The biofilm serves multiple purposes and is attributed to MIC; it acts as a diffusion barrier, preventing oxygen and anion diffusion to cathodic and anodic sites. Detachment of biofilm results in removing a protective film, and the non-uniform nature of biofilm results in differential aeration cells that causes the potential difference, resulting in corrosion current. It also alters conditions of oxidation/reduction at the interface between metal and hydrocarbon (Videla & Herrera, 2005).

2.2 MIC risk-based models

The modeling of MIC is a challenging problem due to the complex behavior of microorganisms responsible for accelerating the corrosion process. Also, the interaction of microorganisms with biotic and abiotic factors is complex. The interaction either leads to an intensification or diminution of MIC over a period of time. The first attempt to model MIC is traced back to 2002, except biological parameters, parameters including operation parameters, water presence and wetting, are considered (Pots et al., 2002). Although the biological parameters were incorporated later to monitor and mitigate MIC threat in pipelines (Maxwell; Campbell et al., 2006). The risk associated with bio-fouling was also evaluated on a section of the pipeline using surface conditions, biological phenomena and hydraulics using the neural network method (Urquidi-Macdonald et al., 2014). The risk matrix method is also used to assess MIC risk (Kaduková et al., 2014) before moving to a semi-quantitative approach using prediction and monitoring factors (Wang & Jain, 2016).

Numerous attempts were made to model the MIC process using BN (Dawuda et al., 2021; Taleb-Berrouane et al., 2019; Taleb-Berrouane et al., 2018; Taleb-Berrrouane et al., 2019). BN is particularly suitable for risk analysis and capturing accident scenarios from cause to consequence (Deyab et al., 2018; Kabir et al., 2019; Taleb-Berrouane et al., 2017; Yang et al., 2020). In contrast, other modeling techniques exist for accident scenarios, such as Petri nets (Kamil et al., 2019; Taleb-Berrouane et al., 2019; Taleb-Berrouane & Khan, 2019) or Markov chains (Taleb-Berrouane et al., 2016). However, adding a new factor can lead to a different structure in the latter, whereas it will remain the same in the former.

In addition, BN offers the flexibility of incorporating evidence into the network with the help of the Bayes theorem (Taleb-Berrouane et al., 2020; Taleb-berrouane, Imtiaz, et al., 2018). (Ayello et al., 2014) proposed a model for internal and external corrosion and incorporated a limited MIC mechanism. Koch et al. developed a BN-based model to assess MIC based on sulfate-reducing microorganisms where other factors responsible for MIC were not considered (Koch et al., 2015). Another BN-based model considers the MIC mechanism a sub-mechanism in internal corrosion modeling (Shabarchin & Tesfamariam, 2016). MIC was quantified by identifying and quantifying internal corrosion causal factors dependencies using BN (Liu et al., 2018). Another BN-based model is developed by considering operational parameters, operating history and indication factors to assess MIC threat (Taleb-Berrouane et al., 2018). However, this comprehensive study is a static model and lacks dynamic behavior. The subsequent research considers the latter (Kannan et al., 2020). Kannan et al., 2020 developed the BN model by considering failure history data, operational data and other parameters in a 60-node structure. The model considers a limited number of dynamic nodes and does not leverage MIC data due to unavailability in the public domain. BN model structure was based on expert opinion, thus introducing uncertainty into the assessment. Another model was

developed to capture changes in the database and adapts those changes in the Bowtie model (Taleb-Berrouane et al., 2021).

The literature reviews show that limited research has been carried out on the failure aspect of the MIC phenomenon compared to MIC monitoring, inhibition and development of biofilm (Hashemi et al., 2018; Taleb-Berrouane et al., 2020). Based on the approaches discussed in section 2.2 to model MIC, the primary concern is using expert opinion due to the lack of field data in determining MIC risk model structure and parameters. The causation factors interaction in modeling MIC is important but is often neglected due to the unavailability of modeling approach. MIC failure risk model is needed that can capture causal factor interaction and learn from data to develop MIC risk model (Skovhus et al., 2017).

2.3 NLP risk-based models

NLP is a field of computer science, artificial intelligence (AI) and linguistics concerned with the interaction between humans and computers (Khatri, 2021). NLP comes from the processing of natural languages used to convey messages and thoughts to another person. For computers to understand the message embedded in natural language, the message needs to process by converting it into a numerical form that machines. In other words, NLP is a domain of extracting information from spoken language or written textual data (Clark et al., 2010).

NLP enables computers to process, interpret and extract meaningful information from a natural language with the help of algorithms, statistical models and computational linguistics. NLP has a wide category of applications, to name a few: text classification, sentimental analysis, machine translation and the most popular chatbots (ChatGPT and LaMDA). Although the application of NLP is wide and evolving, there has been limited research on leveraging NLP for developing risk models.

NLP applications have been widely found in ontology-based studies for risk assessment models. A scenario object model is developed using domain ontology for information

extraction from HAZOP analysis (Wu et al., 2013). An investigation was carried out to assess the performance of domain-based ontology with non-domain-based ontology and concluded that the former provides robust results to automate the extraction of safety regulatory information (Kwon et al., 2013). Developing an ontology is time-consuming and labor-intensive work. However, incorporating machine learning to develop ontology assists in developing ontology with less time.

A semi-quantitative ontology development was proposed that employed machine learning in developing ontology (Guo & Huang, 2016). Developing ontology requires domain expertise for defining the keywords. A pre-defined keywords list based on domain expertise was used to automate feature extraction to gain insights. An application was shown on construction safety reports to automate accident precursors and outcomes to gain insights (Tixier et al., 2016b). A hazard ontology was developed to transform preliminary hazards from natural language into hazard modeling language to specify hazards (Zhou et al., 2017).

NLP is used to text-mine data to analyze a bag of words from adjacent sentences to extract information from textual data for aviation incidents (Nakata, 2017). However, the order of the bag of words was not considered, which assists in developing causation. An ontology-based approach was used to analyze language components such as subject, predicate and object to improve communication in airport operations (Abdullah et al., 2019). An ontology-based framework was developed to automate knowledge extraction from abstracts using bidirectional encoder representations from transformers. The proposed method was advantageous due to gaining more insights than bibliometric analysis, which is restricted to finding relations between co-authors, publications and institutions.

The approaches show the application of an ontology-based approach for gaining insights from textual data. The ontology-based approaches can also be used to assess risk. A pathway is proposed to construct BN comprised of multi-entity to assess risk (Aziz et al., 2019). Later, an

ontology-based approach integrating active learning was developed to extract dependencies (Deshpande et al., 2020). Knowledge acquisition based on ontology was developed and applied to the chemical accident database (Single et al., 2020). Besides ontology-based approaches, another broad area of NLP application comprises machine learning-based methods. However, machine learning demands data in large quantities for better accuracy (Robinson et al., 2015). The random forest method extracts information about injuries reported due to construction (Tixier et al., 2016a). Support vector machine (SVM) transforms natural language features from the textual form into numerical data for classification purposes (Tanguy et al., 2016). Natural gas pipeline incident data was used for developing an integrated spatio-temporal approach to extracting correlations between causation factors and the severity of incidents (Li et al., 2021). Based on the outcome, causal factors related to human shows the highest severity in natural gas pipeline incidents. K-means clustering, and co-occurrence matrix are used to text-mine reports from the PHMSA database (Liu et al., 2021). The advantage is extracting contributory factors and potential causality from the accident reports.

Another text-mining approach was developed based on a semi-supervised method to label unstructured data and (Ahadh et al., 2021). The advantage is labeling data with less manual intervention to analyze reports. The application was shown on pipeline accidents to identify causes and aviation reports to determine the flight stage at the time of the accident (Ahadh et al., 2021). Recent work demonstrated how to employ NLP to analyze subject and action words from their co-occurrences for accident consequence prediction (Wang et al., 2023).

The literature review discussed in section 2.3 revealed that there needs to be a solution for using unstructured data as a data source to evaluate objective risk assessment. NER is a proven method to identify entities that can be used to identify underlying cause-effect scenarios from textual data. Furthermore, based on the literature, no systematic process is available for developing qualitative and quantitative reasoning for learning lessons from past experiences.

There is a need for a method to visualize the hierarchy among different factors in a complex system that assists in making decisions. Quantitative reasoning defines factors' interrelationships and estimates each accident's likelihood with potential pathways based on the given conditions. Developing a systematic approach offers an opportunity to establish a generalized causation model. The generalized causation model can be further developed into a hybrid model comprised of structured and unstructured data sources and serve as a tool for Safety 4.0. Safety 4.0 demands a data-driven approach integrating artificial intelligence techniques (i.e., NLP) to gain insights for better safety management. Collecting data from the database and real-time data from sensors provides a comprehensive view of accident patterns and potential hazards that would otherwise be difficult to detect.

2.4 Identified Knowledge Gaps

The knowledge gaps identified from the literature review are conducted in sections 2.1, 2.2 and 2.3 and are summarized as follows:

1. Modeling of MIC requires an approach to extract the information from available structured data to evaluate MIC risk.
2. Developing a BN-based model for MIC requires a data-driven approach to learning the topology and parameters from structured data.
3. Extracting causation factors interaction from the structured data must be derived from operational and laboratory data.
4. Developing an objective risk assessment methodology is important for analyzing existing database resources and evaluating risk.
5. Automating underlying cause-effect storylines from the domain-specific corpus is needed.
6. Developing a knowledge-based causation model from unstructured data requires a methodical approach.

7. Integrating multi-source heterogeneous data are needed from structured and unstructured data to develop robust likelihood analysis.

2.5 References

1. Abdullah, A., Yahaya, N., Norhazilan, M. N., & Rasol, R. M. (2014). Microbial corrosion of API 5L X-70 carbon steel by ATCC 7757 and consortium of sulfate-reducing bacteria. *Journal of Chemistry*.
2. Abdullah, D., Takahashi, H., & Lakhani, U. (2019). Domain Specific Ontology Enhancing Communication Accuracy in Airport Operation. *Proceedings - 2019 IEEE 14th International Symposium on Autonomous Decentralized Systems, ISADS 2019*.
3. 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*.
4. Ayello, F., Jain, S., Sridhar, N., & Koch, G. H. (2014). Quantitive assessment of corrosion probability - A Bayesian network approach. *Corrosion*.
5. Aziz, A., Ahmed, S., & Khan, F. I. (2019). An ontology-based methodology for hazard identification and causation analysis. *Process Safety and Environmental Protection*.
6. Clark, A., Fox, C., & Lappin, S. (2010). The Handbook of Computational Linguistics and Natural Language Processing. In *The Handbook of Computational Linguistics and Natural Language Processing*.
7. Dawuda, A.-W., Taleb-berrouane, M., & Khan, F. (2021). A probabilistic model to estimate microbiologically influenced corrosion rate. *Process Safety and Environmental Protection*.
8. Deshpande, G., Motger, Q., Palomares, C., Kamra, I., Biesialska, K., Franch, X., Ruhe, G., & Ho, J. (2020). Requirements Dependency Extraction by Integrating Active Learning with Ontology-Based Retrieval. *Proceedings of the IEEE International Conference on Requirements Engineering*.

9. Deyab, S. M., Taleb-berrouane, M., Khan, F., & Yang, M. (2018). Failure analysis of the offshore process component considering causation dependence. *Process Safety and Environmental Protection*, *1*(8), 220–232.
10. Geissler, B., De Paula, R., Keller-Schultz, C., Lilley, J., & Keasler, V. (2014). Data mining to prevent microbiologically influenced corrosion? *NACE - International Corrosion Conference Series*.
11. Guo, J., & Huang, J. C. (2016). Ontology learning and its application in software-intensive projects. *Proceedings - International Conference on Software Engineering*.
12. Hashemi, S., Bak, N., Khan, F., Hawboldt, K., Lefsrud, L., & Wolodko, J. (2018). Bibliometric Analysis of Microbiologically Influenced Corrosion (MIC) of Oil and Gas Engineering Systems. *Corrosion*, *74*(4), 468–486.
13. Johri Prashant and Khatri, S. K. and A.-T. A. T. and S. M. and S. S. and K. A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In O. and V. D. Abraham Ajith and Castillo (Ed.), *Proceedings of 3rd International Conference on Computing Informatics and Networks* (pp. 365–375). Springer Singapore.
14. Kabir, S., Taleb-berrouane, M., & Papadopoulos, Y. (2019). Dynamic Reliability Assessment of Flare Systems by Combining Fault Tree Analysis and Bayesian Networks. *Energy Sources Part A Recovery Utilization and Environmental Effects*, *September*.
15. Kaduková, J., Škvareková, E., Mikloš, V., & Marcinčáková, R. (2014). Assessment of microbially influenced corrosion risk in slovak pipeline transmission network. *Journal of Failure Analysis and Prevention*, *14*(2), 191–196.
16. Kamil, M. Z., Taleb-Berrouane, M., Khan, F., & Ahmed, S. (2019). Dynamic domino effect risk assessment using Petri-nets. *Process Safety and Environmental Protection*, *124*.
17. Kannan, P., Kotu, S. P., Pasman, H., Vaddiraju, S., Jayaraman, A., & Mannan, M. S. (2020). A systems-based approach for modeling of microbiologically influenced corrosion

implemented using static and dynamic Bayesian networks. *Journal of Loss Prevention in the Process Industries*.

18. Kannan, P., Su, S. S., Mannan, M. S., Castaneda, H., & Vaddiraju, S. (2018). A Review of Characterization and Quantification Tools for Microbiologically Influenced Corrosion in the Oil and Gas Industry: Current and Future Trends. *Industrial and Engineering Chemistry Research*, *57*(42), 13895–13922.
19. Koch, G., Ayello, F., Khare, V., Sridhar, N., & Moosavi, A. (2015). Corrosion threat assessment of crude oil flow lines using Bayesian network model. *Corrosion Engineering, Science and Technology*, *50*(3), 236–247.
20. Kwon, J. H., Kim, B., Lee, S. H., & Kim, H. (2013). Automated procedure for extracting safety regulatory information using natural language processing techniques and ontology. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*.
21. Li, X., Penmetsa, P., Liu, J., Hainen, A., & Nambisan, S. (2021). Severity of emergency natural gas distribution pipeline incidents: Application of an integrated spatio-temporal approach fused with text mining. *Journal of Loss Prevention in the Process Industries*.
22. Little, B. J., & Lee, J. S. (2014). Microbiologically influenced corrosion: an update. *International Materials Reviews*, *59*(7), 384–393.
23. 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.
24. Liu, G., Zhang, J., Ayello, F., & Stephens, P. (2018). The application of Bayesian network threat model for corrosion assessment of pipeline in design stage. *Proceedings of the Biennial International Pipeline Conference, IPC*.
25. Maxwell; Campbell, Maxwell, S., & Campbell, S. (2006). Monitoring the mitigation of MIC risk in pipelines. *NACE - International Corrosion Conference Series*, *244*, 1–10.

26. Nakata, T. (2017). Text-mining on incident reports to find knowledge on industrial safety. *Proceedings - Annual Reliability and Maintainability Symposium*.

27. Pots, B. F., John, R. C., Rippon, I. J., Thomas, M. J. J. S. J. S., Kapusta, S. D., Girgis, M. M., Whitham, T., Grigs, M. M., & Whitham, T. (2002). Improvements on de waard-milliams corrosion prediction and applications to corrosion management. *NACE - International Corrosion Conference Series*, *02235*, 19.

28. Robinson, S. D., Irwin, W. J., Kelly, T. K., & Wu, X. O. (2015). Application of machine learning to mapping primary causal factors in self reported safety narratives. *Safety Science*, *75*, 118–129.

29. Salgar-Chaparro, S. J., Darwin, A., Kaksonen, A. H., & Machuca, L. L. (2020). Carbon steel corrosion by bacteria from failed seal rings at an offshore facility. *Scientific Reports*.

30. Shabarchin, O., & Tesfamariam, S. (2016). Internal corrosion hazard assessment of oil & gas pipelines using Bayesian belief network model. *Journal of Loss Prevention in the Process Industries*, *40*, 479–495.

31. 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*.

32. Skovhus, T. L., Enning, D., & Lee, J. S. (2017). *Microbiologically influenced corrosion in the upstream oil and gas industry* (Vol. 1).

33. Sooknah, R., Papavinasam, S., & Revie, R. W. (2008). Validation Of A Predictive Model For Microbiologically Influenced Corrosion. In *CORROSION 2008* (p. 17). NACE International.

34. Taleb-berrouane, M., Imtiaz, S., & Khan, F. (2018). Internal Corrosion Monitoring in the Crude Oil Pipelines. *20th Annual Aldrich Conference*, *March*.

35. Taleb-Berrouane, M., & Khan, F. (2019). Dynamic resilience modelling of process systems. *Chemical Engineering Transactions*, *77*(1), 313–318.
36. Taleb-Berrouane, M., Khan, F., & Amyotte, P. (2020). Bayesian Stochastic Petri Nets (BSPN) - A new modelling tool for dynamic safety and reliability analysis. *Reliability Engineering and System Safety*, *193*.
37. Taleb-Berrouane, M., Khan, F., Eckert, R. B., & Skovhus, T. L. (2019). Predicting Sessile Microorganism Populations in Oil and Gas Gathering and Transmission Facilities-Preliminary Results. *7th International Symposium on Applied Microbiology and Molecular Biology in Oil Systems (ISMOS 7)*.
38. Taleb-Berrouane, M., Khan, F., & Hawboldt, K. (2021). Corrosion risk assessment using adaptive bow-tie (ABT) analysis. *Reliability Engineering & System Safety*, *214*(May), 107731.
39. Taleb-Berrouane, M., Khan, F., Hawboldt, K., Eckert, R., & Skovhus, T. L. (2018). Model for microbiologically influenced corrosion potential assessment for the oil and gas industry. *Corrosion Engineering, Science and Technology*, *53*(5), 378–392.
40. Taleb-Berrouane, M., Khan, F., & Kamil, M. Z. (2019). Dynamic RAMS analysis using advanced probabilistic approach. *Chemical Engineering Transactions*, *77*.
41. Talebberrouane, M., Khan, F., & Lounis., Z. (2016). Availability Analysis of Safety Critical Systems Using Advanced Fault Tree and Stochastic Petri Net Formalisms. *Journal of Loss Prevention in the Process Industries*, *44*, 193–203.
42. Taleb-Berrouane, M., Sterrahmane, A., Mehdaoui, D., & Lounis., Z. (2017). Emergency Response Plan Assessment Using Bayesian Belief Networks. *3rd Workshop and Symposium on Safety and Integrity Management of Operations in Harsh Environments (C-RISE3)*.

43. Taleb-berrrouane, M. (2019). *Dynamic Corrosion Risk Assessment in the Oil and Gas Production and Processing Facility* (Issue October). Memorial University of Newfoundland.
44. Taleb-berrrouane, M., & Khan, F. (2018). Development of MIC Risk Index for Oil and Gas Operations. *C-RISE & Geno-MIC Workshop & Symposium*.
45. 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*.
46. Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016a). Application of machine learning to construction injury prediction. *Automation in Construction*.
47. Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*.
48. Urquidi-Macdonald, M., Tewari, A., & Ayala H, L. F. (2014). A neuro-fuzzy knowledge-based model for the risk assessment of microbiologically influenced corrosion in crude oil pipelines. *Corrosion*.
49. Videla, H. A., & Herrera, L. K. (2005). Microbiologically influenced corrosion: looking to the future. *International Microbiology : The Official Journal of the Spanish Society for Microbiology*, *8*(3), 169–180.
50. Wang, F., Gu, W., Bai, Y., & Bian, J. (2023). A method for assisting the accident consequence prediction and cause investigation in petrochemical industries based on natural language processing technology. *Journal of Loss Prevention in the Process Industries*, *83*, 105028.
51. Wang, Y., & Jain, L. (2016). MIC assessment model for upstream production and transport facilities. *NACE - International Corrosion Conference Series*.

52. Wu, C. G., Xu, X., Zhang, B. K., & Na, Y. L. (2013). Domain ontology for scenario-based hazard evaluation. *Safety Science*.

53. Yang, R., Khan, F., Taleb-Berrouane, M., & Kong, D. (2020). A time-dependent probabilistic model for fire accident analysis. *Fire Safety Journal*, *111*(December 2018), 102891.

54. Zhou, J., Hanninen, K., & Lundqvist, K. (2017). A hazard modeling language for safety-critical systems based on the hazard ontology. *Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017*.

3 Data-Driven Operational Failure Likelihood Model for Microbiologically Influenced Corrosion

Preface

This chapter has been published in the ***Process Safety and Environmental Protection*** Journal. I am the primary author of this manuscript, along with co-authors Drs. Mohammed Taleb-Berrouane, Faisal Khan, and Paul Amyotte. I developed the framework for Bayesian learning and its application in developing the MIC model from operational and laboratory data. I prepared the first draft of the manuscript and revised it based on the co-authors' and peer review feedback. The co-author Dr. Mohammed Taleb-Berrouane provided fundamental assistance in model development, data collection, testing and revision based on peer review feedback. The co-author Dr. Faisal Khan proposed the conceptual framework and helped develop the framework, testing and revising the model. The co-author Dr. Paul Amyotte provided constructive feedback to improve the readability, review and revision based on peer review feedback and finalizing the manuscript.

Reference: Kamil, M. Z., Taleb-Berrouane, M., Khan, F., & Amyotte, P. (2021). Data-driven operational failure likelihood model for microbiologically influenced corrosion. *Process Safety and Environmental Protection*, *153*, 472-485.

Abstract

Corrosion is a threat to asset integrity, with engineering challenges and economic burdens. Since the last decade, microbiologically influenced corrosion (MIC) began to be recognized among corrosion professionals as a severe corrosion form. It is challenging to detect and predict MIC due to the complex behaviour of microorganisms. The current MIC risk assessment models define the dependencies of parameters with their synergic interactions. A data-driven approach is needed to utilize available operational and microbiological data and learn as the data changes. The model proposed in this study is used to strengthen the variables' correlation

and their features to assess MIC likelihood. It can integrate available field and laboratory data into a Learning-based Bayesian network (LBN) model. The model minimizes current research gap and has the advantage of adapting to changes in process operation. It is based on an advanced Bayesian learning algorithm, which develops topology of the Bayesian network (BN) from the input data and its parameters.

This chapter focuses on the development of the LBN model that utilizes available MIC data for likelihood estimation. The model is tested and validated using data reported in the public domain. The application of the model is demonstrated on the processing facility on a Floating, Production, Storage and Offloading (FPSO). The topology and parameter estimation will update as data changes/improve to capture the system behaviour to assess MIC likelihood, which helps in decision-making to control and mitigate MIC threats.

Keywords: Corrosion, Microbiologically Influenced Corrosion (MIC), Learning-based Bayesian network (LBN), Bayesian learning, Floating, Production, Storage and Offloading (FPSO)

3.1 Introduction

Corrosion is a severe threat to asset integrity, especially in oil and gas industry. It has been estimated to cost US$2.5 trillion in 2013 globally (Gerhardus et al., 2016). However, corrosion failures resulting from MIC account for 20% of the global cost (Liengen et al., 2014; Sorensen et al., 2012). Additionally, in the oil and gas industry, 30% of equipment damage is attributed to MIC (Revie, 2015), which can be further divided into internal and external corrosion modes. The former account for 40% of failures in underground pipelines, whereas the latter varies from 20-30% (Kaduková et al., 2014; Revie, 2015).

NACE Standard TM0212 defines MIC as a microorganism activity on a biofilm attached to a corroded metal surface (Tm et al., 2012). A biofilm is a consortium of microorganisms and bacteria attached to a metal surface (Liengen et al., 2014; Sorensen et al., 2012). The activities

of microorganisms in the biofilm change the electrochemical conditions at metal-solution interface, which results in enhancing the corrosion process (Sooknah, et al., 2007; Taleb-Berrouane, et al., 2019). MIC is a process that consists of initiation and propagation. The former occurs when a pipeline fluid is exposed to a pipeline surface containing free-floating microorganisms (planktonic); a portion of these microorganisms gets attached to the metal surface and form a biofilm. The latter occurs when a consortium of microorganisms attracts more microorganisms and forms exopolysaccharides to adhere to the surface. Once the biofilms become mature, they act as a channel for releasing metabolites and nutrient requirements of microorganisms (Skovhus et al., 2017). The corrosive microorganism is not limited to bacteria; it also includes methanogens and fungi. The corrosion process depends on oxidation of anode or reduction of cathode. The microorganisms only require an exchange of electrons from either oxidation or reduction, accelerating the rate of oxidation or reduction (Revie, 2015). They can be further divided based on the source of energy, oxygen requirement and favourability of the environment (B. Little et al., 2000). Common examples are sulphate-reducing bacteria (SRB), manganese-reducing and iron-reducing bacteria. Among them, SRB is considered to be highly responsible for MIC. They reduce sulphate underneath the biofilm, which is highly corrosive (Hashemi et al., 2018). Literature has revealed the presence of the following microorganisms in MIC; sulfate-reducing microorganisms (SRM) (Cord-Ruwisch et al., 1987), thiosulfate-reducing bacteria (TRB) (Liang et al., 2014), nitrate-reducing microorganisms (NRM) (Lahme et al., 2019), acid-producing bacteria (APB) (Gu, 2014), iron-oxidizing and iron-reducing bacteria (IOB, IRB) (Ray et al., 2010; Valencia-Cantero & Peña-Cabriales, 2014) and biofilm-forming microorganisms (Vigneron et al., 2016). The taxonomy of microorganisms helps to identify their presence in FPSO data. The identified type of microorganisms in this study's sample were SRM, methanogens and IRM. SRM is responsible for both chemical MIC (CMIC) and electrical MIC (EMIC) (Nicoletti, 2020). The former is a

mechanism in which microbiological metabolism is indirectly causing corrosion in pipelines, which rely on crude oil for carbon and energy sources (sulfide). The reduction of electron acceptor's metabolism produces highly corrosive products (Enning et al., 2012; Enning & Garrelfs, 2014). However, unlike CMIC, EMIC directly involves microbes with corrosion; they scavenge the electron from iron alloy surface and do not require external electron donors. The corrosion rate of EMIC in comparison to CMIC is significantly higher and thus more technically relevant (Enning et al., 2012). Methanogens produce methane in the presence of carbon dioxide and hydrogen, which also causes EMIC (Kip et al., 2017). IRB, Deferribacteraceae (5% relative presence), can use iron or nitrate as electron acceptors (Vigneron et al., 2016). The biofilm presence is detected with respect to Thermoanaerobacter and Caminicella; both are known to be capable of biofilm and spore-formation in the presence of thio-sulfate reducing members (Peng et al., 2016; Verbeke et al., 2014).

A review of the available literature shows that there is limited research on risk assessment of MIC failures, compared to monitoring, mechanisms, inhibition or prevention and biofilm formation (Hashemi et al., 2018; Taleb Berrouane, 2020). Hence, the challenge is not the availability of data, but to convert it to a robust model for risk assessment (Ben Seghier et al., 2021; Dawuda et al., 2021; Sorensen et al., 2012). The investigation of MIC risk assessment/modelling maturity trend is carried out using two databases: Web of Science and Scopus. Keywords such as MIC risk assessment or MIC modelling were selected to perform an advanced search in the scientific literature. Results obtained from the databases are combined using MS Excel. Figure 3-1 illustrates the investigation of MIC risk assessment. It can be observed that MIC risk assessment studies have significantly increased in the past decade. This shift reflects MIC knowledge evolution and, therefore, models/techniques' development to detect and predict its threat.

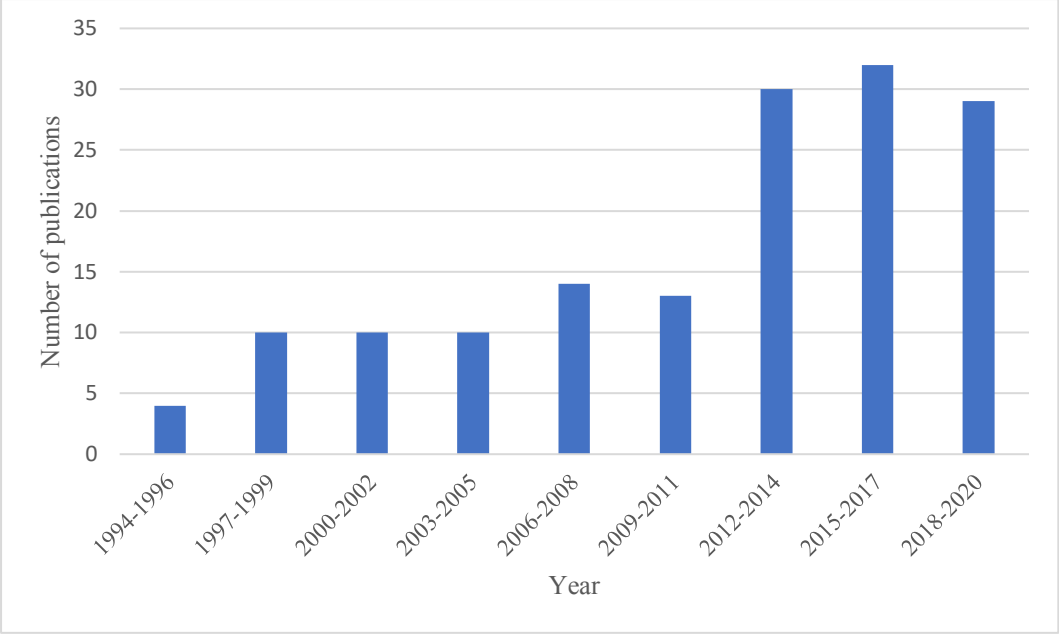


Figure 3-1 The evolution of MIC risk assessment/modelling

Prediction of MIC is a complex task, as the activities of microorganisms change along with their interaction with biotic and abiotic factors, which results in either enhancement or diminution of MIC activity over time. The first attempt to quantify MIC risk, made in 2002, depended on several factors such as water presence, water wetting, pH, salinity and temperature; however, it did not incorporate biological parameters (Pots et al., 2002). Later on, the model was improved by considering the biological parameter to enhance the prediction of MIC and the effect of biocide (Maxwell; Campbell et al., 2006). Another attempt was made, focusing on microbiological growth, to assess MIC occurrence, considering operating pipeline parameters and water chemistry (Sooknah, Papavinasa, et al., 2007). Fuzzy logic and a neural network were combined to evaluate the risk of bio-fouling assisted corrosion at a particular section of a pipeline by considering hydraulic, biological and surface conditions (Urquidi-Macdonald et al., 2014). Other attempts include a risk matrix to assess MIC risk with limited factors (Kaduková et al., 2014). A semi-quantitative analysis was performed that incorporated two factors; prediction factors and monitoring factors, to assess MIC risk (Y. Wang & Jain, 2016). In 2018, MIC causal factors' dependencies were considered using the BN to predict

MIC (G. Liu et al., 2018). Additionally, another study was carried out using BN, which is not limited to dependencies, but also considered synergic interactions to predict MIC (Taleb-Berrouane et al., 2018). The early risk assessment approaches did not consider consequences of failure but only focused on material degradation due to MIC (Andersen et al., 2017; Wolodko et al., 2018).

Molecular modelling techniques were also used to capture the presence and activities of microorganisms. Molecular Microbiological Methods (MMM) monitor microorganisms' distribution and help analyze MIC risk factors and pitting corrosion rates. These calculations estimate the number of MIC microorganisms with reaction stoichiometric and electron flow (Sorensen et al., 2012). Another study estimates MIC causing microorganisms based on DNA enumeration to assess MIC potential (Skovhus et al., 2012). An early work using molecular techniques shows that cultivation independence is reliable for bacteria identification and quantification, in contrast with a Most Probable Number (MPN) method. Other studies conducted by the same research group investigated the similarities and differences of bacterial populations from scale and produced water (Larsen et al., 2008). Also, microbiological activities were measured to design an early warning strategy to detect MIC in pipelines (Larsen et al., 2013). A molecular modelling technique was also applied to study growth of pits underneath biofilm and favorable conditions for this; it also investigated the role of hydrogen sulfide (HS^-) for microbiologically influenced pitting (Ezenwa et al., 2019).

Quantitative Risk Assessment (QRA) has been applied for MIC susceptibility to identify asset integrity threats. However, there is a need to develop a dynamic QRA, which can help minimize the loss by providing a data-based decision-making process. The data obtained from physical, chemical and biological parameters are important to capture synergic interactions, dependencies and causalities (Taleb-berrrouane & Khan, 2018; Wolodko et al., 2018). BN is considered to be a popular tool to model MIC (Dawuda et al., 2021; Taleb Berrouane, 2020;

Taleb-Berrouane et al., 2018; Taleb-Berrouane, et al., 2019; Taleb-berrrouane & Khan, 2018). This modelling tool (i.e., the BN) is particularly suitable for risk analysis (Deyab et al., 2018; Kabir et al., 2019; Kamil, Khan, et al., 2019; Emergency Response Plan Assessment Using Bayesian Belief Networks, 2017; Taleb-Berrouane & Khan, 2019; Yang et al., 2020) and reliability analysis (Bougofa et al., 2021), as it offers flexibility when adding new parameters to the network. Unlike Petri nets (Kamil, Taleb-Berrouane, et al., 2019; Talebberrouane et al., 2016; Taleb-berrouane et al., 2018; Taleb-Berrouane et al., 2020; Taleb-Berrouane, 2019) or Markov chains (Ayello et al., 2014), this addition can be done without disturbing the overall structure of the model or other dependencies (i.e., directed arcs). Besides, the BN can easily incorporate new evidence and generate posterior probabilities useful for the analysis (Kamil, Khan, et al., 2019). Ayello et al. (Koch et al., 2015) used a BN-based model for internal and external corrosion and incorporated a MIC mechanism in the model. However, the development of MIC mechanism is limited in this study. Another study conducted by Koch et al. utilized the BN approach to model MIC, which is limited to considering sulfate-reducing microorganisms as the only causal factor and ignores other known factors in the model (Koch et al., 2015). Another BN-based approach enhances MIC mechanism modelling as a subsystem in internal corrosion assessment by considering the operational conditions, water conditions and bacteria presence (Shabarchin & Tesfamariam, 2016). Recently, a static BN-based approach has been used to quantify MIC susceptibility based on operating parameters, fluid chemistry, settlement parameters, material parameters, operating history, mitigation parameters and symptoms of MIC presence (Taleb-Berrouane et al., 2018). The study incorporated various factors; the disadvantage is an absence of data-driven BN parameters and lack of dynamicity in the model. The latter is significantly addressed in a more recent study (Kannan et al., 2020). The study utilizes a static BN approach to consider failure analysis history, maintenance history, material properties and operational data, which account for the

60 nodes (causal factors) BN model. The advantage of the model is incorporation of dynamic behaviour in the model by considering five nodes as time-dependent, namely, field PCR, iron carbonate, water cut, hydrogen sulfide concentration and wall thickness. The study does not utilize MIC data due to their unavailability in open literature and thus, relies on subjective judgement. The scenario-based results of the study are used in the benchmarking of the proposed model results. A more recent work by Taleb-Berrouane et al. (Taleb-Berrouane et al., 2021) proposed the "Adaptive Bow-Tie (ABT)" approach to adapt (i.e., capture) the changes in a database to the bow-tie structure and applied it to MIC risk assessment. The new study provides a dynamically changing structure in a very simple and innovative approach; however, it cannot capture complex dependencies.

Existing Bayesian approaches demonstrate that heuristic observations and expert judgment are the primary data sources for MIC modelling. The interaction of causal factors observed at a laboratory scale is often missing in high-level heuristic observations (Kannan et al., 2020). There is a need for a learning model that extracts the information from available data to evaluate MIC threat as the next step towards MIC risk modelling (Skovhus et al., 2017). Machine learning methods are gaining attention in engineering risk assessment due to their ability to extract features from data. Data can be available in textual or numerical form. However, the latter case is most common in risk assessment. Machine learning methods often used in engineering include but not limited to artificial neural network (ANN), support vector machine (SVM) and Naive Bayes classifier (Hegde & Rokseth, 2020). A comparison of popular machine learning methods used in process safety engineering (PSE) is shown in Table 3-1.

- ANN algorithms are similar to neurons in human brains. The neurons (contains functions) in ANN are connected by links. Link weightage is adjusted based on training data to improve ANN performance.

- SVM algorithm is based on developing hyperplanes depending upon training data sets. Hyperplanes serve as a basis for data classification.
- Naïve Bayes algorithm uses the Bayes rule with a strong conditional independence assumption, i.e. Conditional independence of feature variables given the class variable (Goh et al., 2018).

Table 3-1 Summary of common machine learning models used in risk assessment

Machine learning models	**Features**	**Limitations**	**References**
Artificial Neural Network (ANN)	Require less training data Able to determine non-linear relationship between input and output variable	Input and output variables relationships are generalized, i.e., overfitting Computational demand is high	(Hegde & Rokseth, 2020; Tu, 1996)
Support Vector Machine (SVM)	Computational demand is less Superior predictions	Performance depends on training data set (support vectors)	(Ma et al., 2009)
Naïve Bayes	Performance is good on small and large data sets	Conditional independence of feature variables given class variable	(Adedigba et al., 2017; Jensen & Nielsen, 2007)

A data-driven model is proposed called the learning-based Bayesian network (LBN) model, which will help minimize the above-mentioned research gap. The topology of a Bayesian network considers expert judgement to identify the interaction between causal factors; however, this is time-consuming and error-prone (Larrañaga et al., 2013). The LBN model is based on the Bayesian learning algorithm, which develops topology from the obtained data and the parameters required in BN. There is a need to leverage operational and microbiological data in determining MIC likelihood. The key advantage would be capturing system behaviour in the process. The parameters will be adapted as the new data becomes available. The BN model will help to assess MIC likelihood and operational decision-making for mitigating and controlling measures. It will also help to identify critical equipment with high risk.

Section 3.2 discusses details about Bayesian learning and selected algorithm. Section 3.3 of the study is devoted to LBN model, followed by its applicability in section 3.4 which includes, application of LBN model with complete and incomplete data sets followed model testing and its requirement in terms of FPSO data. Section 3.5 shows LBN model validation and benchmarking on MIC-induced failures. The conclusion obtained from the study is presented in section 3.6. Figure 3-2 depicts a visual representation of an overview of the study.

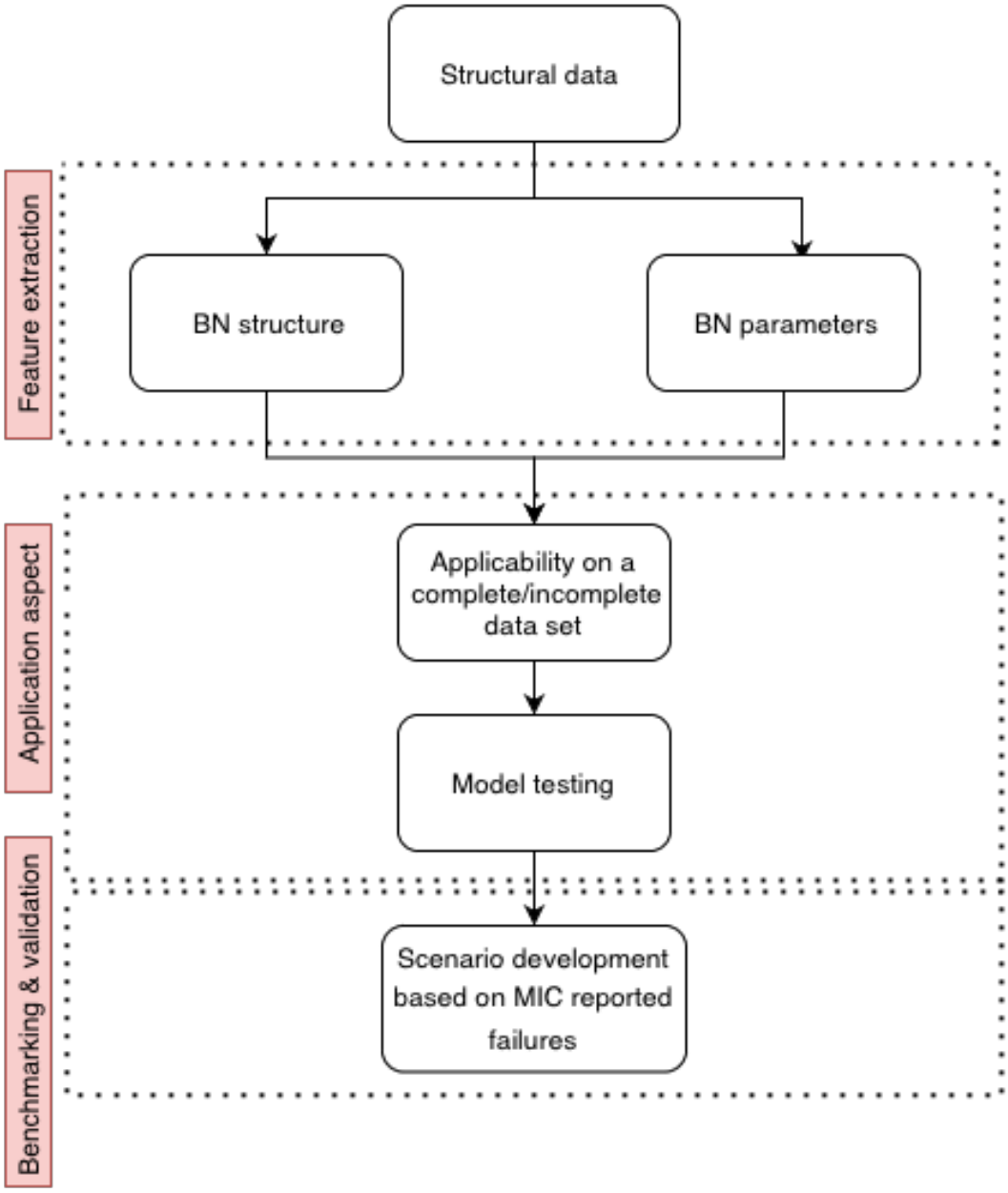


Figure 3-2 Overview of the study conducted

3.2 The Concept of Bayesian Learning

BN is considered to be a prime probabilistic graphical model for reasoning under uncertainty. Its inherent capability, representation of accident scenario and inference makes it unique to apply in multiple domains. Its use in MIC domain is well established in literature but lacks in learning aspect of BN model. This section will discuss ways to learn BN model from structural data with its pros and cons.

3.2.1 Bayesian Network and its Structural Learning

BN is a widely used probabilistic method in safety and risk analysis due to its flexibility in incorporating complex causal dependencies and graphical representation of cause-effect. For more details about BN features and applicability in Process safety, interested readers are referred to (Kamil, Khan, et al., 2019; Khakzad et al., 2013; Taleb-Berrouane et al., 2020).

There are two ways to define the BN structure: expert judgement (knowledge-driven) or utilizing the data to obtain variable correlations (data-driven) (Adedigba et al., 2018). The former method is prone to error due to a lack of expert knowledge about variable correlations and an inability to reach a consensus about its structure. It is also time-consuming and challenging if there are numerous variables (Adedigba et al., 2017; Neapolitan, 2004). However, the data-driven method overcomes the challenges and can learn the BN structure, given that the data set consists of all the variables of interest. The main challenge is an exponential increase in possible structures of BN as the number of nodes (m) increases, as shown in Equation (1) (Jensen & Nielsen, 2007). For example, for m=10, the possible BN structure is approximately $4.2*10^{18}$. A method is needed to maximize BN structure, given the data set.

$$f(m) = \sum_{i=1}^{m} (-1)^{i+1} \frac{m!}{(m-i)!\,m!} 2^{i(m-i)} f(m-1) \qquad (1)$$

There are two methods for learning the Bayesian network structure: the constraint-based method and Bayesian score-based method (Adedigba et al., 2017; Dash & Druzdzel, 1999; Jensen & Nielsen, 2007). Table 3-2 illustrates the difference between both methods. The LBN model uses the Bayesian search-based method based to obtain BN's topology.

Table 3-2 Major differences in Bayesian learning methods

Parameter	Constraint-based method	Bayesian score-based method

Conditional independence	Conditional independence test is required	Bayesian score-based method can be applied if conditional independence test fails BN
Significance level	Arbitrary significance level to determine independencies when independence test does not hold	Score and search to find optimal
Topology of the network	Error in initial stage of search process leads to different structure	The structure with highest score is taken as optimal structure
Features	Quick and can deal with latent variables	Relatively slow and can deal with incomplete data set

The objective is to obtain an optimal BN structure from the data set. The score-based method has two main components, namely, a score function and a search procedure. The score function should have the ability to balance structure's accuracy, given its number of correlations from the input data set and computational tractability. The present study is based on structure learning introduced by (Cooper & Herskovits, 1992) and refined by (Heckerman et al., 1995). Let us assume that D is a data set of cases and V is a set of variables present in data set D. Z_S is a belief structure consisting of variables V from the data set D. To rank the structure, posterior probability can be calculated by Equation (2).

$$P(Z_s|D) = \frac{P(Z_s, D)}{P(D)} = P(Z_S)P(D|Z_S) \quad (2)$$

The P(D) is a constant and does not depend on Z_s; therefore, it is not necessary to evaluate it for comparing two structures. Equation (2) suggests two terms need to be evaluated, the prior probability $P(Z_S)$ and the marginal probability $P(D|Z_S)$. The main computational challenge is to calculate the marginal probability, given the data set D, and deal with the parameters of the model Z_P, assuming that the data set consists of discrete variables.

$$P(D|Z_S) = \int_{Z_P} P(D|Z_S, Z_P) f(Z_p|Z_S) dZ_P \quad (3)$$

Equation (3) integral is over all the parameters and has a BN structure with same Z_S but different conditional probabilities ($f(Z_P|Z_S)$). According to (Cooper & Herskovits, 1992; Jensen & Nielsen, 2007), the integral can be reduced to a counting problem based upon the following assumptions:

- The data set consists of discrete variables
- The cases in data set are independent of BN structure
- The cases in data set are complete (assumption can be relaxed to accommodate missing data, interested reader is referred to (Cooper & Herskovits, 1992))
- The prior distribution of the parameters in BN is uniform

Therefore, for a BN structure, Z_S, given the data set, D, the score function is shown in Equation (4).

$$P(Z_S, D) = P(Z_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(S_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} S_{ijk}! \quad (4)$$

where,

r_i is the state number of V_i

S_{ijk} denotes the number of samples in a data set D with Vi in its k^{th} configuration and pa(Vi) in the j^{th} configuration, Sij is estimated using Equation (5).

$$S_{ij} = \sum_{k=1}^{r_i} S_{ijk} \quad (5)$$

The goal is to obtain an optimal BN structure from all possible configurations of BN. The scoring function helps to convert the Bayesian structural learning task to a parameter optimization problem. Searching the BN in all potential spaces is exponential as nodes increases. A method is needed to maximize Equation (4). A heuristic method such as the K2 algorithm is available in the literature to overcome the challenge. The K2 algorithm first

assumes a prior structure (usually an empty or a randomly chosen structure) and calculates the gain by adding a parent node. The directed arc from parent to child node must result in an acyclic graph. Once the parent node's addition no longer increases the resulting structure probability, no other parent nodes are needed. The scoring functions in Equation (4) calculate the score of candidate BN structures with the K2 algorithm to search for the optimal BN with the highest score (Cooper & Herskovits, 1992; Jensen & Nielsen, 2007).

To show an illustration of score function, a simple example is shown in Figure 3-3 to compare two BN structures and decide based on score function which is more likely to occur. Equation (4) given by (Cooper & Herskovits, 1992) is used to check each structure's score. A hypothetical data are taken in Table 3-3 to calculate each structure score in Figure 3-3.

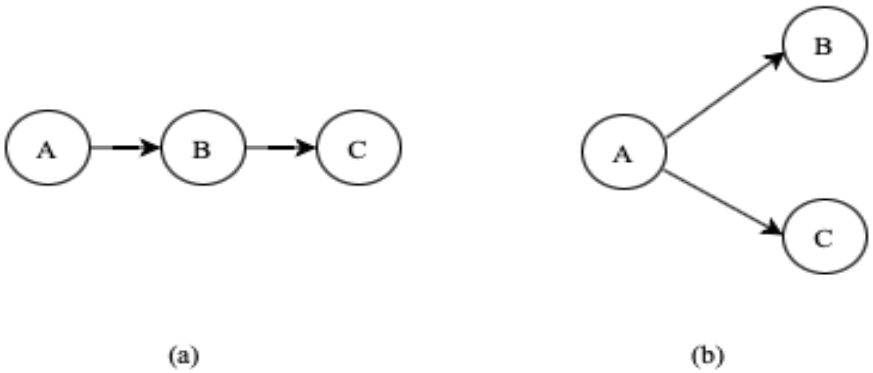


Figure 3-3 Two possible structures of three node BN denoted as (a) and (b)

Table 3-3 A data set to test the likelihood of BN structure's

Number of cases	A	B	C
1	0	1	1
2	0	0	0
3	0	1	1
4	1	0	1

5	0	1	1
6	0	1	1
7	1	1	1
8	1	0	0
9	1	1	1
10	1	0	0

The first BN structure results in $P(a, D) = 2.78 * 10^{-10}$, whereas the other BN structure score, given the data set is $P(b, D) = 5.57 * 10^{-11}$. This shows structure (a) is more likely to occur, compared to structure (b).

3.2.2 Bayesian Parameter Learning

The most common parameter learning method used is maximum likelihood estimation (MLE). When input data are complete, MLE estimates the conditional probability θ of a parameter of interest in terms of log-likelihood. For a BN structure, Z, given the data set, D, the MLE is expressed in Equation (6) as (Jensen & Nielsen, 2007):

$$LL(Z|D) = \sum_{d \epsilon D} log_2 P(d|Z) \tag{6}$$

The MLE ($\widehat{\theta}$) of θ is shown in Equation (7) as

$$\widehat{\theta} = \arg\max_{\theta} \ LL(Z|D) \tag{7}$$

The MLE is a well-suited method when input data are complete. However, in practice and due to sensor malfunction, technical or human error, some variables may not be observed (latent/hidden variables); data often consist of missing values. Details on missing values will be discussed in section 3.4.5. To deal with parameter estimating with missing values, an Expectation-Maximization (EM) algorithm introduced by (Dempster et al., 1977) is considered

for parameter learning. It alters between two steps: the expectation step and the maximization step. The prior guessed a distribution for the parameter of interest (missing value), then the latter estimated the parameter by maximizing lower bound of the likelihood function. The algorithm repeated until it converged (the probability no longer changed) or met the termination criteria. Assume Z is a complete data set with density $P(Z|\theta)$. If the data set is complete, then the objective would be to maximize the following function in Equation (8):

$$L(\theta|Z) \propto p(Z|\theta) \tag{8}$$

When the data set consists of some observed values and missing values, this means Z is partially observed. Therefore, Z can be written to include both observed and unobserved data as, $Z = (Z_{obs}Z_{unobs})$ in Equation (9).

$$L_{obs}(\theta|Z_{obs}) \propto \int p(Z_{obs}Z_{unobs}|\theta)\, dZ_{unobs} \tag{9}$$

The EM steps are as follows:

1. Expectation step: find the expected value of log-likelihood function given the observed and present estimate of parameters (Imtiaz & Shah, 2008):

$$E\left(\theta\middle|\theta^{(t)}\right) = \int L_{obs}(\theta|Z_{obs}, Z_{unobs})p(Z_{unobs}|Z_{obs}|\theta^{t})\, dZ_{unobs} \tag{10}$$

2. Maximizing step: find the value of $\theta^{(t+1)}$ that maximizes the expectation step:

$$\theta^{t+1} = \arg\max_{\theta}\ E\left(\theta\middle|\theta^{(t)}\right) \tag{11}$$

3.3 The LBN Model

The proposed model aims to utilize process operating and microbiological parameters to avoid equipment failure (leakage) due to MIC. Figure 3-4 illustrates each step involved in the model development to assess MIC likelihood. To utilize process operating and microbiological data, steps 3.3.1-3.3.4 describe the data-driven MIC model.

3.3.1 System Identification

The first step is to identify a system to apply the proposed model. It can be identified by analyzing its history, type of process fluid, corrosion signs, and exposure to corrosive conditions.

3.3.2 System based Operational and Microbiological Data Collection

The identified system's operational and microbiological data collection is a crucial step. After reviewing the literature, the range for each parameter has been determined based on reported incidents of MIC. Then, the data collected for each parameter is analyzed to find whether it falls within the specified range. The other critical parameter is microbiological data, which will depict the presence of microorganisms and their types. The data can be obtained from 16 rRNA gene sequencing.

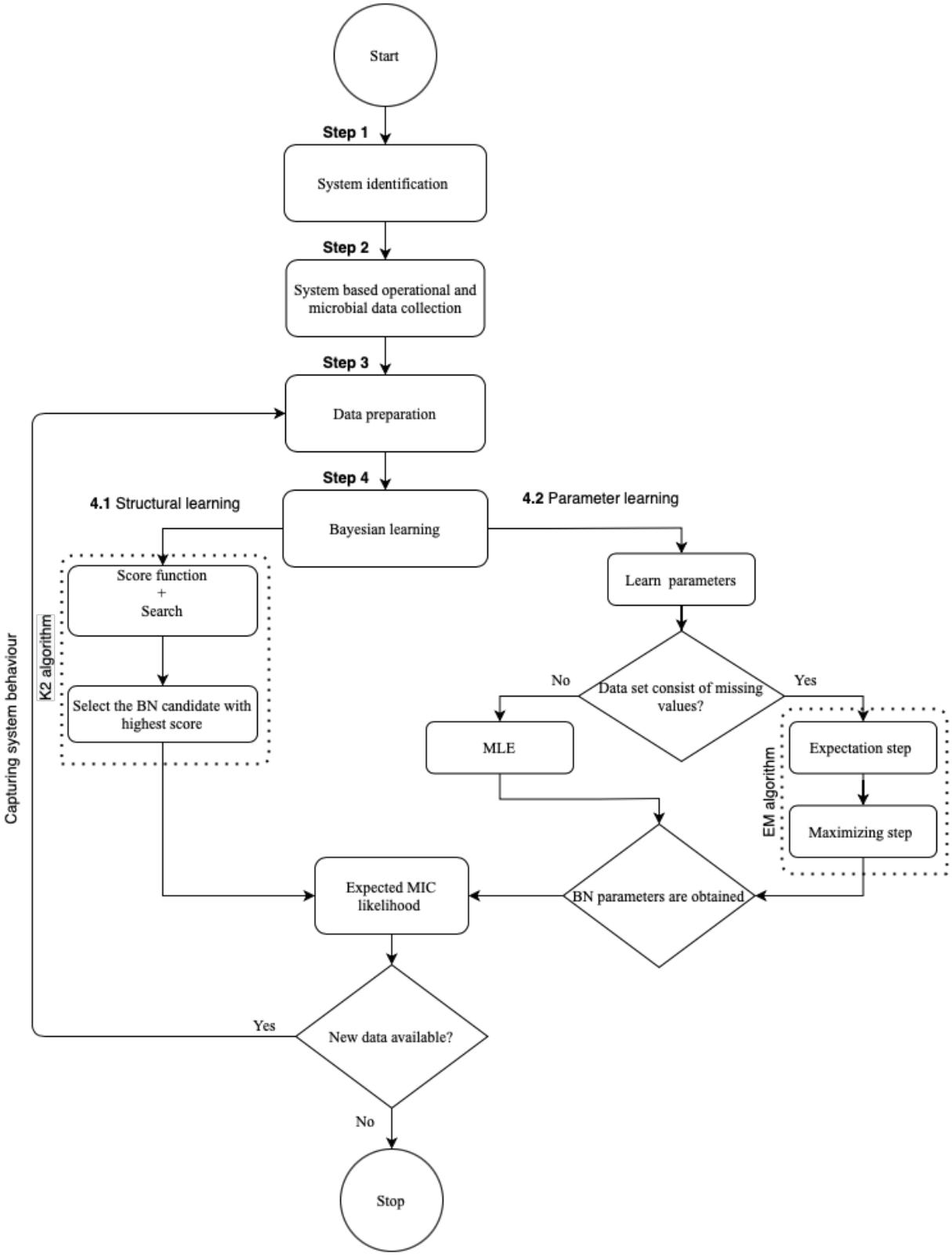


Figure 3-4 The proposed data-driven MIC model

3.3.3 Data Preparation

The collection of operational and microbiological data was an important step. However, there is a challenge with collecting the desired amount of data for the data-driven approach. The data usually unavailable or insufficient for the data-driven approach. The present study has the latter problem. A lower and upper bound of data are taken from each parameter's available data to overcome data scarcity, which allows using n number of randomly generated data sets between the lower and upper bounds.

The parametric data are converted to non-parametric data, which helps to uniform each parameter for Bayesian learning. Therefore, each variable's parametric data are converted into binary state 0 or 1, denoting the variable's absence or presence. If parametric data lies between the lower and upper bounds as discussed before, it will be indicated as 1 or 0.

3.3.4 Bayesian Learning

Bayesian network requires eliciting the BN topology and its parameter estimation. Bayesian learning will help to eliminate the subjective decision in deciding structure and parameters necessary to reduce uncertainty. The same dataset can be used to learn both structure and parameters. Firstly, structured learning can be initiated to obtain an optimal BN structure followed by estimating its parameters.

3.3.4.1 Structural Learning

Once data are prepared for Bayesian learning, the factors such as operational and microbiological parameters can be provided as a text file to GeNIE Modeler (*GeNie Software*, 2023). The presence of microorganisms can be classified based on the taxonomy classes. In addition, expert opinion can be incorporated in terms of background information prior to

learning the structure, such as forced/forbid arcs from parent to child node. The K2 algorithm searches BN with the highest score and is referred to as the optimal BN for the data set.

3.3.4.2 Parameter Learning

Parameter learning is performed using the EM algorithm which is a general method of obtaining MLE with missing values of parameters (Imtiaz & Shah, 2008). When the input data set is complete and does not contain any missing values, the EM algorithm works as MLE. However, in the case of missing data, EM algorithm has two steps, namely, expectation and maximization. It maximizes the Z_{obs} by maximizing the expected value of log-likelihood of the complete data set. The algorithm iterates between two steps, as shown in Equations (10) & (11), with the parameter's initial value θ^t until convergence. The convergence is based on the missing data in the data set. When more data are missing, the convergence will be lower (Imtiaz & Shah, 2008).

3.4 Application of LBN Model

Data utilized in the study is available from the produced water samples obtained from the FPSO platform located in North America for polymerase chain reaction (PCR) to identify relative abundance of microorganisms (Nicoletti, 2020). Figure 3-5 illustrates the schematic of water and crude oil through the topside processing machinery of FPSO platform adapted from Nicoletti (Nicoletti, 2020). The green lines indicate water flow and sampling point, while black lines denote crude oil flow. The produced water samples are collected from the locations shown in Figure 3-5, namely, SC 1013, SC 1032, SC 1035 and SC 1037.

LBN model comprises of four steps: (1) System identification, (2) System based operational and microbial data collection, (3) Data preparation and (4) Bayesian learning-structure and parameter.

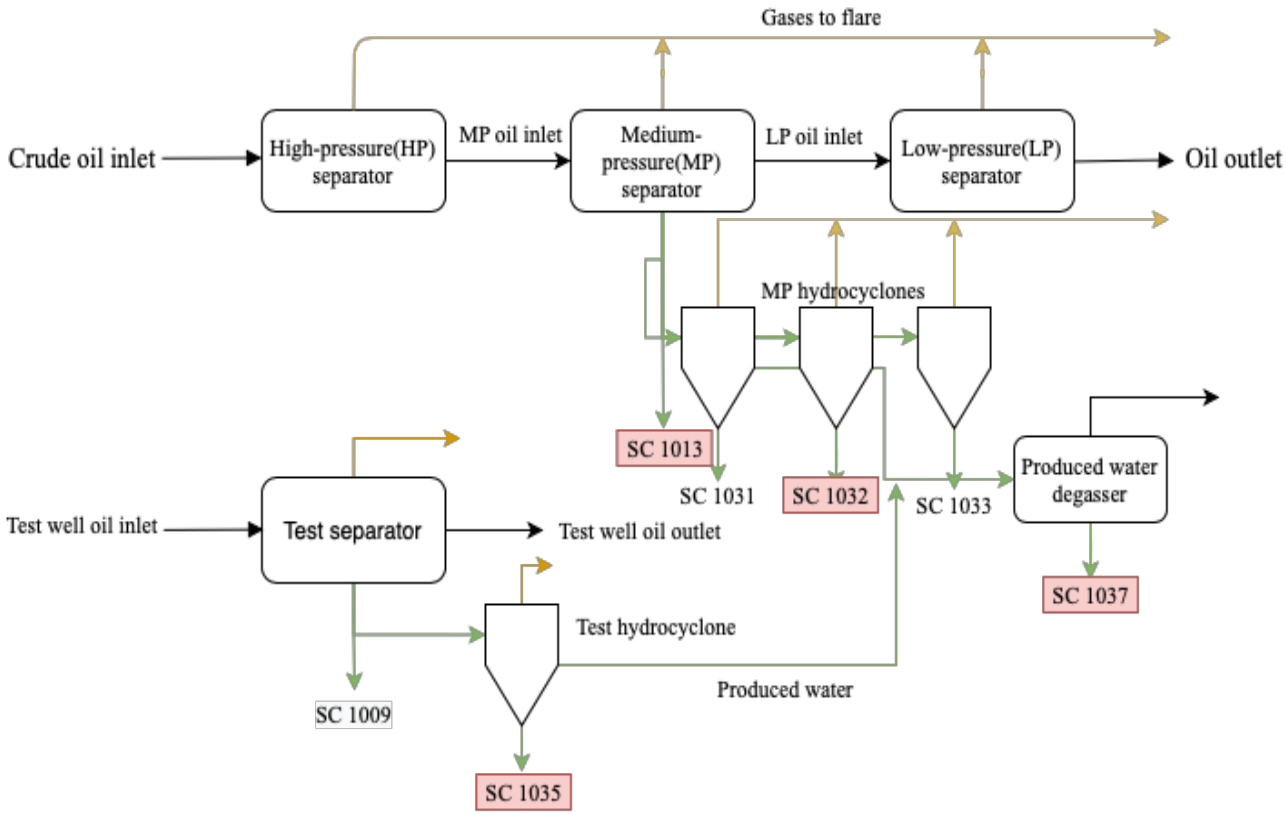


Figure 3-5 A flow diagram of FPSO platform topside view adapted from Nicoletti (Nicoletti, 2020)

3.4.1 System Identification

The first step is to identify a system to apply the proposed model. The offshore FPSO platform's process equipment from where produced water samples were collected are selected due to their exposure to a corrosive environment.

3.4.2 System based Operational Data and Microbiological Data Collection

The three operational parameters considered are temperature, pH and flow velocity of the process fluid. Flow velocity data are not available and assumed to be low as a favourable condition for biofilm growth (validated by the presence of biofilm-

forming- Thermoanaerobacter and Caminicella in the sample). The operational data are not shown due to proprietary issues.

The other critical parameter is microbiological data, which will depict the presence of microorganisms and their types. The present study utilizes the available data from a produced water sample of subsystems from the FPSO facility located in North America, for 16 rRNA gene sequencing. Biofilm film is detected with respect to Thermoanaerobacter and Caminicella. Both are known to be capable of biofilm and spore formation in the presence of thio-sulfate reducing members (Peng et al., 2016; Verbeke et al., 2014). For more details on the microorganisms group and their relative abundance, the reader can refer to the work of Nicoletti (Nicoletti, 2020). The taxonomy is classified as iron-reducing, methanogenic, sulfate-reducing, and biofilm-forming microorganisms.

3.4.3 Data Preparation

The data set consists of mostly one or two data points, based on the experiment conducted to identify the types of microorganisms groups at different time intervals. To obtain an extensive data set for the model, the available data are considered to be the lower and upper bound of a data set. The random numbers generated in the described range will help address scarcity and facilitate learning the BN topology and parameters. Another critical task is determining if the data are in the range of a favourable limit and converting it into a non-parametric form. The range of each parameter is defined based on the lower and upper bound data, as discussed. For example, a favourable temperature for MIC is 10-95 ^{0}C; the identified process equipment operated in this range. If the temperature falls in the favourable range, the non-parametric data are denoted as 1 (yes), if not, 0 (no). A total of 360 data points were used for the model implementation.

3.4.4 Bayesian Learning

Bayesian learning is divided into two aspects: structural learning and parameter learning. The data set prepared in the previous step is used to learn both structure and parameters of the BN. The converted operational non-parametric data are grouped into favourable conditions, which include temperature, pH and flow velocity of a fluid, whereas microbiological groups present are grouped as microbiological activity. The score-based method (K2 algorithm) introduced by (Cooper & Herskovits, 1992; Heckerman et al., 1995) is used to obtain an optimal BN, which corresponds to the highest score among different candidate networks. The EM algorithm is used for parameter learning, which works as a Maximum Likelihood estimation with a complete data set. Figure 3-6 and Figure 3-7 show the BN structure learned from the data set for each location of the FPSO platform. It is worth noting that the data set from the locations SC 1013, 1032 and 1037 consist of same microorganisms group (i.e. Iron-reducing, Sulfate-reducing and Methanogens), thus have the same BN structure. However, in the case of SC 1035, Iron-reducing group is absent. The operational and microbiological data considered in the study vary slightly due to differences in the upper and lower bound of the variables. The learning is performed using default parameters of the algorithms.

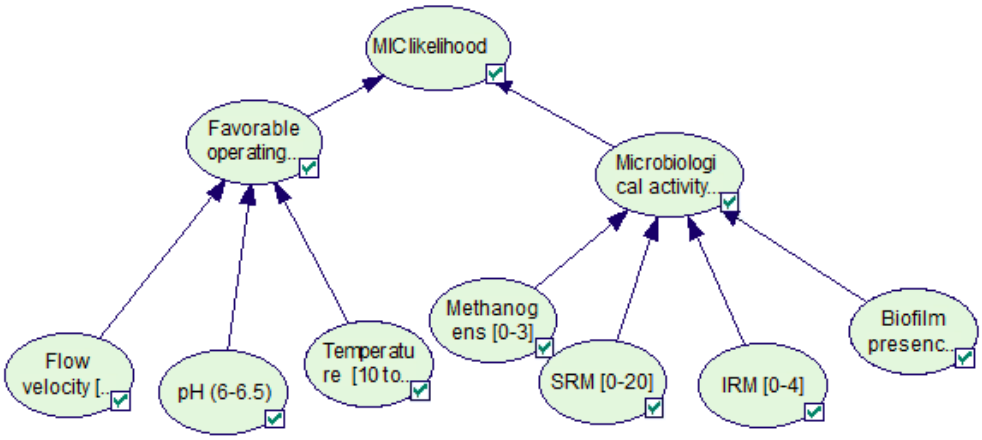


Figure 3-6 BN network learned from data set for SC 1013,1032 and 1037 locations

The MIC likelihood from the LBN model is shown in Table 3-4. The MIC likelihood from all four locations is not different from one and another. The input data plays a sole role in the estimation of parameters. The MIC likelihood from all the identified locations in Figure 3-5 has the same order of magnitude. In contrast, if we consider the study of corrosion coupon testing performed at a laboratory scale (Nicoletti, 2020), also suggests corrosion rate does not significantly varies from one location to another. Therefore, it can be established that the LBN model is capable of extracting BN from data set for MIC likelihood.

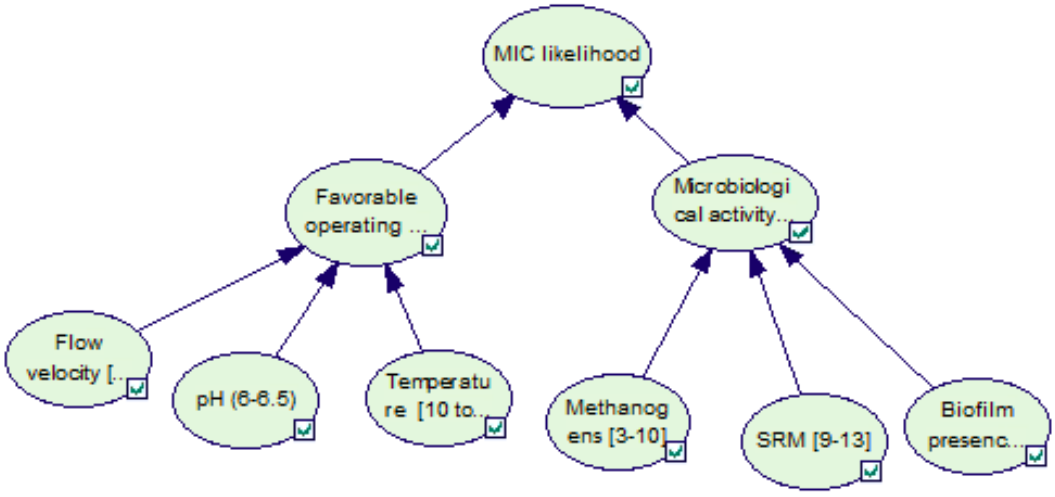


Figure 3-7 BN network learned for SC 1035 location

Table 3-4 MIC likelihood of FPSO platform equipment

Locations of FPSO platform	**SC 1013**	**SC 1032**	**SC 1035**	**SC 1037**
MIC likelihood	4.30E-02	4.70E-02	4.50E-02	5.20E-02

It is vital to use the MIC likelihood and convert it in respect to the maximum pitting rate. MIC is prone to cause pitting corrosion on metal surface that results in leakage of stored material. Therefore, expert opinion is used to develop a relationship of MIC likelihood to maximum pitting rate as shown in Figure 3-8. The two crucial factors in pitting corrosion are pit depth

and pit count on metal surface. The pit depth is categorized as significant or minor, based on the ratio of pit depth over the initial metal thickness. If pit depth reaches to 60% of metal thickness, it is considered as significant, otherwise, it is minor. The pit counts are discretized as P_1, P_2, P_3 and P_4. P_1 denotes the likelihood of pit counts between 0-10 while P_2 indicates 10-20, P_3 signifies 20-30 and P_4 represents 30 or more. The expected risk of pitting is calculated based on NACE standard (NACE RP0775, 2005) as the multiplication of likelihood of pitting rate by the weighted pitting rate. The result for prior risk obtained from the BN model is shown in Table 3-5. The risk is in the same order of magnitude due to less variability of data as discussed earlier. Note that in the maximum pitting rate node, there is one more state which exists due to the directed arc from MIC potential node to pitting rate node. It is called the "No" state, which accounts for the non-occurrence of MIC potential node.

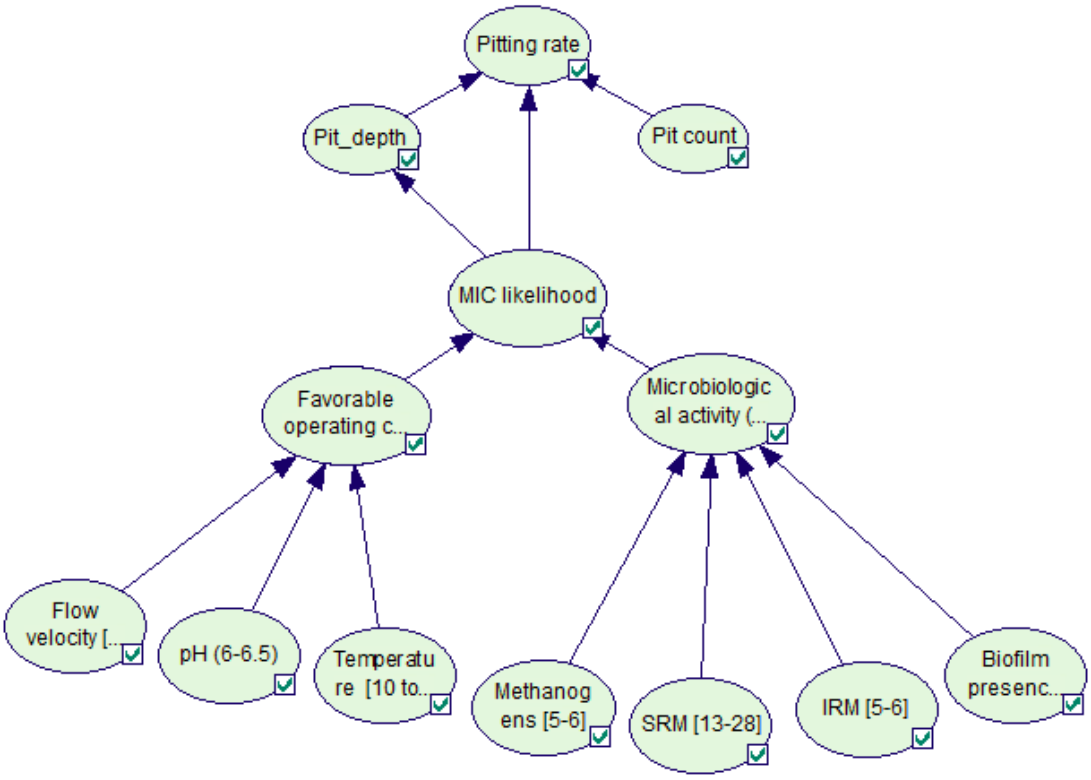


Figure 3-8 BN model for likelihood of pitting rate

Table 3-6 consists of three different risk values: R_A denotes the prior risk from the BN, as shown in *Table 3-5*. R_B represents the expected risk of pitting when microorganisms activity

and favourable operating conditions are detected in the system. R_C denotes the condition when a corrective measure is applied, such as using a biocide to control MIC. This provides a comparison of three states of the system: R_A (actual state), R_B (worst-case) and R_C (ideal case). It also represents the variability of the model. The risk of each pitting rate is increased by one order of magnitude. Note that the system's actual state (R_A) has already shown the risk of MIC at a laboratory scale (corrosion coupon testing) (Nicoletti, 2020). The R_C state is developed to illustrate LBN model behaviour when the operator takes control measures to lower MIC risk. The risk is reduced by two orders of magnitude in most pitting rates, as compared to R_A. It also represents a safe condition for the process to continue. The results also show that the SC 1037 location is slightly more likely to show pitting than others. This can be explained based on the more relative abundance of sulfate-reducing microorganisms than other mentioned locations.

Table 3-5 Expected risk of pitting in FPSO platform

Maximum pitting rate	**Weighted pitting rate (mm/yr)**	**Weighted factor**	**Expected risk**			
			SC 1013	SC 1032	SC 1035	SC 1037
Low	<0.13	0.13	1.02E-04	1.12E-04	1.06E-04	1.24E-04
Moderate	0.13-.20	0.2	2.08E-05	2.28E-05	2.16E-05	2.53E-05
High	0.21-0.38	0.38	1.08E-05	1.18E-05	1.12E-05	1.31E-05
Severe	>0.38	0.55	3.03E-07	3.33E-07	3.15E-07	3.69E-07

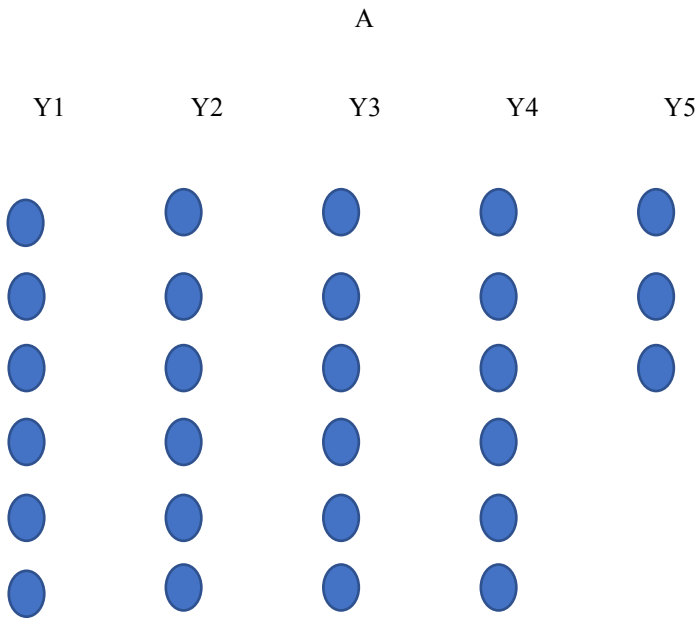
Table 3-6 Expected risk of pitting in FPSO platform, RA (prior probability), RB (microorganisms activity detected) and RC (corrective measure applied)

Maximum pitting rate	SC-1013			SC-1032			SC-1035			SC-1037		
	RA	RB	RC	RA	RB	RC	RA	RB	RC	RA	RB	RC
Low	1.02E-04	2.29E-03	5.15E-06	1.12E-04	2.28E-03	4.80E-06	1.06E-04	2.29E-03	5.11E-06	1.24E-04	2.31E-03	5.22E-06
Medium	2.08E-05	4.68E-04	1.05E-06	2.28E-05	4.66E-04	9.80E-07	2.16E-05	4.67E-04	1.04E-06	2.53E-05	4.71E-04	1.07E-06
High	1.08E-05	2.43E-04	5.47E-07	1.18E-05	2.42E-04	5.09E-07	1.12E-05	2.43E-04	5.42E-07	1.31E-05	2.45E-04	5.54E-07
Severe	3.03E-07	6.83E-06	1.54E-08	3.33E-07	6.81E-06	1.43E-08	3.15E-07	6.82E-06	1.52E-08	3.69E-07	6.88E-06	1.56E-08

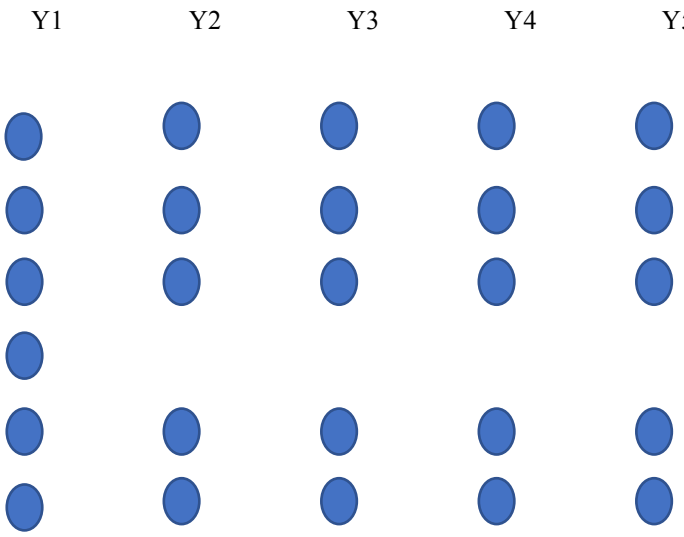
3.4.5 Application of the LBN Model with Missing Values

In the past decade, the oil and gas processing industry has generated enormous data, introducing challenges for process engineers to analyze and convert this information into valuable knowledge. We have demonstrated the application of LBN model on the complete data set. However, the structure and parameter learning are also capable of learning the BN in case of missing values. In process industries, missing values refer to a data entry in a data set with no relationship to process, such as (no data). The incomplete data (i.e., missing values of parameters) may not guarantee a satisfactory model performance, especially if the missing values are large and affect the variables correlation (Xu et al., 2015). In the present study, we will investigate how much model performance will be affected with respect to missing values for a 10 node BN structure (shown in Figure 3-6).

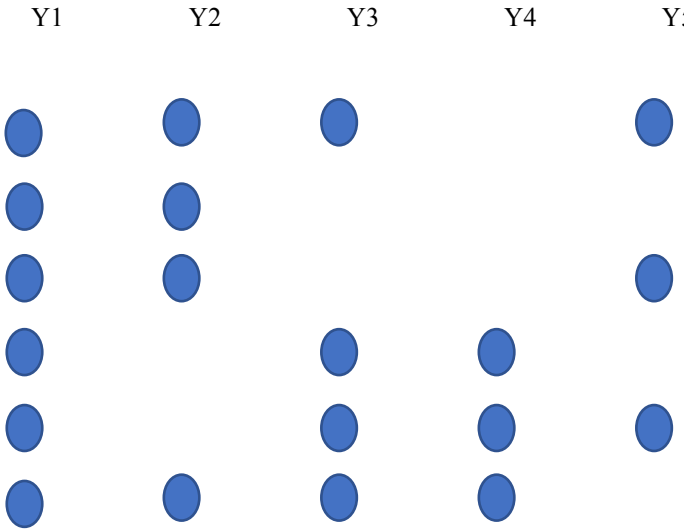
The first step in data cleaning is to analyze the pattern of incomplete data and possible reasons for incomplete data. Imtiaz et al. (Imtiaz & Shah, 2008) discussed commonly missing patterns and their possible causes, as shown in Figure 3-9.



B



C



D

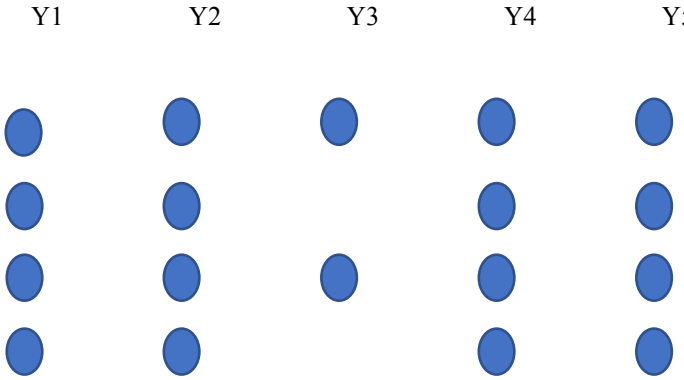


Figure 3-9 Missing data pattern adapted from Imtiaz et al. (Imtiaz & Shah, 2008)

- In case A, only one variable (Y3) contains missing values, this may be due to sensor breakdown.
- In case B, data entries of variables (Y2-Y5) are missing for the same data entry, reflecting a process shutdown due to a fault condition, and time stamps are the only available information.
- Case C depicts an irregular pattern; possible causes are outlier removal and sensor malfunction.
- Case D reflects the condition of multi-rate sampling by a regular missing pattern of one variable.

The present study considers two parameters: pH and biofilm with missing values shown in case A. The most straightforward procedure to deal with incomplete data are deletion of missing data. There are two deletion methods available in literature, namely, list-wise deletion and pairwise deletion. The former eliminates the time stamp containing missing values, whereas the latter only removes the observations for particular variables having missing values. The list-wise deletion will sacrifice a large amount of data and introduce more uncertainty in Bayesian learning. Therefore, pairwise deletion is selected for the investigation. The LBN model steps are followed with incomplete parametric data. Note that, due to similarity and less variability in the input data of SC 1013, 1032, 1035 and 1037, only SC 1037 data set is considered. The present study deals with 5%, 10%, 20% 30% and 40% missing values in parametric data of pH and biofilm. The same BN network shown in Figure 3-6 is obtained with respect to missing values which implies that the correlations of variables are not affected until there are 40% missing values in the data set. The detailed result of analysis is shown in Table

3-7. The percentage change is calculated to show how the expected risk varies compared to the data set consisting of missing values in pH and biofilm parameters. The negative sign indicates a decrease in expected risk when compared to prior risk with the complete data set. Figure 3-10 depicts the expected risk of pitting, corresponding to the percentage of missing values in the data set. Xu et al.(Xu et al., 2015) pointed out that the model performance is affected when large amount of data are missing, which is seen when missing values increases to 40%. The data set does not have a sufficient correlation of variables i.e., percentage change shows -100% in the last column of Table 3-7. However, the LBN model can learn BN when the missing values reach 30% or unless the correlations are affected. This exercise is conducted to show the capability of the model (degraded performance) to work with missing values in the data set, as long as they are not large enough to affect the correlation (i.e., 40% in the present case). It does not aim to answer how to handle missing values in data set.

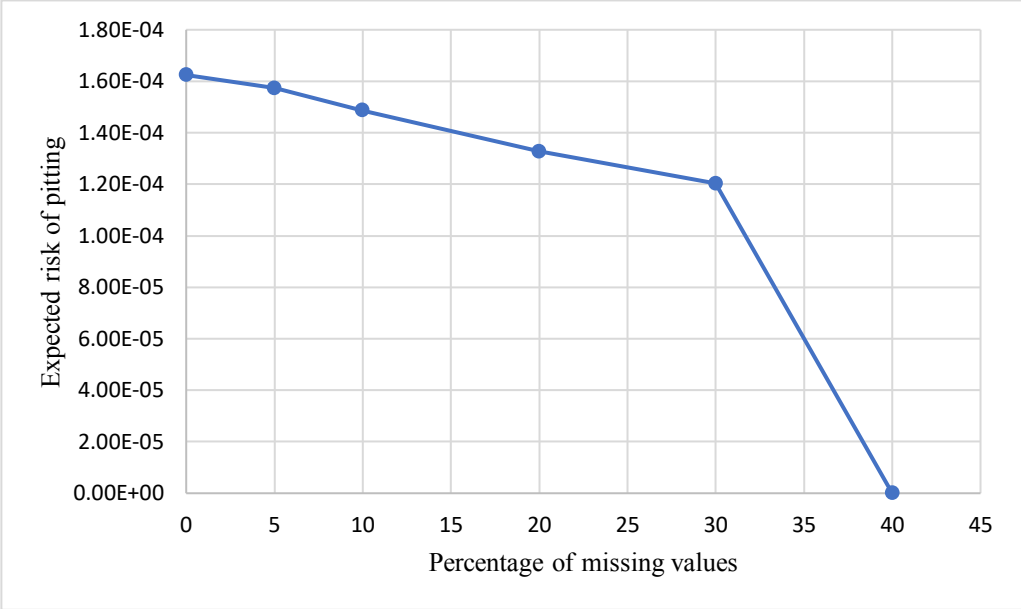


Figure 3-10 Bayesian learning with respect to missing values in data set

Table 3-7 Expected risk of pitting with respect to missing values in data set

Maximum pitting rate	Complete data set	5%	Percentage change	10%	Percentage change	20%	Percentage change	30%	Percentage change	40%	Percentage change
Low	1.24E-04	1.20E-04	-3.1	1.13E-04	-8.5	1.01E-04	-18.3	9.16E-05	-25.9	0	-100.00
Moderate	2.53E-05	2.45E-05	-3.1	2.31E-05	-8.5	2.06E-05	-18.3	1.87E-05	-25.9	0	-100.00
High	1.31E-05	1.27E-05	-3.1	1.20E-05	-8.5	1.07E-05	-18.3	9.72E-06	-25.9	0	-100.00
Severe	3.69E-07	3.57E-07	-3.1	3.38E-07	-8.5	3.02E-07	-18.3	2.73E-07	-25.9	0	-100.00

3.4.6 LBN Model Stability

To investigate data requirement for LBN model implementation shown in Figure 3-6 consists of 10 nodes. Each step mentioned in section 3.2 is carried out to develop and determine when the BN structure stability will be achieved. The initial data points considered were 50, which did not show any directed arcs from one variable to another, due to insufficient correlations. The data points increased with a step size of 25, which resulted in same conclusion. However, when the data points increased to 100, favourable conditions for an intermediate node developed with the directed arcs from all the operational parameters. This showed that the data had sufficient correlations to create directed arcs. Another step size of 25 resulted in the development of directed arcs in the microbiological activity node. At this stage, the directed arcs from intermediate nodes to the target node were unstable and kept changing when different iterations were performed. Another step size increase was the critical point in BN's stability (150 data points). At this stage, BN was stable and same BN is obtained when different iterations were performed. Note that the 10 node BN structure requires at least 150 data points to implement the LBN model. Figure 3-11 illustrates the model's progress in terms of the percentage of BN structure learned with respect to the number of data points.

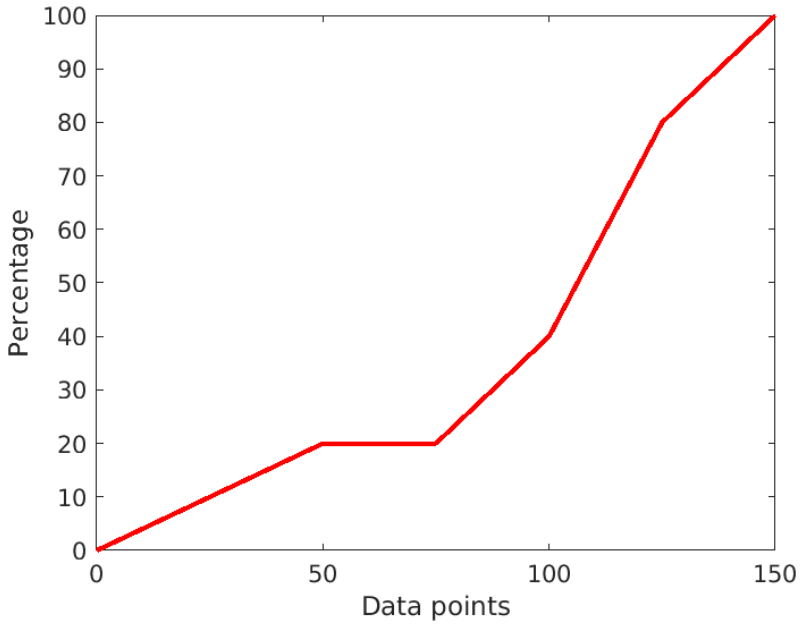


Figure 3-11 LBN progress with respect to data points

3.4.7 Testing of Model on Data Set – Clean and Corrupt Data

The learned BN testing has been performed on 20% of the data set, i.e., 90 data points. Firstly, the actual data were given as a clean testing data set, consisting of 85 cases representing no MIC likelihood in labelled data set. The other 5 represent the likelihood of MIC. It can be observed that the results are shown in Table 3-8 in the form of a 4 by 4 confusion matrix representing the actual (testing dataset) and predicted values from the LBN model. This result is also expected, since the training and testing data set comes from the same dataset. However, data obtained from the process often consists of error, which could be due to faulty sensors or logic-solver errors. Therefore, the testing data set has been modified to test the model's efficacy. A 20% error has been introduced in the clean testing data set and referred to as corrupt data. A total of 18 data points have been modified in labelled data set: 5 (yes) points are replaced with no and 13 (no) with yes. The result is shown in Table 3-9 the model predicted all 85 cases of no MIC likelihood (actual 72 and 13 false negative) and 5 instances of MIC likelihood (false positive), correctly.

Table 3-8 Confusion matrix with clean testing data

Actual (MIC likelihood based on input data)		Predicted (MIC likelihood from BN model)	
		No	Yes
	No	85	0
	Yes	0	5

Table 3-9 Confusion matrix with 80% clean and 20% corrupt data

Actual (MIC likelihood based on input data)		Predicted (MIC likelihood based on input data)	
		No	Yes
	No	72	5
	Yes	13	0

3.5 Validation of the LBN Model

The aim is to conduct validation based on MIC-induced failures reported in (Skovhus et al., 2017). A total of 6 scenarios were generated; scenarios 1-4 were based on the failures, whereas scenarios 5 and 6 correspond to lower and upper bound of the LBN model by considering the presence and absence of all parent nodes to show model variability. The LBN model developed for the FPSO platform is used for validation and comparison to benchmark model performance with the results reported in Kannan et al.(Kannan et al., 2020). Evidence of each scenario is reported in

Table 3-10 as per understanding and relevance of MIC process that are useful for LBN model.

Table 3-10 Scenarios 1-6 evidence

Scenario	Identified parameter

1	Biofilm, Flow rate, Sulfate-reducing microorganism, Methanogens and Iron-reducing microorganism
2	Temperature, Flow rate, pH, Sulfate-reducing microorganism and Iron-reducing microorganism
3	pH, Temperature, Flow rate, Sulfate-reducing microorganism and Biofilm
4	pH, Temperature, Flow rate, Iron-reducing microorganism and Biofilm
5	All parent nodes are present
6	All parent nodes are absent

The results of each scenario were summarized in Table 3-11. Scenario 1 was based upon the failure reported due to MIC in an outlet of a high-pressure production trap at the gas-oil separation plant. The failure occurred due to under-deposit localized pitting where the pit morphology and shape were similar to those caused by MIC. The LBN model resulted in an 80% likelihood compared to 58% reported in Kannan et al.(Kannan et al., 2020). Scenario 2 was based on failure and leaks of heat exchangers in different process units due to localized pitting on the tube side, due to MIC. The failure occurred in a cooling system within a crude oil refinery. The model resulted in a 55% likelihood of MIC, based upon the observables obtained, compared to 47% from Table 3-11. Scenario 3 describes fire hydrants premature failure due to Stress Corrosion Cracking (SCC) and leaching, which was accelerated due to microbes. The LBN model estimated a 70% likelihood of MIC, which denotes the microbiological activity that accelerated SCC compared to 50% reported in Kannan et al.(Kannan et al., 2020). Scenario 4 consists of a pinhole leak observed on a diesel shipping line due to localized pitting, which probably resulted from MIC. The model resulted in a 79% likelihood of MIC compared to 54% reported by Kannan et al. Scenario 5 was simulated based

on the parent nodes presence. The model estimated a 98% likelihood of MIC. However, in scenario 6, all the parent nodes were absent, and the model estimated a 0.22% likelihood of MIC. Scenarios 5 and 6 show the variability of LBN model (upper bound and lower bound values).

Table 3-11 Comparison of MIC scenario-based likelihood results (rounded percentage) compared to Kannan et al.(Kannan et al., 2020)

Scenarios	LBN model results (%)	Kannan et al. results (%)
1	80	57.59
2	55	46.93
3	70	50.24
4	79	53.92
5	98	100
6	0.22	0.05

The comparison shows that the LBN model can predict all the failures in scenarios 1-4. The purpose of comparison is to establish benchmarking which helps validate LBN model performance with the recent study conducted in MIC domain. This exercise also shows the LBN model capability and potential use in MIC domain.

The results show higher variability in predicting scenarios 1-4. However, Kannan et al. results have less variability in the scenarios likelihood estimation possibly due to mean estimation of child node. Their model also considers arc weightage from a parent node to a child node to estimate the mean value of child node. As reported in (Kannan et al., 2020), weighting the arc was conducted based on heuristic estimates. However, the present study gives higher arc weightage to microorganisms activity than favourable operating conditions since microorganisms activity resulted in MIC-induced failures.

3.6 Conclusions

Progressing towards process digitalization demands a data-driven dynamic approach to ensure safety of process systems. This chapter presents an integrated data-driven model that has the advantage of directly using the available field and laboratory data to assess MIC likelihood. LBN model is advantageous compared to other BN-based models in MIC domain due to its learning ability, missing values handling and acceptable performance demonstrated on model testing on the corrupt dataset. LBN model is also benchmarked and validated with the Kannan et al. study conducted in MIC domain to establish its effectiveness in predicting MIC likelihood. The data driven LBN model provides a unique platform for strengthening the variables' correlation and their features to assess MIC likelihood.

The data-driven quantitative analysis of MIC provides insight into determining vulnerable process equipment. As a result, the model can manage and control the MIC risk in industries and enhance the overall safety of the process operation. Future work can consider an incremental Bayesian learning model to improve its practicability on the existing system. Moreover, features from unstructured data can be used in combination with structured data to offer holistic process safety solutions.

3.7 Acknowledgements

The first author would like to thank Dr. Guozheng Song for helping with the GeNie software in the initial stage of the work. The authors acknowledge the financial support provided by Genome Canada and their supporting partners through the Large Scale Applied Research Project and the Canada Research Chair (CRC) Tier I Program in Offshore Safety and Risk Engineering.

3.8 References

1. K. Gerhardus, V. Jeff, N. Thopson, O. Moghissi, M. Gould, and J. Payer, “International Measures of Prevention , Application , and Economics of Corrosion

Technologies Study," *NACE Int.*, 2016.

2. K. B. Sorensen, U. S. Thomsen, S. Juhler, and J. Larsen, "Cost efficient MIC management system based on molecular microbiological methods," in *NACE - International Corrosion Conference Series*, 2012.
3. T. Liengen, D. Féron, R. Basséguy, and I. Beech, *Understanding Biocorrosion: Fundamentals and Applications*, vol. number 66. Elsevier Inc., 2014.
4. R. W. Revie, *Oil and Gas Pipelines: Integrity and Safety Handbook*. wiley, 2015.
5. J. Kaduková, E. Škvareková, V. Mikloš, and R. Marcinčáková, "Assessment of microbially influenced corrosion risk in slovak pipeline transmission network," *J. Fail. Anal. Prev.*, vol. 14, no. 2, pp. 191–196, 2014.
6. N. S. Tm, I. No, T. N. International, T. Nace, and C. Notice, "Standard Test Method TM0212 Detection , Testing , and Evaluation of Microbiologically Influenced Corrosion on Internal Surfaces of Pipelines," *NACE Int. Corros. Soc.*, 2012.
7. R. Sooknah, S. Papavinasam, and R. W. Revie, "Monitoring microbiologically influenced corrosion: A review of techniques," in *NACE - International Corrosion Conference Series*, 2007.
8. M. Taleb-Berrouane, F. Khan, R. B. Eckert, and T. L. Skovhus, "Predicting Sessile Microorganism Populations in Oil and Gas Gathering and Transmission Facilities-Preliminary Results," in *7th International Symposium on Applied Microbiology and Molecular Biology in Oil Systems (ISMOS 7)*, 2019.
9. T. L. Skovhus, D. Enning, and J. S. Lee, *Microbiologically influenced corrosion in the upstream oil and gas industry*, vol. 1. 2017.
10. B. Little, R. Ray, and R. Pope, "Relationship between corrosion and the biological sulfur cycle: A review," *Corrosion*, vol. 56, no. 4, p. 433, 2000.

11. S. Hashemi, N. Bak, F. Khan, K. Hawboldt, L. Lefsrud, and J. Wolodko, “Bibliometric Analysis of Microbiologically Influenced Corrosion (MIC) of Oil and Gas Engineering Systems,” *Corrosion*, vol. 74, no. 4, pp. 468–486, 2018.
12. R. Cord-Ruwisch, W. Kleinitz, and F. Widdel, “SULFATE-REDUCING BACTERIA AND THEIR ACTIVITIES IN OIL PRODUCTION.,” *JPT, J. Pet. Technol.*, 1987.
13. R. Liang, R. S. Grizzle, K. E. Duncan, M. J. McInerney, and J. M. Suflita, “Roles of thermophilic thiosulfate-reducing bacteria and methanogenic archaea in the biocorrosion of oil pipelines,” *Front. Microbiol.*, 2014.
14. S. Lahme *et al.*, “Metabolites of an oil field sulfideoxidizing, nitrate-reducing Sulfurimonas sp. cause severe corrosion,” *Appl. Environ. Microbiol.*, 2019.
15. T. Gu, “Theoretical Modeling of the Possibility of Acid Producing Bacteria Causing Fast Pitting Biocorrosion,” *J. Microb. Biochem. Technol.*, 2014.
16. R. I. Ray, J. S. Lee, and B. J. Little, “Iron-oxidizing bacteria: A review of corrosion mechanisms in fresh water and marine environments,” in *NACE - International Corrosion Conference Series*, 2010.
17. E. Valencia-Cantero and J. J. Peña-Cabriales, “Effects of iron-reducing bacteria on carbon steel corrosion induced by thermophilic sulfate-reducing consortia,” *J. Microbiol. Biotechnol.*, 2014.
18. A. Vigneron, E. B. Alsop, B. Chambers, B. P. Lomans, I. M. Head, and N. Tsesmetzis, “Complementary Microorganisms in Highly Corrosive Biofilms from an Offshore Oil Production Facility,” *Appl. Environ. Microbiol.*, vol. 82, no. 8, pp. 2545 LP – 2554, Apr. 2016.
19. D. S. Nicoletti, “Microbial Nitrogen and Sulfur Metabolism and its Relation to Corrosion Risk on Offshore Oil Production Platforms,” PRISM, 2020.

20. D. Enning *et al.*, “Marine sulfate-reducing bacteria cause serious corrosion of iron under electroconductive biogenic mineral crust,” *Environ. Microbiol.*, 2012.
21. D. Enning and J. Garrelfs, “Corrosion of iron by sulfate-reducing bacteria: New views of an old problem,” *Applied and Environmental Microbiology*. 2014.
22. N. Kip *et al.*, “Methanogens predominate in natural corrosion protective layers on metal sheet piles,” *Sci. Rep.*, vol. 7, no. 1, p. 11899, 2017.
23. T. J. Verbeke *et al.*, “Thermoanaerobacter thermohydrosulfuricus WC1 shows protein complement stability during fermentation of key lignocellulose-derived substrates,” *Appl. Environ. Microbiol.*, 2014.
24. T. Peng, S. Pan, L. P. Christopher, R. Sparling, and D. B. Levin, “Growth and metabolic profiling of the novel thermophilic bacterium Thermoanaerobacter sp. strain YS13,” *Can. J. Microbiol.*, vol. 62, no. 9, pp. 762–771, May 2016.
25. M. Taleb Berrouane, “Dynamic corrosion risk assessment in the oil and gas production and processing facility.” Memorial University of Newfoundland, 2020.
26. A.-W. Dawuda, M. Taleb-berrouane, and F. Khan, “A probabilistic model to estimate microbiologically influenced corrosion rate,” *Process Saf. Environ. Prot.*, 2021.
27. M. E. A. Ben Seghier, B. Keshtegar, M. Taleb-Berrouane, R. Abbassi, and N. T. Trung, “Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines,” *Process Saf. Environ. Prot.*, vol. 147, no. January, pp. 818–833, 2021.
28. B. F. Pots *et al.*, “Improvements on de waard-milliams corrosion prediction and applications to corrosion management,” in *NACE - International Corrosion Conference Series*, 2002, no. 02235, p. 19.
29. Maxwell; Campbell, S. Maxwell, and S. Campbell, “Monitoring the mitigation of

MIC risk in pipelines," in *NACE - International Corrosion Conference Series*, 2006, no. 244, pp. 1–10.

30. R. Sooknah *et al.*, "Modelling the occurrence of microbiologically influenced corrosion," in *NACE - International Corrosion Conference Series*, 2007, no. 07515, pp. 1–12.
31. M. Urquidi-Macdonald, A. Tewari, and L. F. Ayala H, "A neuro-fuzzy knowledge-based model for the risk assessment of microbiologically influenced corrosion in crude oil pipelines," *Corrosion*, 2014.
32. Y. Wang and L. Jain, "MIC assessment model for upstream production and transport facilities," in *NACE - International Corrosion Conference Series*, 2016.
33. G. Liu, J. Zhang, F. Ayello, and P. Stephens, "The application of Bayesian network threat model for corrosion assessment of pipeline in design stage," in *Proceedings of the Biennial International Pipeline Conference, IPC*, 2018.
34. M. Taleb-Berrouane, F. Khan, K. Hawboldt, R. Eckert, and T. L. Skovhus, "Model for microbiologically influenced corrosion potential assessment for the oil and gas industry," *Corros. Eng. Sci. Technol.*, vol. 53, no. 5, pp. 378–392, 2018.
35. E. S. Andersen, T. L. Skovhus, and E. Hillier, "Review of current models for MIC management," in *Microbiologically Influenced Corrosion in the Upstream Oil and Gas Industry*, 2017.
36. J. Wolodko *et al.*, "Modeling of Microbiologically Influenced Corrosion (MIC) in the oil and gas industry - Past, present and future," *NACE - Int. Corros. Conf. Ser.*, vol. 2018-April, 2018.
37. T. L. Skovhus, L. Holmkvist, K. Andersen, H. Pedersen, and J. Larsen, "MIC risk assessment of the halfdan oil export spool," in *Society of Petroleum Engineers - SPE International Conference and Exhibition on Oilfield Corrosion 2012*, 2012.

38. J. Larsen, T. L. Skovhus, A. M. Saunders, B. Højris, and M. Agerbák, “Molecular identification of MIC bacteria from scale and produced water: Similarities and differences,” in *NACE - International Corrosion Conference Series*, 2008.
39. J. Larsen, S. Juhler, K. B. Sørensen, and D. S. Pedersen, “The application of molecular microbiological methods for early warning of MIC in pipelines,” in *NACE - International Corrosion Conference Series*, 2013.
40. N. Ezenwa, F. Khan, K. Hawboldt, R. Eckert, and T. L. Skovhus, “A preliminary molecular simulation study on the use of HS- as a parameter to assess the effect of surface deposits on the SRB-initiated pitting on metal surfaces,” in *CORROSION 2019*, 2019.
41. M. Taleb-berrrouane and F. Khan, “Development of MIC Risk Index for Oil and Gas Operations,” in *C-RISE & geno-MIC Workshop & Symposium*, 2018.
42. R. Yang, F. Khan, M. Taleb-Berrouane, and D. Kong, “A time-dependent probabilistic model for fire accident analysis,” *Fire Saf. J.*, vol. 111, no. December 2018, p. 102891, 2020.
43. S. Kabir, M. Taleb-Berrouane, and Y. Papadopoulos, “Dynamic reliability assessment of flare systems by combining fault tree analysis and Bayesian networks,” *Energy Sources, Part A Recover. Util. Environ. Eff.*, 2019.
44. S. M. Deyab, M. Taleb-berrouane, F. Khan, and M. Yang, “Failure analysis of the offshore process component considering causation dependence,” *Process Saf. Environ. Prot.*, 2018.
45. M. Taleb-Berrouane, A. Sterrahmane, D. Mehdaoui, Z. Lounis., and Z. Lounis, *Emergency Response Plan Assessment Using Bayesian Belief Networks*. St John’s NL, 2017, pp. 1–6.
46. M. Z. Kamil, F. Khan, G. Song, and S. Ahmed, “Dynamic Risk Analysis Using

Imprecise and Incomplete Information," *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng.*, vol. 5, no. 4, 2019.

47. M. Taleb-Berrouane and F. Khan, "Dynamic resilience modelling of process systems," *Chem. Eng. Trans.*, vol. 77, no. 1, pp. 313–318, 2019.
48. M. Bougofa, M. Taleb-Berrouane, A. Bouafia, A. Baziz, R. Kharzi, and A. Bellaouar, "Dynamic Availability Analysis Using Dynamic Bayesian and Evidential Networks," *Process Saf. Environ. Prot.*, 2021.
49. M. Z. Kamil, M. Taleb-Berrouane, F. Khan, and S. Ahmed, "Dynamic domino effect risk assessment using Petri-nets," *Process Saf. Environ. Prot.*, vol. 124, 2019.
50. M. Taleb-Berrouane, F. Khan, and M. Z. M. Z. Kamil, "Dynamic RAMS analysis using advanced probabilistic approach," *Chem. Eng. Trans.*, vol. 77, pp. 241–246, 2019.
51. M. Talebberrouane, F. Khan, and Z. Lounis, "Availability analysis of safety critical systems using advanced fault tree and stochastic Petri net formalisms," *J. Loss Prev. Process Ind.*, vol. 44, pp. 193–203, 2016.
52. M. Taleb-Berrouane, F. Khan, and P. Amyotte, "Bayesian Stochastic Petri Nets (BSPN) - A new modelling tool for dynamic safety and reliability analysis," *Reliab. Eng. Syst. Saf.*, vol. 193, 2020.
53. M. Taleb-berrouane, S. Imtiaz, and F. Khan, "Internal Corrosion Monitoring in the Crude Oil Pipelines," in *20th Annual Aldrich Conference*, 2018, no. March.
54. F. Ayello, S. Jain, N. Sridhar, and G. H. Koch, "Quantitive assessment of corrosion probability - A Bayesian network approach," *Corrosion*, 2014.
55. G. Koch, F. Ayello, V. Khare, N. Sridhar, and A. Moosavi, "Corrosion threat assessment of crude oil flow lines using bayesian network model," *Corros. Eng.*

Sci. Technol., 2015.

56. O. Shabarchin and S. Tesfamariam, "Internal corrosion hazard assessment of oil & gas pipelines using Bayesian belief network model," *J. Loss Prev. Process Ind.*, 2016.
57. P. Kannan, S. P. Kotu, H. Pasman, S. Vaddiraju, A. Jayaraman, and M. S. Mannan, "A systems-based approach for modeling of microbiologically influenced corrosion implemented using static and dynamic Bayesian networks," *J. Loss Prev. Process Ind.*, 2020.
58. M. Taleb-Berrouane, F. Khan, and K. Hawboldt, "Corrosion risk assessment using adaptive bow-tie (ABT) analysis," *Reliab. Eng. Syst. Saf.*, vol. 214, no. May, p. 107731, 2021.
59. J. Hegde and B. Rokseth, "Applications of machine learning methods for engineering risk assessment – A review," *Safety Science*. 2020.
60. Y. M. Goh, C. U. Ubeynarayana, K. L. X. Wong, and B. H. W. Guo, "Factors influencing unsafe behaviors: A supervised learning approach," *Accid. Anal. Prev.*, 2018.
61. J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, 1996.
62. Y. Ma, M. Chowdhury, A. Sadek, and M. Jeihani, "Real-time highway traffic condition assessment framework using vehicleInfrastructure integration (VII) with artificial intelligence (AI)," *IEEE Trans. Intell. Transp. Syst.*, 2009.
63. S. A. Adedigba, F. Khan, and M. Yang, "Dynamic failure analysis of process systems using principal component analysis and Bayesian network," *Ind. Eng. Chem. Res.*, 2017.
64. F. V. Jensen and T. D. Nielsen, *Bayesian networks and decision graphs*

(Information Science and Statistics). 2007.

65. P. Larrañaga, H. Karshenas, C. Bielza, and R. Santana, "A review on evolutionary algorithms in Bayesian network learning and inference tasks," *Inf. Sci. (Ny).*, 2013.
66. N. Khakzad, F. Khan, and P. Amyotte, "Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network," *Process Saf. Environ. Prot.*, vol. 91, no. 1–2, pp. 46–53, 2013.
67. S. A. Adedigba, O. Oloruntobi, F. Khan, and S. Butt, "Data-driven dynamic risk analysis of offshore drilling operations," *J. Pet. Sci. Eng.*, 2018.
68. R. E. Neapolitan, *Learning Bayesian networks*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2004.
69. D. Dash and M. J. Druzdzel, "A Hybrid Anytime Algorithm for the Construction of Causal Models From Sparse Data," *Artif. Intell.*, 1999.
70. G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Mach. Learn.*, 1992.
71. D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Mach. Learn.*, 1995.
72. A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum Likelihood from Incomplete Data Via the EM Algorithm ," *J. R. Stat. Soc. Ser. B*, 1977.
73. S. A. Imtiaz and S. L. Shah, "Treatment of missing values in process data analysis," *Can. J. Chem. Eng.*, 2008.
74. "GeNie software accessed from https://www.bayesfusion.com." 2020. Accessed on January 2020.
75. NACE RP0775, "Recommended Practice Preparation , Installation , Analysis , and Interpretation of Corrosion Coupons in Oilfield Operations," *Nace Int. Houston, TX. USA,* no. 21017, 2005.

76. S. Xu *et al.*, “Data cleaning in the process industries,” *Rev. Chem. Eng.*, 2015.

4 Textual Data Transformations Using Natural Language Processing for Risk Assessment

Preface

This chapter has been published in the ***Risk Analysis*** Journal. I am the primary author of this manuscript, along with co-authors Drs. Mohammed Taleb-Berrouane, Faisal Khan, Paul Amyotte and Salim Ahmed. I developed the framework for objective risk assessment from textual data and its application in developing the models. I prepared the first draft of the manuscript and revised it based on the co-authors' and peer review feedback. The co-author Dr. Mohammed Taleb-Berrouane assisted in model development, testing and revision based on peer review feedback. The co-author Dr. Faisal Khan proposed the conceptual framework and helped develop the framework, testing and revising the model. The co-authors, Drs. Paul Amyotte and Salim Ahmed provided constructive feedback to improve the readability, review and revision based on peer review feedback and finalizing the manuscript.

Reference: Kamil, M. Z., Taleb-Berrouane, M., Khan, F., Amyotte, P., & Ahmed, S. (2023). Textual data transformations using natural language processing for risk assessment. *Risk analysis*.

Abstract

Underlying information about failure, including observations made in free text, can be a good source for understanding, analyzing, and extracting meaningful information for determining causation. The unstructured nature of natural language expression demands advanced methodology to identify its underlying features. There is no available solution to utilize unstructured data for risk assessment purposes. Due to the scarcity of relevant data, textual data can be a vital learning source for developing a risk assessment methodology. This work addresses the knowledge gap in extracting relevant features from textual data to develop cause-effect scenarios with minimal manual interpretation. This study applies natural language

processing (NLP) and text-mining techniques to extract features from past accident reports. The extracted features are transformed into parametric form with the help of fuzzy set theory and utilized in Bayesian networks (BN) as prior probabilities for risk assessment. An application of the proposed methodology is shown in microbiologically influenced corrosion (MIC)-related incident reports available from the Pipeline and Hazardous Material Safety Administration (PHMSA) database. In addition, the trained named entity recognition model (NER) is verified on eight incidents, showing a promising preliminary result for identifying all relevant features from textual data and demonstrating the robustness and applicability of NER method. The proposed methodology can be used in domain-specific risk assessment to analyze, predict and prevent future mishaps, ameliorating overall process safety.

KEYWORDS: Natural language processing (NLP), text mining, unstructured, data, Bayesian network (BN), microbiologically influenced corrosion (MIC), named entity recognition (NER), risk assessment, process safety

4.1 Introduction

Natural language processing (NLP) is concerned with human-computer interaction, including computational methods for automated analysis (Cambria & White, 2014). In order words, NLP is a field of interpreting and understanding human text and speech (Clark et al., 2010). It has a wide range of applications such as sentiment analysis, information extraction, text classification, question answering, speech recognition, machine translation, keyword searching and advertisement matching. Applications of NLP have been widely reported across domains based on two broad classifications: (i) hand-written rules or ontology-based studies and (ii) machine learning algorithms. (Wu et al., 2013) proposed a scenario object model based on domain ontology to capture and effectively utilize information from HAZOP studies. NLP combined with ontology was used to automate the procedure of information extraction. An ontology was developed and compared with and without domain-specific ontology and

concluded that domain-specific ontology exhibits robust results (Kwon et al., 2013). (Guo & Huang, 2016) proposed a semi-automated ontology that utilizes web mining, machine learning and NLP to assist the user in building a domain-specific ontology without much effort. A pre-defined list of keywords by experts was used for automated content analysis for construction safety (Tixier et al., 2016b). A hazard modelling language was proposed to provide a structured pathway for formalizing natural language hazard descriptions in a safety-critical system, i.e., Passenger train (Zhou et al., 2017). Another study proposed by Nakata uses a text mining approach to extract information using bag-of-words from two adjacent sentences. However, this study ignored order within bag-of-words, which will be helpful in constructing the causality of aviation incidents (Nakata, 2017).

A domain-specific ontology was developed, which employed NLP to extract subject, predicate, and object from unstructured textual data to improve human communication in aviation (Abdullah et al., 2019). (Hughes et al., 2019) developed an ontology-based approach capable of using multiple languages (German, French or Italian) to identify safety incidents on railways, such as falling of passengers and being stuck by doors. A framework consisting of ontology and NLP was proposed to automate literature knowledge from abstract instead of bibliometric analysis, which is only limited to critical phrases such as authors, publications, journals, and citations. Bidirectional Encoder Representations from Transformers (BERT) were used to facilitate NLP tasks in the study (Chen & Luo, 2019b). (Aziz et al., 2019) proposed a pathway for conducting causality analysis from an undesired event using an ontology-based approach to construct a multi-entity Bayesian network. This study's advantage is to perform risk estimation by evaluating potential hazards based on operational and environmental conditions. A new method is proposed to extract dependencies using NLP techniques with an ontology-based approach (Deshpande et al., 2020). It shows that the use of NLP in dependency extraction is also feasible. Another approach was shown to extract

information from a chemical accident database using NLP techniques. The authors also suggest that the standard named entity recognition (NER) method cannot extract information from chemical accidents (Single et al., 2020).

Apart from hand-written or ontology-based approaches, machine learning-based approaches are exploited for employing NLP for automated text analysis. (Tixier et al., 2016a) proposed using random forest and stochastic gradient tree-boosting machine learning techniques to predict construction injuries from construction safety reports. (Tanguy et al., 2016) uses a supervised learning method by employing a linear support vector machine (SVM). This study transforms textual data into numerical features and uses SVM for classification purposes of aviation reports. A recent study (Li et al., 2021) analyzed chemical accidents of natural gas pipeline incidents to extract spatial and temporal correlations between natural gas pipeline accident severity and contributory factors. The study's outcome suggests that human-related contributory factors have high incident severity. In addition to supervised learning (Ben et al., 2021), unsupervised machine learning is also used to categorize primary causal factors using a latent semantic analysis approach to aviation narratives. However, this study demands an extensive data set for increased accuracy (Robinson et al., 2015). (Chokor et al., 2016) investigated the k-means clustering approach in rearranging OSHA construction accident reports based on accident types. (Liu et al., 2021) uses two methods for feature extraction from textual data, such as k-means clustering and co-occurrence matrix. The latter approach is advantageous and identifies contributory factors and causality. However, this approach has the demerit of omitting important features from incident data. In addition to supervised and unsupervised machine learning techniques for NLP, a semi-supervised approach was recently introduced by (Ahadh et al., 2021) that can label unstructured data and require less intervention to apply in other domains. The approach is validated on aviation and pipeline incident data.

Structured data from operational parameters and laboratory testing have been used to understand, analyze and convert into the MIC likelihood model (Kamil et al., 2021). However, unstructured data are available in free text, consisting of underlying causes and contributory factors not considered for risk assessment models. A database of incidents is needed to assess the unstructured nature of human language and will act as a data source for risk assessment. Corrosion-related incidents from the database of (Pipeline and Hazardous Materials Safety Administration, 2022) are selected to identify underlying causal factors for pipeline incidents. Corrosion is a challenging and ubiquitous issue that has significantly affected oil and gas industries and poses an economic challenge. One of the severe forms of corrosion is Microbiologically Influenced Corrosion (MIC) which is a phenomenon that involves microorganisms' presence/activity resulting in a corrosive environment that affects the metal's surface (B. J. Little & Lee, 2014). Microbes play a vital role in initiating or accelerating the corrosion process by altering electrochemical conditions at metal's surface (Salgar-Chaparro et al., 2020; Videla & Herrera, 2005). MIC occurrence in pipeline and storage tanks poses a risk of leakage of hydrocarbons, resulting in harm to people, property and the environment (Kannan et al., 2020). There has been a significant increase in the transport of natural gas and crude oil through pipelines in the United States (Allison et al., 2020). Based on data from a hazardous liquid database maintained by the database of Pipeline and Hazardous Material Safety Administration (PHMSA), (Halim et al., 2020) reported that since 2010 there has been a consistent number of pipeline incidents each year; from the reported incidents, corrosion is the second-highest causal factor after equipment failure. Based on public data available at PHMSA (Stover, 2013) reported that since 1986 there was a total spill of 76,000 barrels of crude oil/petroleum products due to pipeline incidents. It shows a dire need to improve safer oil and gas operations through pipelines. Therefore, it is essential to understand causal factors behind incidents in pipeline operations and enable timely actions to prevent such incidents

(Halim et al., 2018). In several studies (Bersani et al., 2010; Bubbico, 2018; Halim et al., 2020), the PHMSA database was used to identify direct causes and contributions based on pre-defined categories of equipment failure, corrosion, natural force damage and others. However, except (Liu et al., 2021), none has identified underlying causal factors beyond pre-defined labels. Natural language processing (NLP) and text mining technique were recently used to identify underlying causes and extract valuable information from unstructured data (Liu et al., 2021; Zhang et al., 2020). Therefore, it will be worthwhile to explore employing NLP methods for risk assessment using textual data.

The approaches discussed leverage NLP to extract relevant information from textual data. Earlier methods were mainly focused on the classification of incident data based on extracted features. A recent study (Ahadh et al., 2021) proposes a semi-supervised keyword extraction method that eliminates the need for a labelled data corpus for training. Nonetheless, it does not provide any potential pathway for utilizing extracted features to improve the safety of a process operation by transforming qualitative features into quantitative reasoning. This chapter introduces a unique approach to using unstructured data (i.e., textual data) as a source to perform an objective risk assessment. NER uses a machine learning method coupled with a domain-specific corpus for automated feature extraction to identify underlying cause-effect scenarios from textual data. The training corpus contains entities with labels for automatic feature extracting and defining causation with minimal manual interpretation. NER method has the advantages of easy implementation, incremental annotation and improving the existing model to capture more entities under each label. In order words, it helps to correct an incorrect entity or add more data when necessary. Furthermore, a methodology is developed to provide a unique risk assessment pathway using textual data by transforming qualitative features into numerical reasoning by employing fuzzy set theory coupled with Bayesian network (BN) to predict objective risk.

The novelty of the proposed method is to leverage the domain corpus into an objective risk assessment approach. Unlike previous studies that employed NLP techniques (Aziz et al., 2019; Liu et al., 2021; Feng et al., 2021), this work utilizes a simple and unique approach. It consists of a custom NER method of identifying causal-effect storyline using five labels, namely, “causal factor”, “scenario”, “consequences”, “emergency response”, and “caution”. The label “caution” helps determine if the entities could be misleading from the fact and determines the abnormality's existence. In cases where abnormalities exist, are determined by the label “causal factor”. Therefore, it depicts NER's ability to determine the cause-effect relationship, which can be further improvised as the need arises. The second point is the application of fuzzy logic in translating features from NER model into a fuzzy probability. The final step is processing numerical data into risk assessment with the help of BN model. The proposed methodology is novel in extracting and transforming unstructured data into structured data that can feed into a probabilistic model such as BN for risk analysis. The method is simple to adopt and can be applied to other domains for risk assessment. Verification of NER model is performed to evaluate and benchmark its performance with the co-occurrence method (Liu et al., 2021). In addition, the proposed methodology is verified by evaluating its outcome with the actual conditions.

Section 4.2 is dedicated to this study's proposed methodology and details NER model. Section 4.3 of the study is devoted to applying the proposed methodology to five cases selected from the database of PHMSA, followed by NER model verification in section 4.4 based on a recent study identifying underlying causal factors. Thus, establishing its use in extracting causal-effect scenarios. Conclusions drawn from this study are mentioned in section 4.5.

4.2 Proposed methodology

The proposed methodology analyzes system states based on textual data found in past incident reports. The methodology illustrated in Figure 4-1 consists of four broad steps. Steps 4.2.1-

4.2.2 apply natural language processing and text mining techniques to extract underlying cause-effect storylines from a domain-specific corpus. Step 4.2.3 acts as a bridge to transform qualitative information into numerical reasoning. The last step helps to improve the subjective quantification of risk to objective risk assessment.

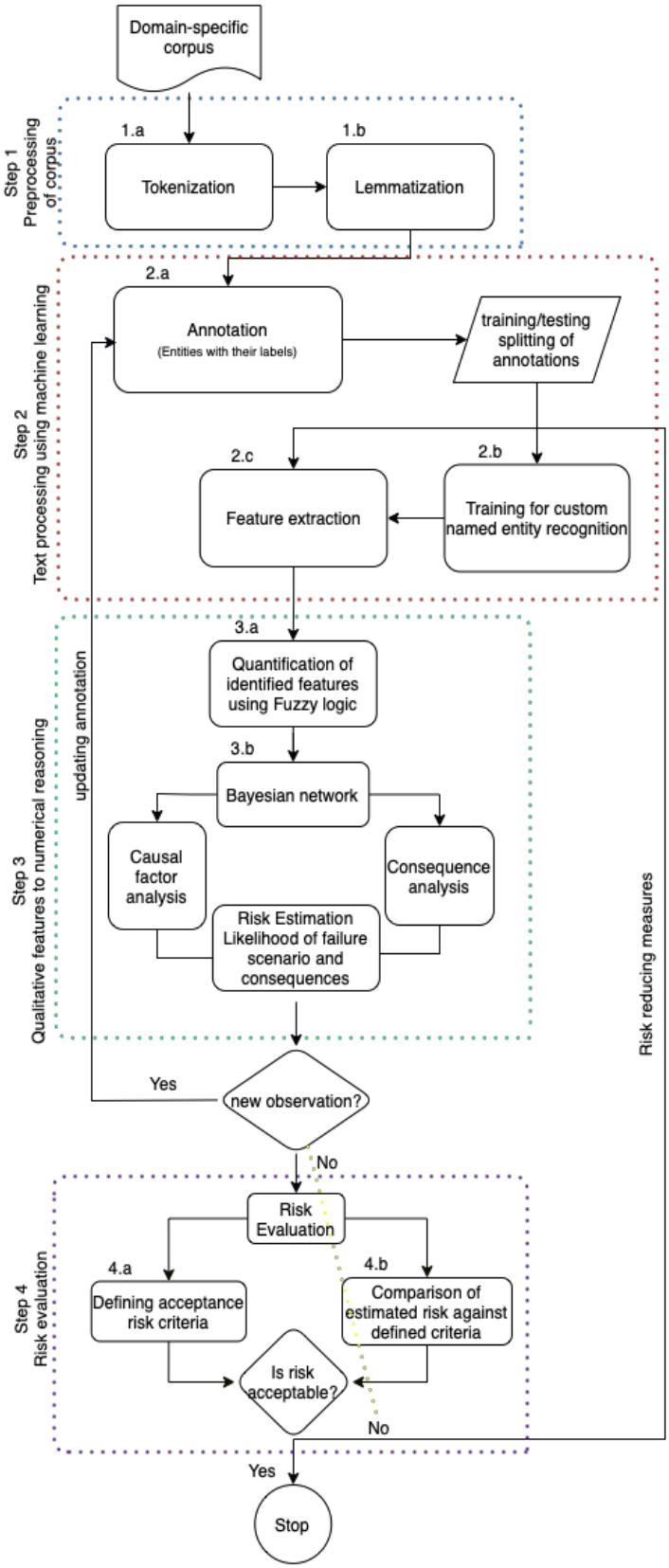


Figure 4-1 The proposed methodology for risk assessment from textual data

4.2.1 Preprocessing of Corpus

Preprocessing a corpus is essential for training and evaluation purposes to interpret the data by machine. In other words, data features can be easily parsed by machines. Preprocessing steps include tokenization of text followed by lemmatization, as described below:

4.2.1.1 Tokenization

The first step is to split a sentence into meaningful segments known as tokens (Honnibal & Montani, 2021c). The individual tokens will make up a string of text. For example, the following sentence is tokenized in 21 tokens as follows:

['John', 'identified', 'on', 'Monday', 'evening', 'at', '8:00', 'p.m.', 'temperature', 'was', 'very', 'cold', 'in', 'St.', 'Johns', 'and', 'decided', 'to', 'halt', 'the', 'operation']

4.2.1.2 Lemmatization

In linguistics, the word "lemma" refers to the representative form of the word, whereas stemming refers to removing common prefixes and suffixes from each word. Morphological analysis of the word differentiates the former from the latter. The present study considers lemmatization over stemming for converting each token (word) to its representative form due to attention to context (Liu et al., 2021) and linguistic accuracy (Toman et al., 2006) differentiates the former from the latter. For instance, in the following example, the word 'caring' is to 'car' by cutting off ‘ing' if stemming is performed (Ullah & Al Islam, 2019).

‘Caring’ > Lemmatization > care

‘Caring’ > Stemming > car

Preprocessing can be performed using spaCy library (Honnibal & Montani, 2021c). SpaCy an open-source library for advanced NLP tasks, including Part-of-speech tagging and NER. SpaCy uses neural network-based models for NLP tasks (Partalidou et al., 2019).

4.2.2 Text Processing using Machine Learning

Machine learning demands a labelled/annotated corpus for training NER model to predict the desired entities. In this study, the tagged corpus is a domain-specific corpus with pre-defined labels. The steps of machine learning of NER model are as follows:

4.2.2.1 Annotation

Extracting valuable information from unstructured data (human language) is a critical task that needs a large corpus for NLP. Annotation is the task of "tagging" or "labelling" a text corpus with a set of labels to the whole corpus or a part of it. This task can be performed manually or automatically based on how much data you want to annotate. This work performed the former method to tag text with labels representing interest semantics. In contrast, the latter case uses defined rules without human interference to annotate a corpus (Grosman et al., 2020). The prodigy tool can use active learning to annotate automatically. The curator, who manages the annotation process (known as curation), plays an essential role, in resolving inconsistencies among different annotators or the same annotator at different periods and annotating based on defined labels to ensure consistency and quality of an annotated corpus (Grosman et al., 2020; Ide & Pustejovsky, 2017). Annotation demands efforts to organize a sizeable domain-specific corpus that needs to be annotated and define categories of entities based on what you would like to extract from the corpus (Grosman et al., 2020). For instance, this work aims to extract a cause-effect storyline from textual data to assess the risk of failure scenario. The prodigy tool (Honnibal & Montani, 2021b) annotates machine learning model data. The ner.manual command (Honnibal & Montani, 2021b) highlights spans from textual data for different labels. Annotation of text is essential to label categories of entities desired to be extracted from textual data.

4.2.2.2 Training for Custom-Named Entity Recognition

Annotations performed in the previous step are exported in the spaCy library to train NER model from scratch to extract predefined entities from textual data. SpaCy uses a deep learning four-step method, “embed, encode, attend and predict.” Each token of a domain corpus or class of an NLP task, such as an entity class, is first converted into a unique integer (ID). Features like prefix, suffix, shape and form of work are used to extract hashed values reflecting word similarity. The model consists of hash values and their vectors at the embedding stage. The next step is to encode context; values pass through a convolutional neural network and encode with their context, resulting in a matrix-vector. Hence, each row represents the information of each token. The attending step matrix passes through the attention layer of a neural network to summarize the input with a query vector. The available class of entity is predicted at the prediction stage (Partalidou et al., 2019).

4.2.2.3 Feature Extraction

The trained NER model from the previous step extracts features from textual data. SpaCy library helps to evaluate the trained NER model and extract underlying features to illustrate causation using less manual interpretation. The entities have labels associated with them that do not demand extensive understanding.

4.2.3 Transform Qualitative Features into Numerical Reasoning

This step aims to use a method to quantify underlying information of causal-effect scenarios identified through the trained NER model.

4.2.3.1 Quantification of Identified Features Using Fuzzy Logic

Qualification of cause-effect storylines extracted from textual data needs to provide a bridge to transform the storyline into numerical reasoning. The challenge arises due to uncertainty and subjectivity in natural language. (Zadeh, 1965) introduced the concept of fuzzy set theory in his pioneering work; the argument was that probability theory alone is insufficient to represent all types of uncertainty. Fuzzy set theory is a well-suited and accepted method in safety and risk assessment to handle subjectivity and vagueness in linguistic variables (Ferdous, Khan, Sadiq, Amyotte, & Veitch, 2013). The fuzzy set theory consists of five tuples to define a linguistic variable for approximate reasoning (Zadeh, 1975). Therefore, this work utilizes the concept of fuzzy set theory to map linguistic grades used in natural language. A detailed example in the application section will help explain how fuzzy set theory will play an essential role in this study.

4.2.3.2 Bayesian Network

Bayesian network (BN) is an effective tool for reasoning under uncertainty (Deyab et al., 2018); (Taleb-Berrouane et al., 2017) (Taleb-Berrouane et al., 2018). The nodes in BN represent each variable; a directed arc from a parent node to a child node depicts conditional dependence, which is defined using a conditional probability table (CPT).

BN can capture conditional dependency to represent cause-effect relations of an incident (Bougofa et al., 2021; Yang et al.,, 2020) which is a significant advantage. The joint probability distribution P(B) of variables B= {B1, …., Bn} can be incorporated into BN as follows (Kamil et al., 2019; Khakzad et al.,, 2013):

$$P(B) = \prod_{i=1}^{n} P\left(B_i \middle| P_{y_{(B_i)}}\right), \quad (1)$$

where Py(Bi) is the parent of variable Bi

The identified causal factors, failure scenario and consequences can be modelled using BN model. Each entity can be represented as a node (i.e., variable) in BN model with a prior probability obtained from the previous step of using fuzzy set theory. Hence, it will provide a unique opportunity to model identified entities from textual data. The result from BN model will give a likelihood of a failure scenario and its consequences. In combination, likelihood and consequence parameters will evaluate risk associated with the scenario.

4.2.4 Risk Evaluation

It is necessary to evaluate the estimated risk from step 4.2.3 due to the subjective nature of natural language. This step aims to improve the subjective quantification of risk to its objective assessment. Risk evaluation consists of the following steps (Khan & Haddara, 2003):

4.2.4.1 Defining Risk Acceptance Criteria

This step requires defining the acceptance criteria to be used in a study. The acceptance criteria depend on the nature and type of system. Some of the commonly used criteria in the literature are ALARP (as low as reasonably practicable), Dutch acceptance criteria and US EPA (Environmental protection agency) acceptance criteria (Khan & Haddara, 2003). Characterization of the likelihood of an abnormal event and the severity of consequences depends on the type and nature of process activity.

4.2.4.2 Comparison of Estimated Risk Against Defined Criteria

This step applies acceptance risk criteria to the estimated risk to evaluate a system state. If a low-risk value does not exceed the acceptance criteria, the system is safe to operate; otherwise, it requires maintenance strategies to reduce risk.

4.3 Application of Methodology

A unique approach to leveraging unstructured textual data are needed to perform a risk assessment. This approach will provide an avenue to utilize observations recorded in human language, which is often neglected in risk assessment due to the unavailability of a technique to leverage them. This work introduces a way to use unstructured data as a source of data for risk assessment and enriches the identification of causation from past incidents by employing NER model.

4.3.1 Data Preparation for Custom NER Model

Data preparation includes preprocessing and labelling of a text corpus. A training corpus consists of reported information to the (Pipeline and Hazardous Materials Safety Administration, 2022), which shows the cause of the incident and provides a narrative description that will be used as a corpus for this study. For verification purposes, incident data consist of corrosion-related incidents and incidents shared in recent work (Liu et al., 2021). These incidents' descriptions are a good source of past recorded observations.

4.3.1.1 Preprocessing

Preprocessing of the corpus is conducted using the spaCy pre-trained model available (Honnibal & Montani, 2021c). Tokenization and lemmatization are performed to prepare the data for the annotation task. The textual data will be separated by each word, converted into its base form, and saved in a text file used to annotate the corpus.

4.3.1.2 Annotation and Training

This work proposed using a custom NER model to highlight features in terms of cause-effect storyline. Pre-trained models available from the spaCy library consist of a real-world object

such as a person, location, monetary value, time expression and quantity (Partalidou et al., 2019). However, this work demands a different set of entities; one possible way is to train a spaCy neural network for custom entities used in this study.

The ner.manual recipe is used to invoke the prodigy server to start annotation for different labels (Honnibal & Montani, 2021a). There are five labels assigned to the training corpus to extract the cause-effect storyline:

- Causal factors- Events (where abnormality exists) responsible for causing incidents
- Scenario - An incident that occurred due to an anomaly arising in a process operation results in a failure
- Consequences - possible outcomes associated with success/failure of safety barriers
- Emergency response (ER) - sudden action due to an unexpected incident to mitigate its impact
- Caution - phrases or words that are not causal factors but give more information about the abnormality

Once the corpus is annotated, it can train a convolutional neural network (CNN) for a custom NER model. The annotated corpus can be exported from the prodigy tool (Honnibal & Montani, 2021b) to spaCy library (Honnibal & Montani, 2021c). Training can be initiated using a command prompt on Windows or Terminal on Mac OS. After completion of training, spaCy saves the best and last model trained. The training steps can be found in (Honnibal & Montani, 2021c).

4.3.2 Automated Feature Extraction and Causation Construction

A short description of cases shown in Table 4-1 are taken from failure investigation reports (Pipeline and Hazardous Materials Safety Administration, 2022) to show the application of the proposed methodology. The identified entities from the incident description will form an

underlying cause-effect with less manual interpretation. NER model will extract causal factors that lead to failure scenarios and associated consequences.

Table 4-1 Selected description of cases narratives taken from the PHMSA database available in the public domain (Pipeline and Hazardous Materials Safety Administration, 2022) with NER model output

Case 1: Internal corrosion, possibly MIC	
Incident narrative	NER model output
At approximately 10:04 p.m. central standard time (CST) on April 8, 2012, operations personnel for Enterprise Crude Pipeline, LLC (Enterprise) discovered a leak on their 24-inch C75 line located in their Cushing West Terminal Facility located in Cushing, Oklahoma …Enterprise shut the line in… Though the pit wall contained some viable anaerobic bacteria… The internal pipe surface around the hole revealed it was the result of an internal corrosion pit that had grown through the pipe wall. The pit wall was covered with smaller pits. This indicated the pit grew under an occlusion such as a deposit or a biofilm. As corrosion-related bacteria were detected, there is a possibility that these bacteria entered the pipe after the pit formed. The presence of MIC bacteria, itself, is inconclusive as to the cause. The bacteria found in the test could have entered the pipe when the line was unpressurized or when the failed section was cut out in the ditch. Therefore, the test results for MIC could not adequately determine if MIC was a causal factor…Enterprise also indicated that vacuum	leak scenario viable/ relate/detect c aution anaerobic bacteria/bi ofilm/internal corrosi on/ MIC/pit causal fa ctors affected soil consequ ence excavate ER

trucks were en route to pick up free product and affected soil would be excavated.	
Case 2: Internal corrosion	
The well line was loaded at the time of failure but was not flowing gas. The valve at the tie into the field line was closed; the well gate was open. The failure apparently occurred 4 hours after the line was pressurized from ambient to 1720 psig. A local resident near the incident location reported the failure to Columbia Gas Transmission… The pipe ruptured due to internal corrosion pitting complicated by low impact toughness of the pipe material. The corrosion pitting was the result of sulfur and chloride containing compounds, and third party investigator speculated that the low point in the pipeline under creek retained free liquids. Future plans are to replace the entire well line in 2012. A means for liquid removal will be considered in the replacement project… There were no fatalities, injuries, or supply issues as a result of the incident.	failure scenario Sulfur/chloride/ inter nal corrosion pitting/ toughness causal fact ors Injury Consequence
Case 3: Internal corrosion, possibly MIC	
The East Bernard Compressor Station 17 has 3 pipelines (24”, 26”, and 30”) entering via a suction header and 3 pipelines (24”, 30”, and 30”) exiting via a discharge header. All systems operate as a single system for gas flow. The systems were flowing gas normally when the rupture occurred with no warning or abnormal situation occurring. The rupture occurred at approximately 4:25 pm on December 8, 2010. TGP’s Control Center took immediate actions	Rupture scenario Abnormal caution moisture/internal corr osion/microbiologica l causal factors Shut down conseque nce

to ESD Station 17 and isolate the piping associated with the rupture. The local station crew was still on site and secured the area. The discovery and isolation was prompt and operator's actions appear to be appropriate. The failure initiated from a section of dead leg piping established in 2000 when a 24" lateral was disconnected from the stations downstream header. There were obvious indications of residual moisture gathering in the dead leg, contributing to internal corrosion and a thinning of the pipe wall. The internal corrosion was caused by microbiological organisms present due to free moisture in the pipe. Evaluation of all other dead leg segments of pipe in the Station yard found no additional areas affected. TGP's Station Emergency Shut Down Device (ESD) activated immediately upon line failure and TGP's control Center isolated and shut in the 100 System from Wharton to Cleveland.	Initiate/isolate ER
Case 4: External corrosion	
The Fairfax West KCI pipeline operates intermittently to supply jet fuel to MCI. On the day of the failure, the pipeline had been shut down since 8:48 a.m. at a pressure of 227 psig. Throughout the day the pressure increased gradually as the temperature rose, eventually reaching approximately 235 psig. The Magellan control center was notified at 1:02 p.m. by contract workers from MVS Services, who were performing maintenance on a natural gas transmission pipeline operated by Southern Star, of an active leak on what appeared to be a 6-inch pipeline on the 7th Street Bridge…	failure scenario pressure causal factor increase caution temperature rise caus al factor shutdown consequen ce Initiate ER

Operations Control initiated the shutdown of all incoming and outgoing pipeline operations at the Magellan Kansas City facility.	
Case 5: External corrosion	
Visual examination of the leak showed corrosion pitting on the outside surface. Some of the pits had a fibrous appearance indicating preferential attack following the pipe axis. The fibrous appearance was characteristic of microbiologically influenced corrosion (MIC). Metallographic analysis of the corrosion pits showed undercutting and pits within pits, which are also unique characteristics of MIC... The leak was caused by MIC, which occurred after the coal tar coating had been degraded exposing the bare pipe. The presence of sulfur and moisture in the soil around the leak created an ideal environment for MIC to occur…Transco responded to the potential leak by shutting-in and isolating the pipeline valve segment.	Leak scenario coating caution sulfur/moisture/MIC causal factors shut/affected consequence isolate ER valve causal factor

The features extracted from cases are shown in Table 4-1. Hence, these features can be interpreted to extract a cause-effect storyline. Case 1 result shows causal factors, “anaerobic”, “bacteria” and “biofilm” are responsible for “internal corrosion” or possibly due to “MIC”. The possibility that “corrosion” “related” "bacteria" entered after the "pit" already formed due to unknown reasons resulted in "pit" outgrow, which resulted in the “leak” of crude oil. It results in crude oil supply “shut", "down” and “affect" nearby "soil” which refers to consequences associated with it. The “affected”, “soil” is “excavated” by emergency personnel. There are mainly four causal factors identified from incident description. The manual interpretation is needed to define how these factors depend on each other. For instance, biofilm provides an avenue for microorganisms to grow and deliver nutrition, causing MIC. An illustration of the

cause-effect entities relationship is presented in Figure 4-2. As shown in Figure 4-2, it requires less effort to interpret causation from textual data. Entities with their labels are self-explained; their role in depicting their relationship and manual interpretation is necessary to define dependency among the extracted features.

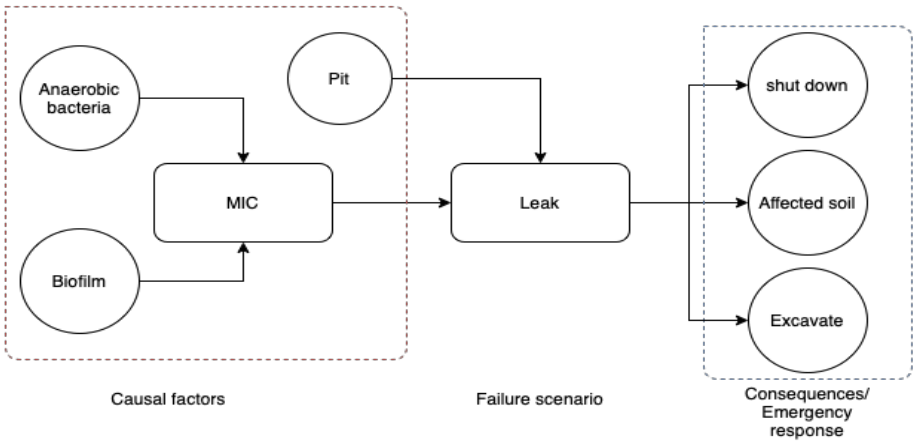


Figure 4-2 A cause-effect relationship from identified entities of case 1

Case 2 causation: a pipeline "rupture" occurred in a well line due to "internal corrosion pitting". It occurred due to "sulfur" and "chloride" containing compounds. Domain expert and practitioner knowledge will help better understand the context of the low "impact" "toughness" of pipeline material in the narration. The effect of internal corrosion pitting on the toughness of pipeline material shows dependency among entities. There was no “injury” reported due to this incident. As illustration depicted in Figure 4-3 shows induced causation from identified features.

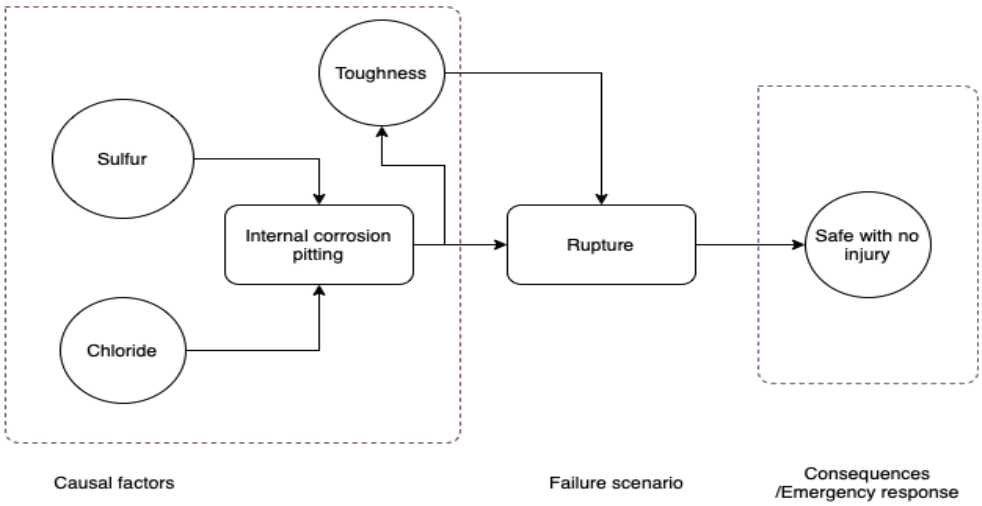


Figure 4-3 A cause-effect relationship from identified entities of case 2

Case 3 incident NER output suggests no "abnormal" situation occurred before "rupture" occurred. It has been found that "rupture" occurred due to residual "moisture" promoting the growth of "microbiological organisms" leading to "internal corrosion". Consequences and emergency response due to this incident include "initiation" of an emergency "shut down" system and isolation of the control center. The induced causation is shown in Figure 4-4, where moisture influences microbiological organism growth.

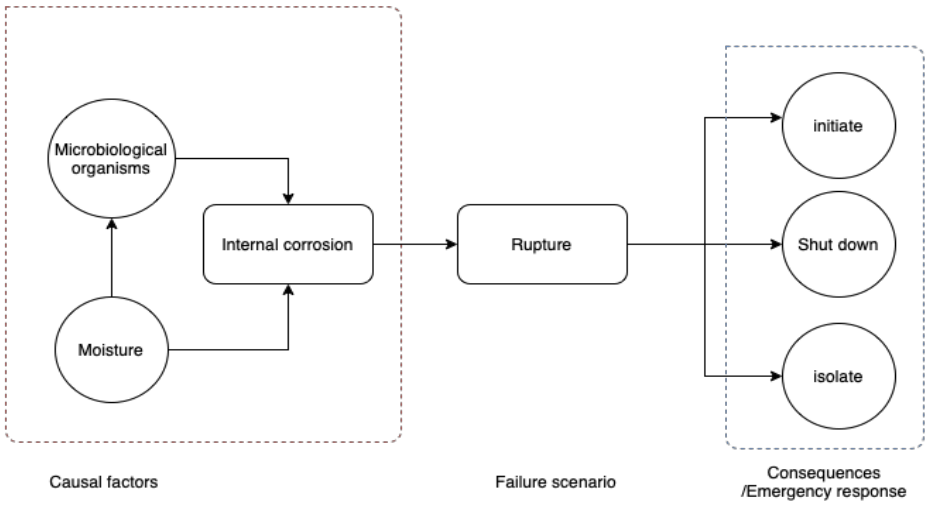


Figure 4-4 A cause-effect relationship from identified entities of case 3

Case 4 incident extracted features of an unfolded causation responsible for a "leak" that occurred due to an "increase" in "pressure" and "rise" in "temperature"; both operational parameters were responsible for causing failure. Emergency personnel were "notified" about the situation; they immediately "initiated" "shut down" of operations to minimize loss. This induced causation is depicted in Figure 4-5.

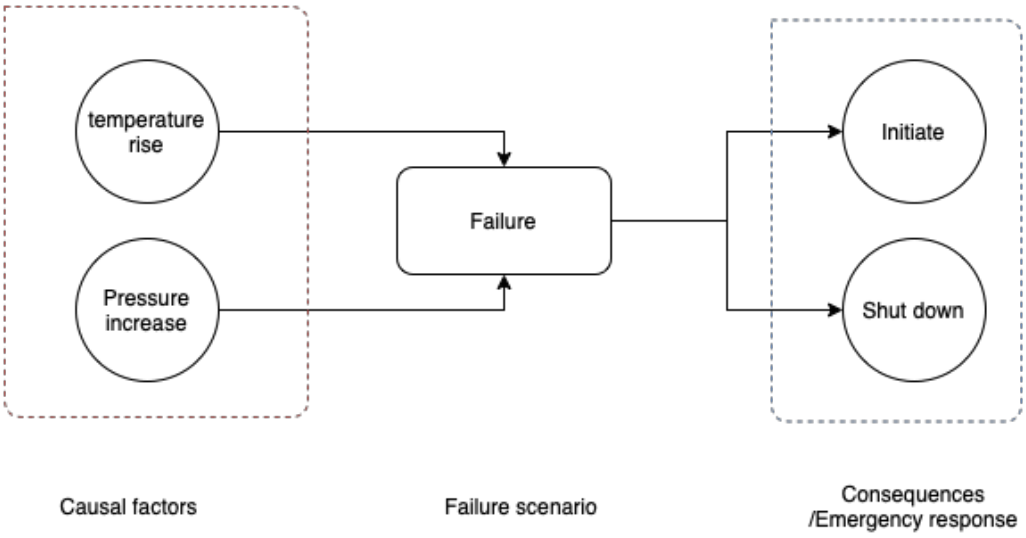


Figure 4-5 A cause-effect relationship from identified entities of case 4

Case 5 extracted entities that unfolded a "leak" scenario attributed to "MIC" and "coating" degradation from the pipe, exposing the metal surface. "MIC" is a result of "moisture" and "sulfur;" sulfate-reducing bacteria are proven to be responsible for MIC. Emergency response includes "shutting" and "isolating" affect pipeline "valve" segment. To better understand this causation, an illustration is shown in Figure 4-6.

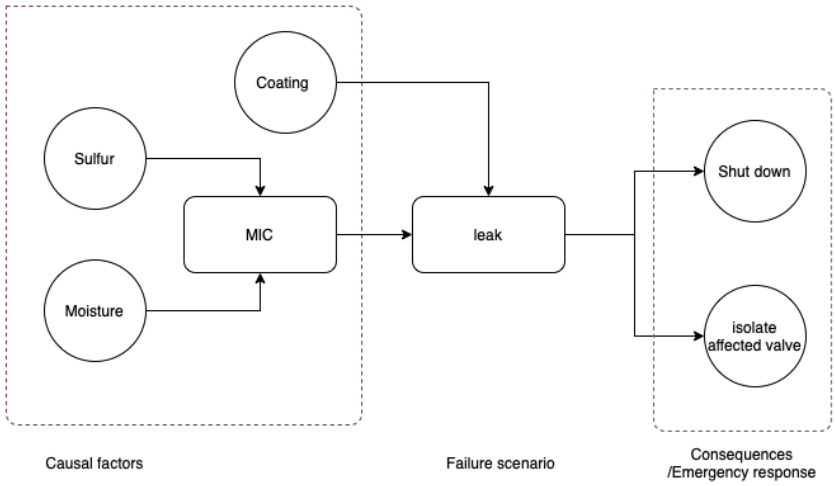


Figure 4-6 A cause-effect relationship from identified entities of case 5

There are five incident narratives selected from PHMHSA incident reporting ("Pipeline and Hazardous Materials Safety Administration," 2021) to show the robustness and applicability of NER model. It can extract all the relevant information and successfully unfold induced causations from textual data. Less manual interpretation is needed in most of the cases shown in Table 4-1 due to the labelling of entities. Labelling/annotating helps differentiate causal factors from failure scenarios, consequences, and emergency responses and eliminates the need to interpret labels manually. This is shown in the construction of the cause-effect relationship of identified features demonstrated in Figure 4-2 to Figure 4-6. Thus, NER model can automate feature extraction and narrow it down to cause-effect scenarios with less manual interpretation.

4.3.3 Transforming Qualitative Features to Quantitative Reasoning

4.3.3.1 Quantification of Identified Features Using Fuzzy Logic

The extracted causal factors from NER model will be used to transform features into quantitative information. The potential use of fuzzy logic can be seen from extracted features in Table 4-1. Human language conceptualization often consists of vagueness, as seen in extracted features. The identified entities such as viable bacteria, biofilm and the possibility of

pit presence before bacteria entered the pipeline are examples of entities from Figure 4-2 that can utilize the mathematics of fuzzy set theory introduced in the pioneering work by (Zadeh, 1975) to transform them into quantitative information. Fuzzy set theory is used for the following reasons:

1. Handling of vagueness in textual data
2. Quantification of subjective qualifications

To exploit the numerical relationship between an indefinite quantity (e.g., viable bacteria - basic event), the fuzzy set theory uses fuzzy numbers and membership functions ranging from 0 to 1. The selection of membership functions depends on available data and expert opinion. Triangular or trapezoidal fuzzy numbers (TFZ) are used in the present study due to their easy ability to model subjectivity and vagueness in identified entities. The identified entities from textual data are uncertain quantities, such as the likelihood of causal factors. The identified causal factors can be transformed into quantitative information. Figure 4-2 causal factor biofilm can be defined in terms of biofilm thickness. Since biofilm thickness is considered a vital engineering parameter related to biofilm growth rate ingress or egress of biofilm mass depends upon diffusion distance and biofouling in pipelines (Bakke & Olsson, 1986; Cunningham et al., 2012).

A linguistic variable is used to determine values in words or sentences. According to (Zadeh, 1975), a linguistic variable is a quintuple.–Linguistic grades and fuzzy numbers are used to describe a likelihood of an event. In the present study, we have used 9 linguistic grades with a scale of 7, as shown in (Chen et al., 1992; Zarei et al., 2019). Firstly, expert elicitation is performed using three different experts. In the aggregation of expert opinion, their weightage is considered based on (Zarei et al., 2019). Secondly, opinion aggregation is transformed into a fuzzy possibility (FPs) followed by fuzzy probability (FPr). The center-of-area method is

used for the defuzzification step (Sugeno & Kang, 1986). The defuzzification of the trapezoidal function is calculated using the following equation.

$$Y = \frac{1}{3} \times \frac{(a_4 + a_3)^2 - a_4 a_3 - (a_1 + a_2)^2 + a_1 a_2}{(a_4 + a_3 - a_1 - a_2)}$$

Lastly, fuzzy possibility from the aggregation is transformed into a fuzzy probability that can be directly used in BN using the function developed by (Onisawa, 1988).

$$FPr = \begin{cases} \frac{1}{10^K} \text{ if } FPs \neq 0 \\ 0 \text{ if } FPs = 0 \end{cases}$$

$$K = \left[\left(\frac{1 - FPs}{FPs} \right)^{\left(\frac{1}{3}\right)} \right] \times 2.301$$

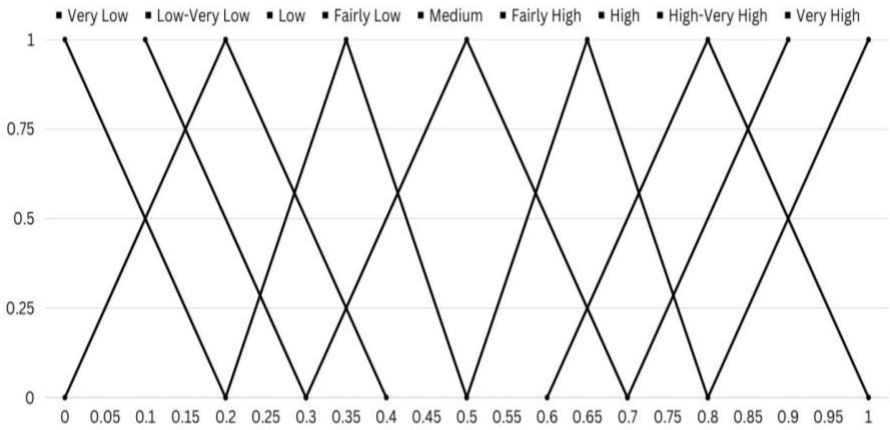


Figure 4-7 Conversion of a linguistic variable into the likelihood of an event

Figure 4-7 shows the conversion scale of linguistic grades into estimating the likelihood of an event. Fuzzy theory acts as a bridge between qualitative knowledge and numerical reasoning. For brevity, the conversion of causal factor entities shown from Figure 4-2 till Figure 4-6 into fuzzy probability is shown in Table 4-2. However, for detailed calculations, interested readers

are suggested to see (Zarei et al., 2019) as in the present work same method has been used. The resulting quantitative fuzzy probability is given as failure probability.

4.3.3.2 Mapping from NER to BN

This work demands a risk assessment technique that has the following advantages, as identified from analyzing five cases:

1. To model entities' dependence on each other
2. To incorporate a fuzzy probability of each causal factor
3. To model common cause failure

Bayesian network (BN) can process numerical data into a risk assessment (Kamil et al., 2021). It has the advantage over Bow-tie analysis (Khakzad et al., 2013;Dawuda et al., 2021) in modeling the statistical dependence of variables and handling discrete multi-states of a variable. These advantages make it advantageous to consider for risk assessment as opposed to other modelling tools such as Fault tree (Berrouane & Lounis, 2016; Taleb-Berrouane et al., 2021), Petri-nets (Kamil et al., 2019; Taleb-Berrouane et al., 2020; Taleb-Berrouane et al., 2019; Talebberrouane et al., 2016) and copula functions (Ramadhani et al., 2021).

The features extracted from NER model can develop BN model. BN serves the purpose due to its advantages mentioned above to model the relationship of entities to construct a likely failure scenario based on the identified features. Entities corresponding to their label are self-explained to depict causation. Figure 4-2 till Figure 4-6 illustrates how each entity with less manual interpretation can be presented in a diagram to understand its relationships better. BN model consists of two sub-models, causal analysis, and consequence analysis. The former can be modelled using fault tree analysis, while the latter uses event tree analysis when both former and latter are combined, called Bow-tie analysis. The literature shows a mapping of bow-tie analysis into BN (Khakzad et al., 2013).

The entities from NER model can also be mapped using a similar approach to BN model, using a simplified mapping algorithm based on entity labels illustrated in Figure 4-8. Based on manual interpretation, underlying causal factors extracted and selected can be used as BN's root cause and intermediate nodes. For instance, Figure 4-2 illustrates three basic causal factors: the root cause nodes in BN model and the MIC causal factor as an intermediate node. The scenario entity is a failure scenario node in BN (commonly referred to as the top event in fault tree analysis), which shows the failure scenario of the incident. Features identified as consequences and emergency response determine loss associated with the incident and the response taken to minimize it. The consequence entity will be represented in consequence node in the constructed BN model. One element of BN is missing in this mapping algorithm, i.e., safety barriers. In all the pipeline incidents narratives considered in this study, safety barriers are assumed, based on the nature of incident. A consistent number of safety barriers are used in the present study due to similar pipeline incidents. In BN, a dependency of one entity on another can be established by drawing an arc from a parent entity to a child entity. Once BN is constructed, an arc from the failure scenario node to the consequences node adds another state in the state set called the "NO" state, which accounts for the non-occurrence of the failure scenario node.

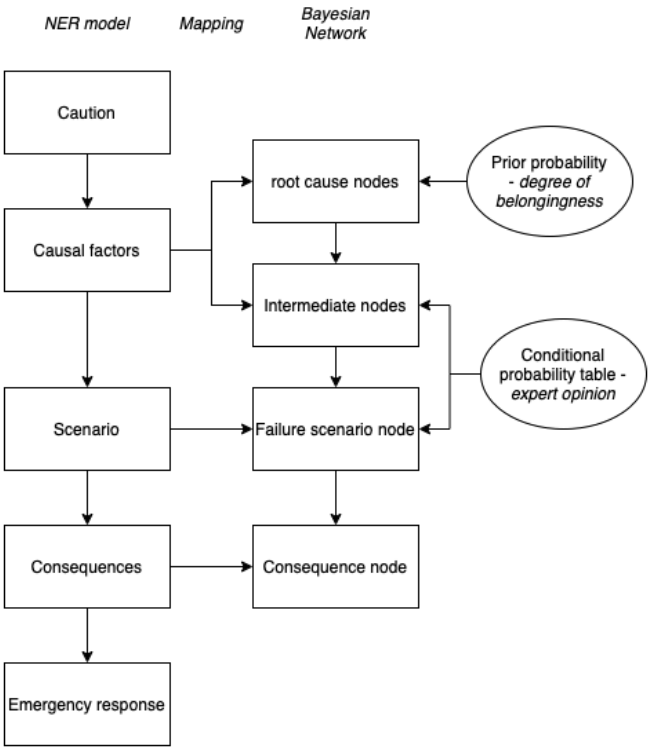


Figure 4-8 Mapping algorithm from NER to BN

BN demands prior probabilities of root nodes and conditional probability to define the relationships among each node (or entity). The former value can be given to the root node by directly providing the fuzzy probability of each causal factor from the fuzzy set theory. In contrast, the latter is based on expert opinion in defining the logical relationship of nodes. Each fuzzy probability of the root node can be given as a discrete state. For example, biofilm thickness and other entities' probabilities can be provided in each discrete state to predict the likelihood of a failure scenario. Table 4-2 shows all the entities' fuzzy probability used in the present study.

Table 4-2 Basic event probability for root nodes

Name of root node	Expert Opinion			K	Fuzzy possibility	Fuzzy probability
	A	B	C			
Bacteria	H	M	VH	1.61	0.74	0.0245

Biofilm	H	H	FH	1.60	0.75	0.0254
Pit	FL	H	FH	2.01	0.60	0.0098
Sulfur	M	H	FH	1.87	0.65	0.0134
Chloride	H	FH	FH	1.74	0.70	0.0184
Moisture	VH	H-VH	FH	1.37	0.82	0.0423
Pressure increase	L	FL	FH	2.63	0.40	0.0023
Rise in temperature	FL	M	L	2.83	0.35	0.0015
Coating	M	FH	M	2.15	0.55	0.0070

Table 4-3 Events along with their symbols

Events	Symbols
Safe evacuation with manual shut down and less property damage	C1
Safe evacuation with emergency shut down and moderate property damage	C2
Fire, low fatalities, moderate loss of property and injury	C3
Fire, fatalities, high loss of property and environmental damage	C4

Let us consider causation depicted in Figure 4-2 for equivalent BN construction. Three basic causal factors are shown as basic events in the cause-effect relationship. These are denoted as root cause nodes in the equivalent BN model based on the mapping algorithm. The same holds for the intermediate event represented as an intermediate node. Once causal analysis is constructed in BN model, the next step is the development of consequence analysis Safety barriers such as ignition, alarm, manual, and emergency shutdown are used to define consequences' states. Consequence events, along with their symbols, are shown in Table 4-3. These states consist of consequences extracted from NER model and what may go wrong when each safety barrier fails. BN model developed for causation shown in Figure 4-2 is depicted in

Figure 4-9. As can be seen, causal analysis can be illustrated similarly, as shown in Figure 4-2. BN model's flexibility in modelling incident scenarios makes it unique in risk assessment approaches. An arc from the leak node is drawn to the alarm safety barrier to establishing conditional dependence of the latter to the former. In contrast, Bow-tie analysis cannot model this dependence due to considering a leak as an initiating event that cannot influence safety barrier success or failure (Khakzad et al., 2013).

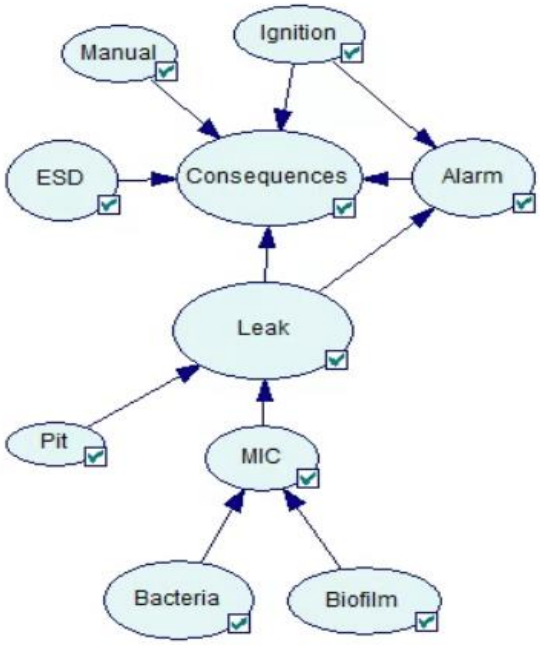


Figure 4-9 BN structure for case 1

Similarly, case 2 causation, shown Figure 4-3, is mapped to construct BN model. The developed BN is demonstrated in Figure 4-10; causal analysis is constructed based on causal factors and failure scenarios identified with less manual interpretation needed, as illustrated in Figure 4-3. In consequence analysis, safety barriers are assumed to model the consequences' states defined in Table 4-3 Events along with their symbols, as in the previous case. Case 3 causation from Figure 4-4 is mapped into BN model to perform the risk assessment. Figure 4-11 shows BN, highlighting another advantage of considering common causes or redundant failures. Moisture promotes microorganisms' growth and contributes to corrosion. Therefore,

to reduce model uncertainty, BN allows drawing an arc from the parent node (i.e., moisture) to the child nodes that will be affected due to the parent node's presence. Therefore, it also justifies the use of BN technique in this study.

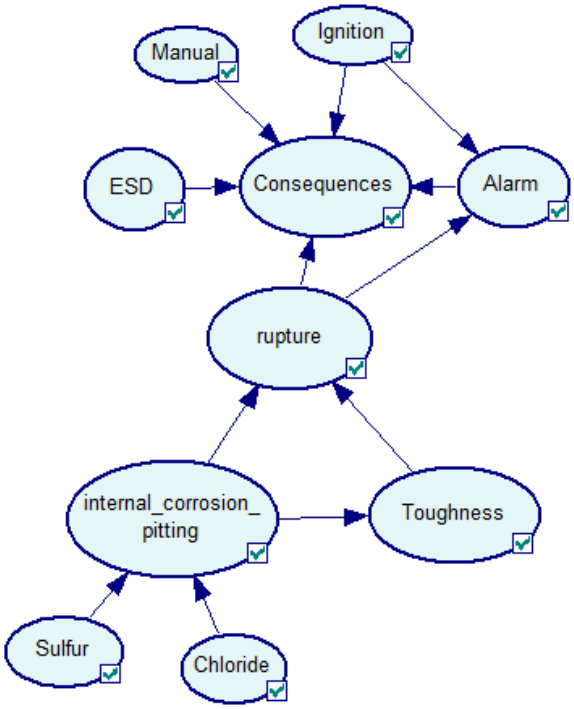


Figure 4-10 BN structure for case 2

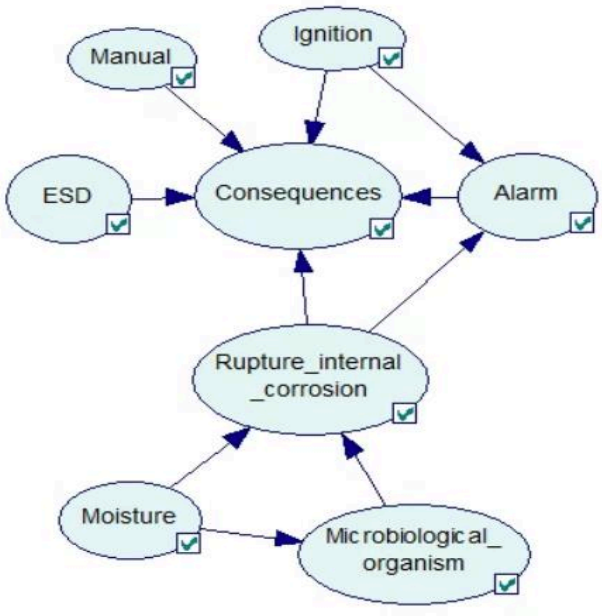


Figure 4-11 BN structure for case 3

In case 4, causation is shown in Figure 4-5; NER model obtains two causal factors and consequences. The failure scenario shown in Figure 4-5 does not need manual interpretation since the extracted entities already narrow it down to a cause-effect scenario. Similarly, BN is constructed in Figure 4-12 based on NER model mapping. Case 5 causation in Figure 4-6 suggests that due to moisture and sulfur availability, sulfate-reducing microorganisms contribute to external MIC. In addition, coating degradation also occurs, due to which a leak occurs rapidly. This incident does not cause a significant loss of property and human life. Figure 4-13 shows the developed BN model for this causation.

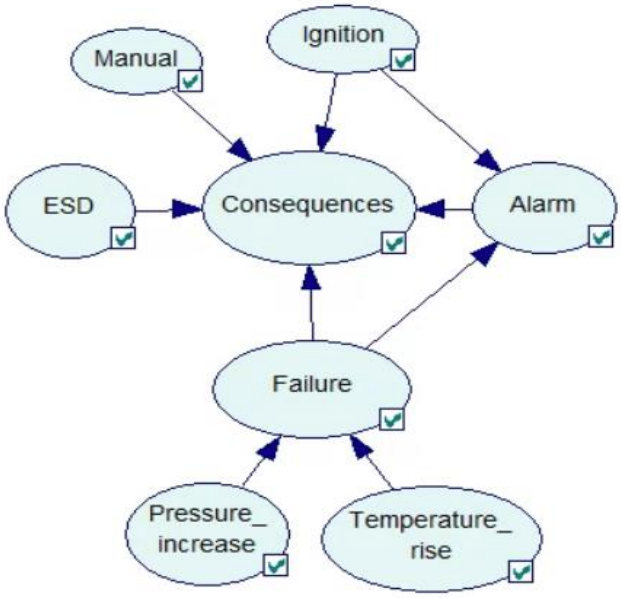


Figure 4-12 BN structure for case 4

BN models developed (from Figure 4-9 till Figure 4-13) for each case, shown in Table 4-1, depict the model's structure. BN model also demands probabilities for root cause nodes and conditional probability tables for intermediate nodes and failure scenario nodes (or pivot nodes). Root cause nodes' probabilities are estimated using fuzzy set theory; an example has been shown in Table 4-2, whereas conditional probability tables are given based on domain expertise.

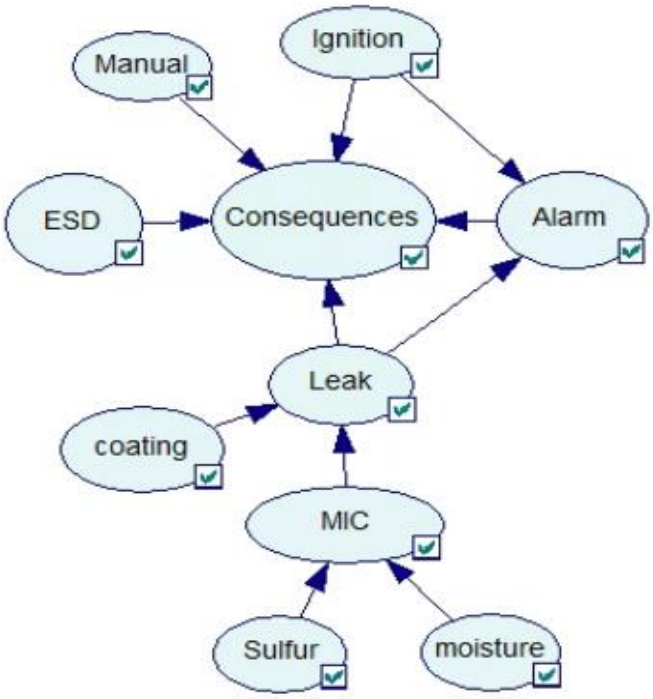


Figure 4-13 BN structure for case 5

4.4 Results and Discussion

In a risk assessment approach, three questions must be answered: what can go wrong? (failure scenario); its likelihood? And associated consequences? (Paté-Cornell, 1996). The present study answers all three questions using textual data as a source to extract relevant information for constructing BN structure with less manual interpretation and root cause probabilities, using fuzzy set theory to handle vagueness in human language. Table 4-4 shows the likelihood of failure and consequences obtained from BN model.

Table 4-4 Likelihood of failure and consequences for cases shown in Table 1

Event	Cases				
	1	2	3	4	5
Failure likelihood	0.021674	0.002158	0.042300	0.0029631	0.005804

C1	0.014891	0.001483	0.029062	0.002036	0.003988
C2	0.004616	0.000460	0.009008	0.000631	0.001236
C3	0.002135	0.000213	0.004166	0.000292	0.000572
C4	0.000033	0.000003	0.000064	0.000004	0.000009

It is necessary first to define the term 'risk'. According to the CCOHS (Canadian Centre for Occupational Health and Safety), the risk is the probability of harm from exposure to any potential event (hazard) which can cause loss of human life, property damage and adverse effects on the environment (CCOHS, 2021). Risk is estimated based on the results shown in Table 4-4. Earlier risk assessment questions are answered by analyzing five cases using NER model. Causations developed using NER model are mapped into BN model, which provides a way of quantitative risk assessment. It is essential to determine how failure likelihood and consequence values will convey a message to practitioners. Therefore, risk is evaluated by transforming subjective risk assessment into objective risk assessment.

This work advocates a 4×4 matrix consisting of four categories for the likelihood of an event and severity of consequences, as shown in Table 4-5. The range's description can be based on expert opinion. It can vary based on the nature and scope of the study. For example, consequence analysis can be performed based on system performance, financial loss, human health and environmental and ecological loss (Khan & Haddara, 2003).

Table 4-5 Categorization of likelihood and severity of consequences

	Category	Range
Likelihood (L)	Probable	$10^{-2} < \mathrm{L} \leq 1$
	Possible	$10^{-4} < \mathrm{L} \leq 10^{-2}$
	Unlikely	$10^{-6} < \mathrm{L} \leq 10^{-4}$
	Rare	$\mathrm{L} \leq 10^{-6}$

Severity of consequences (C)	Critical	$10^{-2} < C < 1$
	Major	$10^{-4} < C \leq 10^{-2}$
	Moderate	$10^{-7} < C \leq 10^{-4}$
	Minor	$C \leq 10^{-7}$

Categorizing the likelihood and severity of consequences provides a risk tolerance zone that constitutes a risk matrix. It is equally important to describe and categorize the scaling of output risk value. Therefore, Table 4-6 shows a risk assessment matrix used for risk evaluation purposes in this work and will identify risk levels based on the likelihood and consequence.

Table 4-6 Proposed risk assessment matrix

Consequence Likelihood	Minor	Moderate	Major	Critical
Rare	Acceptable	Acceptable	Acceptable	Tolerable-acceptable
Unlikely	Acceptable	Tolerable-acceptable	Tolerable-acceptable	Tolerable-unacceptable
Possible	Tolerable-acceptable	Tolerable-unacceptable	Unacceptable	Unacceptable
Probable	Tolerable-unacceptable	Tolerable-unacceptable	Unacceptable	Unacceptable

The study conducted (Markowski & Mannan, 2008; Ruge, 2004) acts as a guide for categorizing risk level and action required. According to (Markowski & Mannan, 2008), there are four levels for risk values: acceptable, tolerable-acceptable, tolerable-unacceptable and unacceptable. An acceptable risk level suggests that no risk-reducing strategies are needed to

continue operating. A tolerable-acceptable level requires regular monitoring based on the principle of ALARP (as low as reasonably practicable). In contrast, a tolerable-unacceptable level demands monitoring risk considering risk-reducing measures on short notice. An unacceptable risk level indicates that implementing risk-reducing strategies is necessary, and process must be halted until risk is brought to a tolerable-acceptable level.

Table 4-4 presents BN model's likelihood of failure and consequences. In case 1, a leak affected nearby soil and shut down the process (represented by C2 consequence state). If an economic loss is considered, the leak also affected the operator's facility and cost more than $1 million USD. Therefore, the risk is estimated as a product of failure likelihood and the C2 consequence state. All risk levels for each case are summarized in Table 4-7. As can be seen, the likelihood of a leak is probable, and the severity of the consequence C2 state is major. Risk categorization suggests that the risk level is unacceptable. Accordingly, risk-reducing strategies are necessary. Risk-reducing methods can include biocide to remove MIC bacteria present in the pipeline. The implementation will result in a lower risk value. To estimate the revised risk value, the observation in textual data will be first identified and then quantified for numerical reasoning. Risk cannot be eliminated from the operation but can be reduced to an acceptable limit to continue operating. The objective risk value helps in deciding the necessary action by the operator. The proposed methodology can predict the actual condition of the incident by evaluating the risk value from BN.

Table 4-7 Risk level of PHMSA incidents cases

Case	Characterization		Risk level
	Likelihood (L)	Severity of consequence (C)	
1	Probable	Major	Unacceptable

2	Possible	Major	
3	Probable	Major	
4	Possible	Major	
5	Possible	Major	

The case 2 incident caused a pipeline rupture with no injuries and less than $30,000 in losses. The incident happened in a rural area, which also significantly reduces the severity of the effects. Consequence state C1, shown in Table 4-3, best represents this case. Therefore, the risk is estimated from a product of rupture likelihood and the C1 state. Table 4-7 suggests an unacceptable risk level due to possible rupture likelihood and major severity of consequence. The risk reduction strategies must include removal of the sulfur and chloride compounds to reduce risk to a tolerable limit. Case 3 incident occurred due to the presence of moisture and microbiological organisms. Rupture of the pipeline occurred due to MIC with an economic loss of more than $700,000 USD. The incident caused no fire or injuries. The emergency shutdown was initiated. The estimated risk is a product of failure likelihood and consequence state C2. The failure scenario is probable, with a major severity of consequence. The risk reduction strategy is the same as in case 1, i.e., the removal of microorganisms from the pipeline. In case 4, two operational parameters, pressure and temperature, failed. This incident occurred in an urban area, but no injury was reported; there was a shutdown and an economic loss of more than $60,000 USD. The risk was evaluated in terms of failure likelihood and consequence state C1. The unacceptable risk was obtained due to defined risk criteria, which demanded risk-reducing measures. Case 5 was caused by the presence of sulfur and moisture. Sulfate-reducing bacteria caused an external MIC on the pipeline's surface. The coating on the pipeline degraded, thus exposing the bare surface of the pipe. The failure caused damage to property, resulting in an economic burden of more than $50,000 USD. The result also shows an unacceptable risk value and suggested reducing risk to continue the operations.

This step translates subjective risk obtained from BN model to objective risk assessment. It helps determine the state of a system based on the scaling of estimated risk. The predictions from the proposed approach and the actual condition of cases shown in Table 4-1 determine the same outcome. All the cases show an unacceptable risk level. The failure likelihood of cases 1 and 3 is probable. Additionally, the economic losses are highest in these cases. This demonstrates the robustness and applicability of the approach in determining objective risk levels.

NER model plays a vital role in extracting relevant features from textual data and constructing causation with less manual interpretation due to the associated labels with each extracted entity. The use of caution labels helps to understand if the causal factor effect is positive or negative. In other words, No abnormal pressure and abnormal pressure have the opposite meaning and can deviate narrative from reality. Furthermore, affected soil cannot be mislabelled as a causal factor, as it has the label of consequence. Therefore, labelling /annotating data helps recognize entities and their respective labels and does not need much interpretation to construct causation. Another advantage of using a NER model is the incremental learning process. When more or different sets of entities are required, NER model can easily incorporate those requirements. The use of NER model in the safety and risk domain shows a possible way of extracting and constructing causation from textual data. Two widely used and accepted methods in safety and risk are fuzzy set theory and BN. Five cases are used to demonstrate the application of the proposed approach. It predicts the correct risk level of all the failures that occurred. This indicates that methodology can be of potential use by operators/practitioners to predict objective risk levels based on their observations.

4.5 Verification of NER Model

4.5.1 Purpose of Verification

The aim of verifying NER model is to demonstrate the capability to identify features for causality and consequence analysis in the safety and risk domain. Textual data needed for verification are taken from a recent study published by (Liu et al., 2021), which focuses on automated feature extraction from textual data by employing a co-occurrence network. Therefore, a comparison is established with the co-occurrence matrix method to verify the proposed NER model performance and its potential to extract relevant features from textual data.

4.5.2 Verification Results and Discussion

Table 4-8 The incidents narrative reported in (G. Liu et al., 2021) with NER model result

1. Cause: Corrosion		
Incident narrative	Liu et al. co-occurrence networks output	NER model output
Internal corrosion … this segment of the pipe was removed entirely. [details available at (G. Liu et al., 2021)]	Internal corrosion Release Crude oil Impact soil Excavated	Internal corrosion causal fa ctor Release/spill scenario Impact soil consequence Excavate ER
2. Cause: Corrosion		
Lion oil called Sunoco control… Re-submitted on 3/19/2013 to include part e5f per phmsa	Pressure Internal corrosion	High pressure/internal corr osion causal factors Failure scenario

request. [details available at (G. Liu et al., 2021)]		Initiate/notify ER
3. Cause: Material failure		
A report of ammonia smell was phoned … This report was mailed 2/12/10 as the online reporting was not active. [details available at (G. Liu et al., 2021)]	No narrow-down co-occurrence network shown, assuming all relevant features extracted	Ammonia smell causal factor Release scenario shut down consequence notify/evacuation/repair/metallurgical/notification ER
4. Cause: Material failure		
On April 17, 2010 at approximately 11:30 am local time … The amount of contaminated soil removed from the leak site was 30 cubic yards. [details available at (G. Liu et al., 2021)]	Assuming all relevant features identified	Brushfire/leak scenario Defect/external/hook crack/fatigue causal factors evidence caution mechanical/corrosion flaw causal factors Shut down/impact/contaminate soil consequences Dispatch/excavate/notification/ metallurgical/restoration/repair ER
5. Cause: Equipment Failure		
The spill was a result of a crack in the flange … [details available at (G. Liu et al., 2021)]	Assuming all relevant features identified	Spill scenario Crack flange/valve causal factors

		Replace ER valve causal factor
6. Cause: Equipment Failure		
The location of …The threaded nipple on the pulsation dampener was replaced. [details available at (G. Liu et al., 2021)]	Assuming all relevant featu res identified	Release scenario Loose fitting/ thread nipple causal factor Properly caution thread nipple causal factor shut down/spray conseque nce manually/recovery/initiate/ mobilize/notification/inspe ct/replace ER
7. Cause: Natural Force damage		
A lightning strike caused a power outage…Impacted soils were place in drums and will be hauled off site to an approved facility. [details available at (G. Liu et al., 2021)]	Lightning strike Valve Sump Release impact soil	Lightning strike/power out age/valve opening/gasoline causal factor Overfill/release scenario Impact soil consequence
8. Cause: Natural Force damage		
Tank 824 water drain was leaking crude. …l due to changes in phmsa reporting	Crude oil Leak Roof Water	extreme temperature/ice ex pansion causal factor leak scenario

form. [details available at (G. Liu et al., 2021)]	Drain remove	

A constructed storyline by (Liu et al., 2021) for incident no.1 shown in Table 4-8 is as follows: "The abnormal pressure leads to internal corrosion of pipeline valves, the crude oil is then leaked to the ground and soil is contaminated, and the emergency response team is notified to excavate the soil for remediation, and the corroded valve is replaced." (Liu et al., 2021).

In the expression mentioned above, bold text denotes each word highlighted to define induced causation verified by incident no.1 of Table 4-8 (Liu et al., 2021).

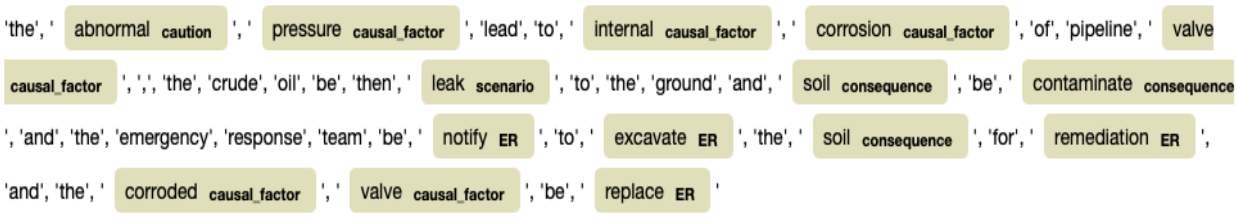


Figure 4-14 Highlighted entities from constructed induced causation by (G. Liu et al., 2021)

NER model result is shown in Figure 4-14 with labels- caution, causal factor, scenario, consequence and emergency response. Causal factor "pressure" along with caution "abnormal" led to "internal corrosion" of "valve," which resulted in the "leak" of crude oil. The consequences that resulted were "soil contamination". The emergency personnel were "notified" about the incident and "corroded" "valve" is "replaced" and soil is "excavated" for "remediation". Underlying entities reflecting the accident scenario can be easily extracted from induced causation. Moreover, NER model also extracts more information from textual data than highlighted by (Liu et al., 2021). This approach also demands minimal manual interpretation due to the use of labels to reflect each entity's belongingness with the respective labels. For instance, the word "abnormal" has the label caution, which defines the existence of abnormality and where it exists is defined by the causal factor label.

For brevity, the outcome of NER model for each incident is shown in column 3 of Table 4-8. In incident 1, the identified entities clearly show underlying features from the expression. The causal factor “internal corrosion” is responsible for crude oil "release". It resulted in "impacting" nearby "soil" which was "excavated" by emergency personnel. The extracted features show all the essential information that can be extracted for further analysis. As reported in (Liu et al., 2021), co-occurrence networks (PHMSA incident network and narrow-down corrosion network) were unable to identify "spill" from the narration. In incident 2, "failure" occurred due to abnormality present in "pressure," and it happened where "internal corrosion" had already appeared in the past. Emergency personnel were "notified" about the situation. NER model shows all the relevant information from the incident narrative. (Liu et al., 2021) result shows pressure and internal corrosion as the extracted features from co-occurrence networks. In incident 3 "release" scenario occurred is detected due to "ammonia smell". Emergency personnel are "notified" and respond in "evacuation" of people near the "release" location. (Liu et al., 2021) highlight that a key feature in this narration, i.e., "ammonia smell," was not extracted using the co-occurrence network. For incidents 3-6, (Liu et al., 2021) result is not reported since the study does not show a narrow-down co-occurrence network for causes under material and equipment failure labels. Therefore, it is assumed that all relevant features were extracted otherwise, as stated in incident 3. Incident 4 is a "brushfire" scenario that occurred due to the causal factor "crack" extended due to "fatigue". The "contaminate soil" was "excavated" from the site. The use of label caution can be helpful to know if identified entities could mislead from the fact. For example, for the present case, no "evidence” suggests "mechanical" or "corrosion" responsible for the "flaw". Caution labels consist of defects, abnormal or evidence when combined with the word "No" can result in opposite meanings. Hence, when a caution entity is detected from textual data, it needs more interpretation. In incident 5, a "spill" occurred due to a "crack" in a "flange" of a "valve". The "valve" is

"replaced" as an emergency response to the incident. Hence, a cause-effect scenario can be established from the features. Incident 6 entities suggest that due to "loose-fitting" of "threaded nipple" "release" occurred in a pump station. The consequence of this incident resulted in a "spray" of oil in the nearby area; the pump was "manually shut down" to prevent further "leak". When emergency personnel were "notified," "threaded nipple" was not "properly" installed, and they "replaced" it. Here, another caution entity "properly" is identified. This narrative meaning can be opposite to properly installed. Therefore, it is necessary to give more attention when dealing with a “caution” entity. The identified entities narrow down the incident to their defined labels of cause-effect.

Due to the narrow-down co-occurrence network (Liu et al., 2021), incidents under Natural force damage cause can be easily compared. Incident 7 extracted features suggest causal factors, "lightning strike" cause "power outage" which leads to "valve opening". This series of events leads to the "release" of "gasoline" which "impacted" nearby "soil". Causation can be constructed from extracted features, unlike (Liu et al., 2021) study in which "power outage" and "overfill" features were missing. Last incident 8 suggests that "extreme temperature," which leads to "ice expansion," are the reason for the "leak" of crude oil. Interestingly, these two features were missing in (G. Liu et al., 2021) study. However, other features of less importance in illustrating causation were present. In co-occurrence network developed by (Liu et al., 2021) omits essential features such as identifying "ammonia smell" and "evacuation" from incident 3 and "power outage" from incident 7, unlike NER model. Hence, it can be established that NER model can effectively extract relevant features from textual data.

This work demonstrates the application of a novel methodology that can utilize textual data by employing NLP and text mining techniques. This study investigates a new way of performing risk assessment from unstructured data. Structural data are not often available to predict the risk of process operations. The scarcity of relevant data motivates the investigation of

unstructured data available and seldom used to perform risk assessment. For instance, an observation made by the operator at a given time will convey ground reality about the process operation. Therefore, this study demonstrates a potential pathway for translating ground information into an objective risk assessment. It has been shown using five cases taken from the PHMSA database, and the outcome obtained from this approach predicts the risk levels were unacceptable for all the cases. Combined with operator expertise, the proposed methodology is a tool to predict objective risk in chemical process industries to determine system state and take necessary action if the risk is above defined criteria.

4.6 Conclusion

Automated feature extraction offers insights to analyze what resulted in the incident and its consequences. The fundamental challenge is to develop an approach that facilitates automated feature extraction based on trained data to understand and extract meaningful features from textual data and predict objective risk. This study demonstrates a unique method of combining NER with BN, using the defined mapping algorithm and employing fuzzy set theory as a bridge to transform features into their fuzzy probability for quantitative analysis. The following are the unique aspects of this methodology:

a) It is an easy-to-implement approach to predicting objective risk.

b) Features are identified, and their respective labels are self-explained to illustrate causation with minimal intervention needed.

c) Incremental annotations are based on information evolution or domain expertise by correcting NER model predictions.

d) It is applicable to other domain-specific areas like the aviation industry, in which human communication plays a vital role.

Five PHMSA incident results indicate risk levels were unacceptable and coherent with actual conditions. This demonstrates the methodology's ability to effectively identify features from textual data and transform them into objective risk assessment, unlike previous studies that are only limited to feature extraction and do not provide any pathway for predicting risk from textual data.

The verification exercise shows NER model's ability of relevant feature extraction, making it a potential enrichment of the co-occurrence network technique in which relevant features are omitted. A comprehensive verification is pending due to the unavailability of ground information that practitioners can perform to evaluate model performance. The advantages of using BN have been seen in accommodating multiple states of each causal factor, common-cause failure, conditional dependence of the parent node to the child node and quantitative analysis. All these advantages were exploited in the applicability aspect of the methodology. Further enrichment of this work can be done by incorporating an automated relation extraction of entities, illustrating a cause-effect scenario using named and relations' entity recognition together and eliminating manual interpretations. The expected result will provide automated causation extraction from textual data.

4.7 Acknowledgements

The first author would like to thank Explosion software company for providing a prodigy research license. The authors acknowledge the financial support provided by Genome Canada and their supporting partners through the Large Scale Applied Research Project and the Canada Research Chair (CRC) Tier I Program in Offshore Safety and Risk Engineering.

4.8 References

1. Abdullah, D., Takahashi, H., & Lakhani, U. (2019). Domain Specific Ontology Enhancing Communication Accuracy in Airport Operation. *Proceedings - 2019 IEEE 14th International Symposium on Autonomous Decentralized Systems, ISADS 2019*. https://doi.org/10.1109/ISADS45777.2019.9155591
2. 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*. https://doi.org/10.1016/j.psep.2021.09.022
3. Allison, E., & Mandler, B. (2018). *Transportation of Oil, Gas, and Refined Products*. The Methods, Volumes, Risks, and Regulation of Oil and Gas Transportation.
4. Aziz, A., Ahmed, S., & Khan, F. I. (2019). An ontology-based methodology for hazard identification and causation analysis. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2018.12.008
5. Bakke, R., & Olsson, P. Q. (1986). Biofilm thickness measurements by light microscopy. *Journal of Microbiological Methods*. https://doi.org/10.1016/0167-7012(86)90005-9
6. Ben Seghier, M. E. A., Keshtegar, B., Taleb-Berrouane, M., Abbassi, R., & Trung, N. T. (2021). Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines. *Process Safety and Environmental Protection*, *147*(January), 818–833. https://doi.org/10.1016/j.psep.2021.01.008
7. Berrouane, M. T., & Lounis, Z. (2016). *Safety assessment of flare system by fault tree analysis*. 229–234.

8. Bersani, C., Citro, L., Gagliardi, R. V., Sacile, R., & Tomasoni, A. M. (2010). Accident occurrance evaluation in the pipeline transport of dangerous goods. *Chemical Engineering Transactions*. https://doi.org/10.3303/CET1019041
9. Bougofa, M., Taleb-Berrouane, M., Bouafia, A., Baziz, A., Kharzi, R., & Bellaouar, A. (2021). Dynamic Availability Analysis Using Dynamic Bayesian and Evidential Networks. *Process Safety and Environmental Protection*. https://doi.org/https://doi.org/10.1016/j.psep.2021.07.003
10. Bubbico, R. (2018). A statistical analysis of causes and consequences of the release of hazardous materials from pipelines. The influence of layout. *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2018.10.006
11. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. In *IEEE Computational Intelligence Magazine*. https://doi.org/10.1109/MCI.2014.2307227
12. CCOHS. (2021). *Canadian Center for Occupational Health and Safety, Government of Canada*. https://www.ccohs.ca/oshanswers/hsprograms/hazard_risk.html
13. Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Advanced Engineering Informatics*. https://doi.org/10.1016/j.aei.2019.100959
14. Chen Shu-Jen and Hwang, C.-L. (1992). Fuzzy Multiple Attribute Decision Making Methods. In *Fuzzy Multiple Attribute Decision Making: Methods and Applications* (pp. 289–486). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-46768-4_5

15. Chokor, A., Naganathan, H., Chong, W. K., & Asmar, M. el. (2016). Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning. *Procedia Engineering*. https://doi.org/10.1016/j.proeng.2016.04.200
16. Clark, A., Fox, C., & Lappin, S. (2010). The Handbook of Computational Linguistics and Natural Language Processing. In *The Handbook of Computational Linguistics and Natural Language Processing*. https://doi.org/10.1002/9781444324044
17. Cunningham, A. B., Lennox, J. E., & Ross, R. J. (2012). Biofilms: The Hypertextbook. In *Http://Www.Hypertextbookshop.Com/Biofilmbook/V005/R001/*.
18. Dawuda, A.-W., Taleb-berrouane, M., & Khan, F. (2021). A probabilistic model to estimate microbiologically influenced corrosion rate. *Process Safety and Environmental Protection*. https://doi.org/https://doi.org/10.1016/j.psep.2021.02.006
19. Deshpande, G., Motger, Q., Palomares, C., Kamra, I., Biesialska, K., Franch, X., Ruhe, G., & Ho, J. (2020). Requirements Dependency Extraction by Integrating Active Learning with Ontology-Based Retrieval. *Proceedings of the IEEE International Conference on Requirements Engineering*. https://doi.org/10.1109/RE48521.2020.00020
20. Deyab, S. M., Taleb-berrouane, M., Khan, F., & Yang, M. (2018). Failure analysis of the offshore process component considering causation dependence. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2017.10.010
21. Ferdous, R., Khan, F., Sadiq, R., Amyotte, P., & Veitch, B. (2013). Analyzing system safety and risks under uncertainty using a bow-tie diagram: An innovative

approach. *Process Safety and Environmental Protection*, *91*(1–2), 1–18. https://doi.org/10.1016/j.psep.2011.08.010

22. Grosman, J. S., Furtado, P. H. T., Rodrigues, A. M. B., Schardong, G. G., Barbosa, S. D. J., & Lopes, H. C. V. (2020). Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*. https://doi.org/10.1016/j.is.2020.101553
23. Guo, J., & Huang, J. C. (2016). Ontology learning and its application in software-intensive projects. *Proceedings - International Conference on Software Engineering*. https://doi.org/10.1145/2889160.2889264
24. Halim, S. Z., Janardanan, S., Flechas, T., & Mannan, M. S. (2018). In search of causes behind offshore incidents: Fire in offshore oil and gas facilities. *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2018.04.006
25. Halim, S. Z., Yu, M., Escobar, H., & Quddus, N. (2020). Towards a causal model from pipeline incident data analysis. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2020.06.047
26. Honnibal, M., & Montani, I. (2021a). *Prodigy*. https://prodi.gy/docs/recipes#ner-manual
27. Honnibal, M., & Montani, I. (2021b). *Prodigy · An annotation tool for AI, Machine Learning & NLP*. https://prodi.gy/ accessed on September 2021
28. Honnibal, M., & Montani, I. (2021c). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. In *To appear* (3.0). MIT. https://spacy.io/ accessed on August 2021
29. 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*. https://doi.org/10.1016/j.ssci.2019.05.029

30. Ide, N., & Pustejovsky, J. (Eds.). (2017). *Handbook of Linguistic Annotation* (1st ed.). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2
31. Kamil, M. Z., Khan, F., Song, G., & Ahmed, S. (2019). Dynamic Risk Analysis Using Imprecise and Incomplete Information. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, *5*(4). https://doi.org/10.1115/1.4044042
32. Kamil, M. Z., Taleb-Berrouane, M., Khan, F., & Ahmed, S. (2019). Dynamic domino effect risk assessment using Petri-nets. *Process Safety and Environmental Protection*, *124*. https://doi.org/10.1016/j.psep.2019.02.019
33. Kamil, M. Z., Taleb-Berrouane, M., Khan, F., & Amyotte, P. (2021). Data-driven operational failure likelihood model for microbiologically influenced corrosion. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2021.07.040
34. Kannan, P., Kotu, S. P., Pasman, H., Vaddiraju, S., Jayaraman, A., & Mannan, M. S. (2020). A systems-based approach for modeling of microbiologically influenced corrosion implemented using static and dynamic Bayesian networks. *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2020.104108
35. Khakzad, N., Khan, F., & Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. *Process Safety and Environmental Protection*, *91*(1–2), 46–53. https://doi.org/10.1016/j.psep.2012.01.005

36. Khan, F. I., & Haddara, M. M. (2003). Risk-based maintenance (RBM): A quantitative approach for maintenance/inspection scheduling and planning. *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2003.08.011
37. Kwon, J. H., Kim, B., Lee, S. H., & Kim, H. (2013). Automated procedure for extracting safety regulatory information using natural language processing techniques and ontology. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*.
38. Li, X., Penmetsa, P., Liu, J., Hainen, A., & Nambisan, S. (2021). Severity of emergency natural gas distribution pipeline incidents: Application of an integrated spatio-temporal approach fused with text mining. *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2020.104383
39. Little, B. J., & Lee, J. S. (2014). Microbiologically influenced corrosion: An update. In *International Materials Reviews*. https://doi.org/10.1179/1743280414Y.0000000035
40. 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. https://doi.org/https://doi.org/10.1016/j.psep.2021.05.036
41. Markowski, A. S., & Mannan, M. S. (2008). Fuzzy risk matrix. *Journal of Hazardous Materials*. https://doi.org/10.1016/j.jhazmat.2008.03.055
42. Nakata, T. (2017). Text-mining on incident reports to find knowledge on industrial safety. *Proceedings - Annual Reliability and Maintainability Symposium*. https://doi.org/10.1109/RAM.2017.7889795

43. Onisawa, T. (1988). An approach to human reliability in man-machine systems using error possibility. *Fuzzy Sets and Systems*, *27*(2), 87–103. https://doi.org/https://doi.org/10.1016/0165-0114(88)90140-6

44. Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., & Diamantaras, K. I. (2019). Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy. *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*. https://doi.org/10.1145/3350546.3352543

45. Paté-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering and System Safety*. https://doi.org/10.1016/S0951-8320(96)00067-1

46. *Pipeline and Hazardous Materials Safety Administration*. (2022). https://www.phmsa.dot.gov/incident-reporting

47. Ramadhani, A., Khan, F., Colbourne, B., Ahmed, S., & Taleb-Berrouane, M. (2021). Environmental load estimation for offshore structures considering parametric dependencies. *Safety in Extreme Environments*.

48. Robinson, S. D., Irwin, W. J., Kelly, T. K., & Wu, X. O. (2015). Application of machine learning to mapping primary causal factors in self reported safety narratives. *Safety Science*, *75*, 118–129. https://doi.org/https://doi.org/10.1016/j.ssci.2015.02.003

49. Ruge, B. (2004). Risk Matrix as Tool for Risk Assessment in the Chemical Process Industries. In *Probabilistic Safety Assessment and Management*. https://doi.org/10.1007/978-0-85729-410-4_431

50. Salgar-Chaparro, S. J., Darwin, A., Kaksonen, A. H., & Machuca, L. L. (2020). Carbon steel corrosion by bacteria from failed seal rings at an offshore facility. *Scientific Reports*. https://doi.org/10.1038/s41598-020-69292-5
51. 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*. https://doi.org/10.1016/j.ssci.2020.104747
52. Stover, R. (2013). *AMERICA'S DANGEROUS PIPELINES*. Center for Biological Diversity. https://www.biologicaldiversity.org/campaigns/americas_dangerous_pipelines/
53. Sugeno, M., & Kang, G. T. (1986). Fuzzy modelling and control of multilayer incinerator. *Fuzzy Sets and Systems*, *18*(3), 329–345. https://doi.org/https://doi.org/10.1016/0165-0114(86)90010-2
54. Taleb-Berrouane, M., Khan, F., & Amyotte, P. (2020). Bayesian Stochastic Petri Nets (BSPN) - A new modelling tool for dynamic safety and reliability analysis. *Reliability Engineering and System Safety*, *193*. https://doi.org/10.1016/j.ress.2019.106587
55. Taleb-Berrouane, M., Khan, F., & Hawboldt, K. (2021). Corrosion risk assessment using adaptive bow-tie (ABT) analysis. *Reliability Engineering & System Safety*, *214*(May), 107731. https://doi.org/10.1016/j.ress.2021.107731
56. Taleb-Berrouane, M., Khan, F., Hawboldt, K., Eckert, R., & Skovhus, T. L. (2018). Model for microbiologically influenced corrosion potential assessment for the oil and gas industry. *Corrosion Engineering, Science and Technology*, *53*(5), 378–392. https://doi.org/10.1080/1478422X.2018.1483221

57. Taleb-Berrouane, M., Khan, F., & Kamil, M. Z. (2019). Dynamic RAMS analysis using advanced probabilistic approach. *Chemical Engineering Transactions*, *77*, 241–246. https://doi.org/10.3303/CET1977041

58. Talebberrouane, M., Khan, F., & Lounis, Z. (2016). Availability analysis of safety critical systems using advanced fault tree and stochastic Petri net formalisms. *Journal of Loss Prevention in the Process Industries*, *44*, 193–203. https://doi.org/10.1016/j.jlp.2016.09.007

59. Taleb-Berrouane, M., Sterrahmane, A., Mehdaoui, D., & Lounis., Z. Emergency Response Plan Assessment Using Bayesian Belief Networks, 3rd Workshop and Symposium on Safety and Integrity Management of Operations in Harsh Environments (C-RISE3) 1 (2017).

60. 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*. https://doi.org/10.1016/j.compind.2015.09.005

61. Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016a). Application of machine learning to construction injury prediction. *Automation in Construction*. https://doi.org/10.1016/j.autcon.2016.05.016

62. Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*. https://doi.org/10.1016/j.autcon.2015.11.001

63. Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT*.

64. Ullah, S., & Al Islam, A. B. M. A. (2019). A framework for extractive text summarization using semantic graph based approach. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3362966.3362971
65. Videla, H. A., & Herrera, L. K. (2005). Microbiologically influenced corrosion: Looking to the future. *International Microbiology*. https://doi.org/10.2436/im.v8i3.9523
66. Wu, C. G., Xu, X., Zhang, B. K., & Na, Y. L. (2013). Domain ontology for scenario-based hazard evaluation. *Safety Science*. https://doi.org/10.1016/j.ssci.2013.06.003
67. Yang, R., Khan, F., Taleb-Berrouane, M., & Kong, D. (2020). A time-dependent probabilistic model for fire accident analysis. *Fire Safety Journal*, *111*(December 2018), 102891. https://doi.org/10.1016/j.firesaf.2019.102891
68. Zadeh, L. A. (1965). Fuzzy Sets. *Informational and Control*, *8*, 338–353.
69. Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning-I. *Information Sciences*. https://doi.org/10.1016/0020-0255(75)90036-5
70. Zarei, E., Khakzad, N., Cozzani, V., & Reniers, G. (2019). Safety analysis of process systems using Fuzzy Bayesian Network (FBN). *Journal of Loss Prevention in the Process Industries*, *57*, 7–16. https://doi.org/https://doi.org/10.1016/j.jlp.2018.10.011
71. Zhang, J., Fu, J., Hao, H., Fu, G., Nie, F., & Zhang, W. (2020). Root causes of coal mine accidents: Characteristics of safety culture deficiencies based on accident statistics. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2020.01.024

72. Zhou, J., Hanninen, K., & Lundqvist, K. (2017). A hazard modeling language for safety-critical systems based on the hazard ontology. *Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017*. https://doi.org/10.1109/SEAA.2017.48

5 A methodical approach for knowledge-based fire and explosion accident likelihood analysis

Preface

This chapter has been published in the ***Process Safety and Environmental Protection*** Journal. I am the primary author of this manuscript, along with co-authors Drs. Faisal Khan, S. Zohra Halim, Paul Amyotte and Salim Ahmed. I developed the methodical approach for fire and explosion accident likelihood analysis from past experiences. I prepared the first draft of the manuscript and revised it based on the co-authors' and peer review feedback. The co-author Dr. S. Zohra Halim assisted in model development, testing and revision based on peer review feedback. The co-author Dr. Faisal Khan proposed the conceptual framework and helped develop the framework, testing and revising the model. The co-authors, Drs. Paul Amyotte and Salim Ahmed provided constructive feedback to improve the readability, review and revision based on peer review feedback and finalizing the manuscript.

Reference: Kamil, M. Z., Khan, F., Halim, S. Z., Amyotte, P., & Ahmed, S. (2023). A methodical approach for knowledge-based fire and explosion accident likelihood analysis. *Process Safety and Environmental Protection, 170, 339-355.*

Abstract

An accident database is an excellent data source if appropriately leveraged along with domain expertise. However, a proper framework and tools are required to extract data from a database. The current work aims to develop such a framework by systematically introducing a unique approach to integrate three techniques. First, Natural Language Processing (NLP) is used to extract causal and contributing factors from an accident database. Second, an Interpretive Structural Model (ISM) establishes the interrelationship and hierarchy of the extracted factors. Third, a probabilistic method for quantitative reasoning and accident analysis is employed.

This integrated approach is applied to the US Chemical Safety and Hazard Investigation Board (CSB) oil and refining (downstream) incident database to develop a generalized accident causation model. The model provides insight into the factors responsible for accidents (i.e., commonalities among casualties), interactions, and accident pathways. It can also be used to develop strategies for preventing accidents. The model is tested on ten scenarios from the CSB and verified on six incidents from the IChemE database. The results are promising in establishing the model's efficacy in predicting adverse events. Sensitivity analysis shows that management of change and lack of procedure and training have the highest sensitivity towards fire and explosion, and therefore need proper attention. This approach will be an essential tool for Safety 4.0, enabling process safety in the digitalization process.

Keywords*: Lessons learnt, process safety excellence, natural language processing, interpretive structural modelling, Bayesian network, Safety 4.0*

5.1 Introduction

A chemical plant consists of many highly complex process systems. Due to their own complexity and that arising from process digitalization, it is challenging to model and predict process accidents. The complex interaction among equipment, operators, hardware, software, procedures, and operating conditions leads to potential hazards (Rathnayaka et al., 2011). In 2022 alone, thirty-five process incidents occurred in the US. From 2020 till today, one hundred and fifty-three incidents have been reported to the CSB (U.S. Chemical Safety and Hazard Investigation Board, 2022). Recently, a chemical explosion occurred in an electronic industry, resulting in ten fatalities and 22 injuries due to an unknown chemical (under investigation) at Hapur, India 37 miles from New Delhi. Disregarding process safety and regulatory oversight are believed to be the main reasons behind process accidents in India (Reuters, 2022a). More than six thousand people were evacuated due to a fire at a fertilizer plant in North Carolina; it was reported that roughly 600 ton of ammonium nitrate was present, which could have led to

a major catastrophe (The Guardian, 2022). Last month, a catastrophic incident in Bangladesh resulted in forty-one fatalities, including nine firefighters and $100 million in losses. A series of fires and explosions occurred due to hydrogen peroxide drums and garments for export purposes. This was caused due to haphazard safety norms ("The New York Times," 2022). An explosion in a chemical plant in Slovenia, producing chemicals for paints, rubber and other industries, caused six deaths; the plant manager blamed human error as the root cause (Reuters, 2022b). In Louisiana, a chemical plant explosion occurred in a storage tank. It was an empty ethylene dichloride tank and led to six people being injured; and the fire was controlled shortly after the explosion ("U.S. News," 2022a). This incident invoked the federal regulator (the U.S. Department of Labor's Occupational Safety and Health Administration) to give citations to four companies (two Texas-based and two Louisiana-based) and impose a total of $139,000 in penalties ("U.S. News," 2022b).

These are a few examples of accidents that occured this year. If analyzed properly, there seem to be commonalities between these accidents. Many of the accidents happened due to flouting safety norms due to management oversights or lax regulators' inspections.

Accident causation is performed to answer two fundamental questions. First, what went wrong? Second, how did it happen? In the past century, efforts have been made to answer these questions. Accident causation models are divided into linear accident models such as System hazard identification, prediction and prevention (SHIPP) proposed by (Rathnayaka et al., 2011) and non-linear accident models. The latter are further divided into the following categories. (Fu et al., 2020):

- Human-based accident model.
- Statistics-based accident model.
- Energy-based accident model.
- System-based accident model.

Interested readers are referred to (Fu et al., 2020) to learn about the evolution of linear and non-linear accident causation models. In addition to the aforementioned accident causation models, there are similar terms for the prevention of accidents, such as accident investigation models (Pasman et al., 2018), accident analysis (Haghighattalab et al., 2019) and accident prediction (la Torre et al., 2019). These models aim to discover what caused an accident and its potential pathway. Past studies mainly focused on understanding a single accident's causation and ways to prevent identified causal or contributing factors. However, this approach needs to evolve with the complexities in process systems. An accident happens due to multiple flaws when the contributing factors form a hazard pathway. As cited by John Mogdorf in the speech after the Texas City refinery incident, "This was a preventable incident, as I will explain. It should be seen as a process failure, a cultural failure and a management failure" (Knegtering & Pasman, 2009). Modelling of the BP Texas refinery incident highlighted key factors. These factors include overpressurization of the raffinate splitter, hydrocarbon venting, insufficient procedures, lack of supervision, trailer presence and lack of information/notification (Khan & Amyotte, 2007). As can be seen, it is a combination of a process failure, lack of safety culture and management oversight. Repetition of process incidents highlights that the lessons learned from past incidents are not implemented to improve process safety. Accident databases are great learning sources for acquiring new information. However, these databases must be used appropriately to extract industry-specific information. Interested readers may refer to (Mannan & Waldram, 2014). As Dr. Mannan said, “The old saying is that you can take a horse to water but you cannot make it drink. The same will be true of the database that we advocate” (Mannan & Waldram, 2014). To investigate incidents through a database, this study focuses on the CSB database and related studies that use the database for lessons learned for industry implementation. The investigation of incidents has been done to analyze the hierarchy of

controls and the concept of inherent safety. The prominent studies conducted to leverage the CSB database are discussed as follows:

The concept of inherent safety in a plant should follow through with a systematic hierarchy of controls at the design and operation levels. The broad steps include the identification of a hazard, and its avoidance, followed by reducing its likelihood and severity (via inherent safety principles), segregation, use of passive and active safety barriers (add-on safety), procedural safeguards and residual risk reduction measures (Amyotte et al., 2009). A total of eighty-eight incidents over the period of 1998-2016 from the CSB database were reviewed based on the hierarchy of controls. The analysis is conducted in two rounds; round 1 consists of sixty-three reports (Amyotte et al., 2011), and the second-round accounts for twenty-five incidents (P. Amyotte et al., 2018). The findings were that 36% of examples were related to inherently safer design, 10% were passive safety measures, 16% were active safety measures, and 48% were procedural safeguards (Amyotte et al., 2018). This shows the attention to loss prevention via a hierarchy of controls. The role of inherently safer design at different stages is significant in avoiding the hazard; interested readers can refer to it (Amyotte & Khan, 2021). Another study is conducted on the CSB database to analyze twenty-one reports over ten years for oversights in process hazard analysis (PHA). The outcome indicated that in 19% of the cases PHA team did not evaluate proper prevention and control measures, and future recommendations were remotely operated valves, worker involvement and freeze protection (Kaszniak, 2010). A study is performed on sixty reports from the CSB to analyze incidents and commonalities among them. Non-routine operations and stationary sitting sources were the highest incidents (Baybutt, 2016).

The investigation of databases, as discussed earlier, gives insight into what went wrong. The implementation is shown in the hierarchy of controls and inherent safety measures (Amyotte & Khan, 2021). However, learnings from investigations are still lacking short of

implementation. Therefore, the key question remains: why do accidents still happen? It is due to lacking implementation of knowledge to avoid accidents, insufficient procedure and training and the introduction of process digitalization (Amyotte et al., 2016). Therefore, new methods are needed for capturing accident causations that are a combination of factors such as organizational failure, lack of competency and equipment failure (Knegtering & Pasman, 2009). A simple and universal accident investigation method does not exist (Pasman et al., 2018). Future accident causation approaches need to give more attention to safety culture's role (as a subset of organizational culture). Safety culture and its relationship with process safety are well-studied (Olive et al., 2006). As Trevor Kletz observed, "Organizations have no memory. Only people have memory and they move on". Another observation: "Accidents are not due to lack of knowledge, but failure to use the knowledge we have" (IChemE Safety and Loss Prevention, 2022). A simple and universal approach is needed to implement database information and domain expertise knowledge to develop a generalized causation model. Based on the developed generalized mode, accidents can be predicted. This study also served the purpose of people's memory retention, as mentioned by Dr. Kletz. Information from a database and knowledge from domain expertise is required in the proposed approach.

The following research questions are answered through this study:

- How to automate information extraction from Chemical Safety and Hazard Investigation Board (CSB) reports using NLP?
- How to combine information from previous accidents with domain expertise knowledge to develop a generalized causation model systematically for accident prediction?
- Is there a way to analyze commonalities between incidents and their interrelationship to assist in learning from incidents?

This study introduces a unique approach. This approach utilizes the NLP technique called Named entity recognition (NER) to extract relevant features. A custom NER using spaCy has a good performance of custom NER compared to other available methods, such as Bluemix and Stanford NLP (Shelar et al., 2020). This is followed by a systematic process of developing qualitative (ISM method) and quantitative reasoning (BN model) for learning lessons from past experiences. The ISM method is developed to visualize a hierarchy among different factors in a complex system that assists in making decisions. It can establish interrelationships among studied factors (Warfield, 1974). BN is a well-established safety and risk technique used to model accident scenarios from cause to consequence (Kamil et al., 2019). Previous studies that leverage the ISM method in the safety and risk domain rely on a literature review to list the factors for the study (Huang et al., 2020; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019). However, this process can be automated with the increasing availability of NLP and text mining tools, like NER (Shelar et al., 2020) and the co-occurrence matrix (G. Liu et al., 2021). Moreover, it can be easily implemented in various fields, from aviation safety databases to process accidents, making it universal but domain dependent.

The key to learning lessons from previous incidents is first to analyze underlying causal and contributing factors leading to an incident. It is immensely important to learn lessons in order to prevent accidents (U.K. HSE, Investigating Accidents and Incidents, 2004). Furthermore, it is equally vital to identify complex interactions among accident causations and their commonalities and interrelationships among the same incident type. Quantitative reasoning defines factors' interrelationships and estimates each accident likelihood with potential pathways based on the given conditions. Section 5.2 of this chapter shows a proposed methodology based on the approach shown in Figure 5-1. Its application is performed on the CSB database of oil and refining (downstream) incidents in section 5.3. Section 5.4 deals with results and discussion of the generalized causation model, model testing and verification,

followed by sensitivity analysis. Conclusions drawn from the study are presented in section 5.5.

5.2 Methodology to Develop Knowledge-based Accident Causation Model

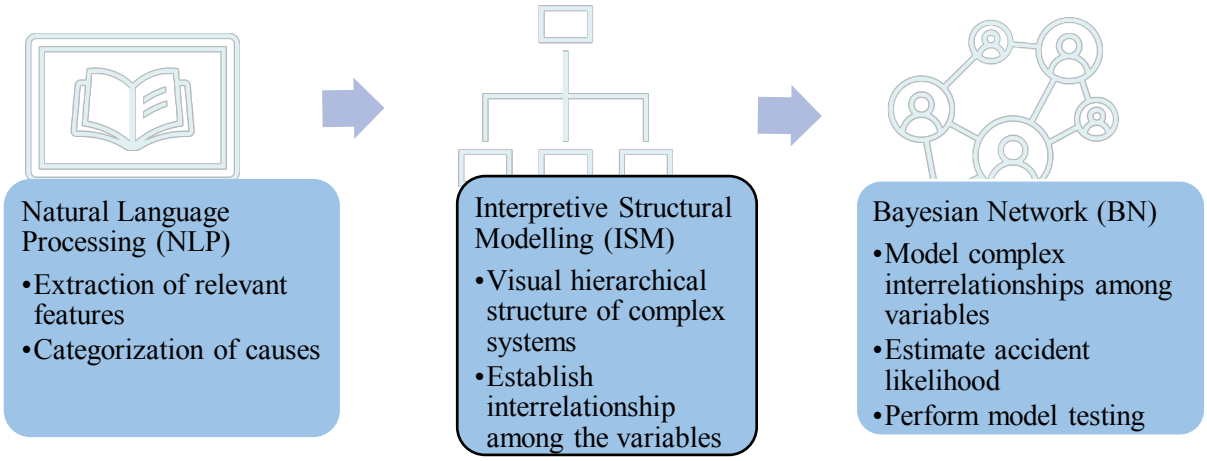


Figure 5-1 A three-step systematically integrated approach to develop a generalized causation likelihood model

A methodical approach involving three steps is shown in Figure 5-1. Step 1 is leveraging NLP capabilities in text mining from the CSB database and categorizing them using a custom-named entity recognition model. This serves as an input to the ISM process, which in previous works relied on manual literature review for the input factors. The ISM process provides two important aspects of this study: establishing interrelationships among factors and developing a hierarchical structure in a complex system. These two steps serve as qualitative analyses in this study. Qualitative analysis is further transformed into quantitative

reasoning through the developed mapping algorithm. The advantage of quantitative analysis is to model interrelationships and estimates the likelihood of an accident.

Figure 5-2 presents the proposed methodology to systematically analyze past incidents to develop a generalized causation model that can determine the influence of identified factors on one another and the accident pathways in a hierarchical structure. This qualitative analysis is further developed into quantitative reasoning to understand factors' interrelationships better. Step 5.2.1 applies NLP and text mining technique - NER to extract relevant features from the CSB database of oil and refining (downstream) incidents. Step 5.2.2 develops a hierarchical qualitative structure through the ISM process, followed by quantitative reasoning via the BN model in step 5.2.3. The details of these steps are as follows:

5.2.1 Application of Natural Language Processing (NLP)

5.2.1.1 Report section selection and data pre-processing

Incident databases are a source of learning that can avoid future incidents in an engineering discipline. Every database has its way of storing data. Pipeline failure investigation reports by PHMSA are organized on four separate databases based on material transported, such as hazardous liquid, natural gas transmission, natural gas distribution and liquified natural gas (Pipeline and Hazardous Materials Safety Administration, 2022). However, CSB databases are categorized by incident types such as chemical distribution - fire and explosion, flammable vapor, combustible dust explosion and fire etc. (Chemical Safety and Hazard Investigation Board , 2022). Each database has different rules for incident reporting that also change with time. When selecting a database, it is essential to see what section of a report contains relevant information for the causation of an incident. For instance, a recent study analyzing the PHMSA database considered the "comment " section as a data source for their research (G. Liu et al., 2021).

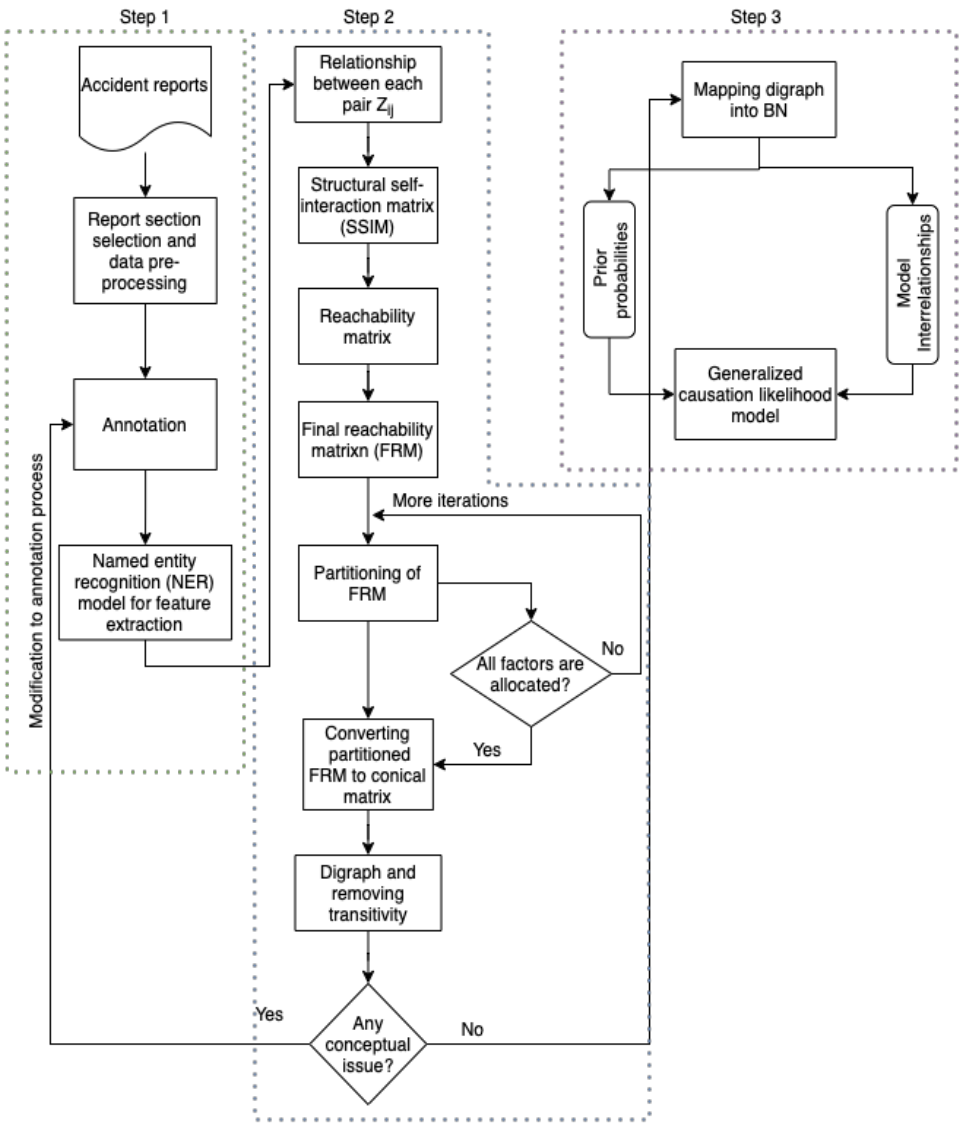


Figure 5-2 The proposed methodology for learning lessons from past experiences and predicting adverse events

The next task is to pre-process a selected section of reports to make it more compatible with NLP tasks. The sequence of steps is not necessarily the same, due to widely available NLP tasks. This work advocates the use of tokenization and lemmatization steps as pre-processing sequences. The former will tokenize each word from a sentence; thus, each token represents each word. The latter converts each token into its base form, such as caring to care. Moreover, it considers the context of terms, whether they are used as nouns or verbs; for instance, am, are and is all represented as be (G. Liu et al., 2021). This also differentiates

lemmatization from stemming, which does not consider the context in a sentence, leading to the wide acceptance of the former.

5.2.1.2 Annotation

To train the NER model to extract features from textual data are to annotate or label textual information after pre-processing it. The pre-trained NER model consists of real-world objects such as a person, geopolitical entity, location, quantity, date and time (Partalidou et al., 2019). Annotation is a task to tag a dataset with a set of labels for entities based on the aim of the study. For instance, this study requires relevant information about an incident's causation. Therefore, three labels were selected: caution, causal/contributing factor (CF) and failure scenario. Caution labels help to understand CF's positive/negative impact and the presence of abnormality based on textual data. CF denotes where abnormality exists that resulted in an incident. This can be performed manually or automated. The present study performed the former method of tagging the dataset using the prodigy tool, a web-based annotation tool (Honnibal & Montani, 2021b). However, if readers wish to automate the process, it is possible to use the same tool and rely on the previous set of annotated data for annotation suggestions. The ner.manual command invokes the annotation process via the Mac terminal (Honnibal & Montani, 2021a). For more details on the annotation process, readers can refer to (Grosman et al., 2020; Honnibal & Montani, 2021b).

5.2.1.3 Named entity recognition model for feature extraction

Annotations performed in the previous step are exported to the spaCy library to train the NER model with defined entity labels. SpaCy has four deep learning stages: embed, encode, attend and predict. Firstly, tokens are converted into integer ID in which extraction of hash values based on shape, prefix, suffix and form, matching a similar word with their vectors, are

embedded, followed by encoding with context through a convolutional neural network, resulting in a matrix vector. At the attending stage, the developed matrix passes through a convolutional neural network, generating a query vector for predicting a class of entity at the final stage (Partalidou et al., 2019). Training can be initiated through a Mac terminal or Windows command prompt, and spaCy will save the best and last trained model to the system's library for feature extraction.

5.2.2 Interpretative Structural Model (ISM)

5.2.2.1 Relationship between each pair of Zij

This study's factors for the ISM process are identified from the NER model, unlike previous studies that employed the ISM method and relied on a literature review to elicit input factors (Huang et al., 2020; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019). A pairwise relation is used to develop a relationship among factors. These relationships can be expressed in terms of yes or no. For example, if A influences B, it is represented as yes, whereas if B does not influence A, then it is described as no. Expert opinion based on previous incidents and process safety knowledge serves as inputs to develop the relationship among factors.

5.2.2.2 Structural self-interaction matrix (SSIM)

Based on the contextual relationship developed in the previous step, an SSIM can be developed to depict a directed relationship between each pair of Zij. Predefined variables (V, A, X, O) express directed relationships. V denotes i is influencing j, but the opposite is not valid, whereas A denotes j's influence on i, but the reverse is invalid. When i and j influence each other, this can be characterized by X. A case where there is no relationship between them is represented by O. These predefined variables are used for a binary matrix. Similarly, a tertiary matrix can be developed using a similar approach (Sajid et al., 2017).

5.2.2.3 Reachability matrix

A reachability matrix (RM) consists of relationships developed in the SSIM step in the binary form. The relationship is described and limited to 0 or 1 from the SSIM step. The matrix entry can be formulated as follows:

1. If SSIM is V then the (i,j) entry is 1 and (j,i) entry is 0
2. If SSIM is A then (j,i) entry is 1 and (i,j) becomes 0
3. If SSIM is X then (i,j) and (j,i) both entries becomes 1
4. If SSIM is O then (i,j) and (j,i) both entries becomes 0

5.2.2.4 Final reachability matrix (FRM)

An assumption in the ISM method is the introduction of transitivity. Once an initial reachability matrix is developed in the previous step, transitivity is checked. Consider three factors, management oversight, organizational culture and lack of procedure and training. When management oversight influences organizational culture and organizational culture influences lack of procedure and training, these are incorporated in the initial reachability matrix as 1. In contrast, an indirect relationship of management oversight influencing lack of procedure and training through organizational culture is represented as 1*, called transitivity. All these indirect relationships are addressed and denoted by 1* in the reachability matrix, resulting in a final reachability matrix (FRM).

5.2.2.5 Partitioning of FRM

Partitioning FRM is essential due to the assignment of levels to each factor. The reachability set, R(Mi), and antecedent set, A(Mi), are developed from FRM. The former consists of the factor itself and other factors i that are influenced (factors i in the row of FRM), whereas the latter consists of the factor itself and other factors that influenced it (factors i in the column of

FRM). The intersection of a reachability set, R(Mi), and the antecedent set, A(Mi), i.e., R(Mi)∩A(Mi), is also derived for all the factors. The factor for which the intersection of R(Mi) and R(Mi)∩A(Mi) is the same occupies the first level (top level). The first level factor would not influence other factors. Partitioning FRM is an interactive process; once a factor is identified, it is removed from the pool of factors in the subsequent interaction. The same steps are repeated until all factors are allocated to their levels. These levels determine the hierarchy in the ISM process.

5.2.2.6 Converting partitioned FRM to conical matrix

A conical matrix is developed by rearranging factors from FRM according to their associated levels. Therefore, all factors associated with each level are pooled together and result in a conical matrix in ascending order of level, with the upper half from the diagonal consisting of factors with the most zeroes while the other half is unitary (1).

5.2.2.7 Digraph and removing transitivity

In this step, a digraph is developed based on the conical matrix. The structure will consist of each factor in the conical matrix and the factors that influenced it. If factor A affects another factor, B, represented as 1 in the conical matrix, it can be identified, and a directed arc is developed from the former to the latter. Similarly, all the factors influencing B are represented through a directed arc, and the process continues until all factors' relations are considered in the digraph. Also, the transitivity introduced in FRM is removed in this step. The resulting structure depicts a visual structure of complex interrelationships among all factors.

5.2.3 Quantitative reasoning using Bayesian Network (BN)

5.2.3.1 Mapping digraph into BN

A BN is a probabilistic technique representing an accident scenario from causal factors to consequences. It has the flexibility to represent complex model interrelationships using a conditional probability table. A BN with a very complex nodal structure is simplified to a sub-network hierarchy called Object-Oriented Bayesian Network (OOBN) so that it is easy to follow. A recent study (Saeed et al., 2022) used OOBN to estimate an idea as compound as cell death in polar cod. The ISM technique provides a digraph for generalized caution from factors extracted using the NER model. However, it is important to quantify the resulting structure to understand each factor's influence and accident pathways of scenarios. Various studies employing the ISM method leveraged BN as a possible way to model complex interrelationships and estimate failure likelihood (Huang et al., 2020; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019). A quantitative relationship among nodes (factors) is represented using a conditional probability table (CPT). A joint probability distribution P(D) of a random variable D= {D1, …., Dn} is given as

$$P(D) = \prod_{i=1}^{n} P\left(D_i \middle| P_{x(D_i)}\right), \quad (1)$$

where Dx(Ai) is the parent of the random variable Di (Pearl, 1988b).

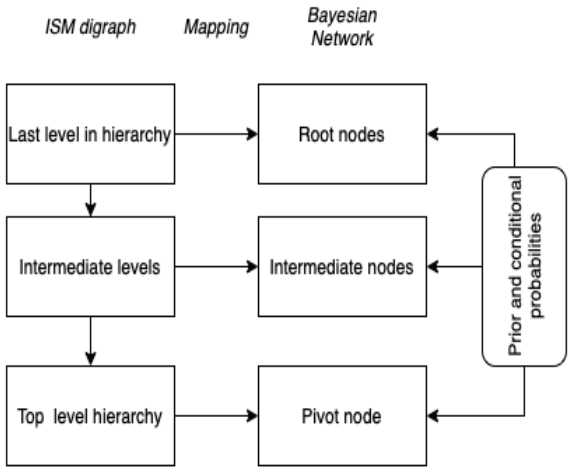


Figure 5-3 Mapping algorithm from ISM into BN

BNs are popular techniques for causal analysis in safety and risk engineering due to their ability to model cause-effect relationships or dependencies among factors. A node in BN represents a random variable, and a direct arc is drawn from a parent node to a child node to model the influence of the former on the latter. Moreover, a BN benefits due to Bayes' theorem, which gives it an advantage for a forward and backward propagation. The former propagation is used for predictive analysis based on the marginal probabilities of intermediate nodes and leaf nodes, according to the CPT. When evidence is provided to the BN model, it updates prior probabilities and reflects more accurate results (Li et al., 2019). In backward propagation, the state of a pivotal node is instantiated to calculate posterior probabilities of root nodes (Khakzad et al., 2013). Based on the similarities in an ISM digraph and BN, this relationship makes BN an obvious choice for quantitative reasoning from the ISM model, which is evident from the previous studies (Huang et al., 2020; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019). However, one of the core differences based on the structure is that BN can only model a directed acyclic graph representing a problem's joint probability, whereas a digraph resulting from the ISM method can be cyclic or acyclic. Therefore, two rules need to be considered while mapping ISM to BN.

- Elimination of single-parent arcs from ISM digraph
- Check for cyclic relationships in ISM digraph

Nodes represent factors in the digraph, which is also the same in BN. Therefore, nodes and directed arcs can be directly mapped into BN from the digraph obtained from the ISM method. A mapping algorithm is illustrated in Figure 5-3. After developing an equivalent BN model, there needs to be a check for a cyclic relationship among nodes. If a cyclic relationship is found in the structure, a modification in the mapped BN is needed for quantitative reasoning.

Developing a duplicate dummy node can handle the cyclic relationships in a BN. This technique has been employed in previous studies, such as (Amin et al., 2018; Yu & Rashid, 2013). The next step is eliminating single direct arcs on a child node from a parent node. Once both rules are implemented for mapping, the next step is quantifying BN.

5.2.3.2 Generalized causation likelihood model

BN developed from the ISM method requires two parameters, prior probabilities and CPTs. The former determines each factor's failure probability, whereas the latter determines the relationship of a parent node to a child node. It defines the interrelationship of factors causing incident scenarios. Expert judgement can input these parameters in BN (Huang et al., 2020; Li et al., 2019; Wu et al., 2015; Yuan et al., 2019). The resulting model will be a generalized causal likelihood model that can establish complex interrelationships among factors, different accident paths and the influence of each factor to mimic the actual incident condition.

5.3 Application to CSB database (oil and refining - downstream)

This section applies the systematically integrated approach to incidents included in the CSB database. There are thirty-six accidents related to fire and explosion. Of them, eight are connected to chemical distribution, sixteen are attributed to chemical manufacturing, two are drillings, and ten are oil and refining (downstream), that are used to develop a generalized causation likelihood model in the present study. This application aims to understand what we can learn from CSB investigation reports. How much of this learning can assist in foreseen future adverse events? The step-by-step methodology depicted in Figure 5-2 is applied to ten CSB incidents (downstream) to answer these questions.

5.3.1 Development of NER model

In step 5.2.1 of the approach, three reports sections are considered: Executive summary/Abstract, Key findings/root causes and incident/causal analysis. Incident reporting in the CSB databases is not consistent in terms of contents. Every report does not have the same sections. Therefore, a combination of sections is considered to overcome this challenge and extract relevant information that can be used for causation modelling. After their selection, these sections are processed in tokenization and lemmatization. Textual data are separated by words in tokenization and converted into their base form considering the context in a sentence. The next step, annotation, is essential in highlighting key entities with their labels. Words such as inoperable, remove, open, close, lack, deviation etc., are used in the "caution" label. Other relevant keywords, such as pressure, temperature, process hazard analysis, corrosion, pressure relief devices etc., are used as "CF" (causal/contributing factor). The former entity shows the presence of an abnormality using phrases or words that are not causal or contributing factors but provide more information about them. The latter describes an abnormality responsible for an incident. A combination of these two entities, caution and CF, helps to analyze previous incidents. The failure scenario is fire & explosion resulting from oil and refining incidents. The annotation process can be time-consuming and depends upon the size of the corpus. Annotation is followed by training convolutional neural network (CNN) for a custom NER model. The spaCy library in Python trains the NER model using entities and labels assigned in the annotation process (Honnibal & Montani, 2021c).

Figure 5-4 Highlighted entities from NER model

An example of NER model output in report 1, the incident of ExxonMobil Refinery Chemical Release and Fire, is shown in Figure 5-4. The identified features can be coupled together and form phrases. For instance, identified entities in the first line, consisting of CF and caution, are coupled together to form a deficiency of the safety management system. Similarly, other features, include reliability issues with gearbox valve, inoperable gearbox, lack of human factor, no written procedure to remove gearbox valve to manually open and close valve, organizational culture- operator remove malfunction valve gearbox, and lack of procedure and training for safe removal, from the shown paragraph. Practitioner domain expertise is helpful in annotation and making sense of extracted features using the NER model.

Table 5-1 shows extracted features from each incident which will serve as a base for the ISM method. An individual causation model can also be developed based on each incident's features; however, this work advocates a generalized causation model to investigate the similarities and influence of each CF in causing fire & explosion.

Table 5-1 Relevant features from each incident using NER model

Report	Name of incident	Causes/contributing factors extracted from NER model	Failure scenario
1	ExxonMobil Refinery Chemical Release and Fire	Deficiency of safety management system	Fire
		Reliability issue with gearbox valve	
		Inoperable gearbox	
		Human factor	
		No written procedure to remove gearbox valve to manually open and close valve	
		Organizational culture- operator remove malfunction valve gearbox	
		lack of procedure and training for safe removal	
		vapor cloud ignition	
2	Delaware City Refining Company	Nonroutine equipment maintenance preparation activity	Fire
		Lack of procedure for non-routine work	
		Leak from valve	
		vapor ignition	
		Human factor- open the valve, operator failed to recognise drop in pressure level	
3	ExxonMobil Refinery	Process safety management system allows without pre safe operating limits	Explosion
		Isolating equipment for maintenance which caused pressure deviation	

	Explosion	pressure deviation caused due to maintenance activity	
		Safety critical safeguard effectiveness	
		Hazard analysis for procedure	
		Lack of safety instrumentation	
		Detect management permit issue	
		Spark - ignition source	
4	Chevron Refinery Fire	High temperature	Fire
		Corrosion	
		Corrosion prevention safeguard effectiveness	
		Not implement recommendations to prevent corrosion failure - organizational culture	
		Lack of safety culture	
		substandard equipment maintenance	
		Vapor cloud ignition	
5	Tesoro Refinery Fatal Explosion and Fire	Rupture due to High temperature hydrogen attack (HTHA)	Fire & explosion
		Leak due to nonroutine hazardous startup activity	
		Lack of mechanical integrity program	
		Weak and ineffective safeguard	
		Lack of process hazard analysis for startup activity	
		Ineffective control and prevent equipment due to HTHA	
		Ineffective leak prevention from flange and gasket	
		Process safety culture deficiency	
		HTHA operating condition safety effectiveness	
		Human factor - operator during startup	

6	Valero Refinery Propane Fire	No formal written procedures for freeze protect dead-legs	Fire
		No Emergency isolation of valve procedure	
		No sufficient distance for fireproof in handling high pressure	
		Ineffective PHA	
		No freeze protection program for freeze hazard for dead leg	
		Lack of remote isolation of valve - prevent operator to close valve or pump control to control high pressure	
		Management of change (MoC)	
		Chlorine release from crack	
		High pressure	
		ignition	
7	BP America Refinery Explosion	Regulatory oversight	Fire & explosion
		Ineffective safety culture (No effective reporting and learning culture)	
		No effective accident prevention plan	
		Impaired process safety performance	
		No flare	
		Lack of automate control to prevent unsafe level	
		Lack of supervisory oversight and technically train for startup and hazardous operations	
		Hazardous startup	
		Inadequate instrumentation for overfilling	

		Lack of effective mechanical integrity program	
		No vehicle traffic policy	
		operator training program inadequate	
		No pre startup safety review	
		Ineffective procedure for operational problem during startup	
		No relief valve system safety	
		vapor cloud ignition	
8	Giant Industries Refinery Explosions and Fire	Corrosion lead to fouling or scoring of pump seal	Fire & explosion
		Ineffective mechanical integrity program to prevent corrosion	
		Valve was open or close by gear but removed and replaced with valve wrench	
		Operator decide valve wrench	
		Release	
		Ignition	
9	Tosco Avon Refinery Petroleum Naphtha Fire	Nonroutine work	Fire
		removal of closed valve leak	
		Work permit - not identify ignition hazard	
		level control valve leaked	
		Evaluation of work as low maintenance	
		Supervisor not involved in permit and lack of supervisory oversight during safety critical activity	
		Corrosion in valve	

		Work permit - not identify ignition and hazardous material	
		Management oversight does not detect deficiency in maintenance and process change	
		No MOC review of operational change which led to corrosion	
		Corrosion control program is inadequate	
		Leak	
		Ignition	
10	Sonat Exploration Co. Catastrophic Vessel Over pressurization	High pressure	Fire
		Pressure relief device and over pressurization lead to failure	
		Ineffective engineering design review and PHA	
		No written operating procedure for startup	
		Lack of adequate pressure relief system	
		Ignition	
		Lack of valve	

5.3.2 Establishing hierarchy and interrelationships among factors

Step 5.2.1 of the approach is based on the role of the ISM method in establishing hierarchical structure and interrelationships in identified features. The ISM method requires factors to establish the interrelationships among them. These factors from ten incidents are used from step 5.2.1 output. Many factors are common in ten incidents; all the factors are summarized in Table 5-2. The output from NER serves as the input for the ISM method. These factors are the basis for the ISM method.

Table 5-2 List of factors from NER output and their probabilities

Serial number	Factors from NER model	Prior probability
1	Human factor	OR gate
2	Organizational culture	OR gate
3	Reliability issue with valve	1.50E-02
4	substandard equipment maintenance	OR gate
5	vapor cloud	OR gate
6	Leak due to nonroutine maintenance activity	5.00E-02
7	Leak from valve	OR gate
8	Safety critical safeguard ineffectiveness	OR gate
9	Lack of safety instrumentation	1.00E-02
10	Ignition source- spark or vehicle	OR gate
11	Abnormal operating conditions (temperature/pressure)	OR gate
12	Corrosion	8.00E-03
13	Lack of safety culture	OR gate
14	Rupture due to High temperature hydrogen attack (HTHA)	AND gate
15	Ineffective control and prevention of equipment	OR gate
16	Ineffective leak prevention from flange and gasket	5.00E-02
17	Lack of adequate pressure relief system	1.20E-02
18	No sufficient distance for fireproof in handling high pressure	4.00E-02
19	No freeze protection program for freeze hazard for dead leg	5.00E-02
20	Lack of remote isolation of valve	5.50E-02

21	Regulatory oversight	1.5E-02
22	No effective accident prevention plan	5.00E-02
23	No vehicle traffic policy	2.50E-02
24	Deficiency of PSM	OR gate
25	Lack of procedure and training	9.00E-02
26	No pre startup safety review	1.80E-02
27	Management permit issue	5.00E-03
28	Leak due to nonroutine hazardous startup activity	3.00E-03
29	Lack of mechanical integrity program	7.00E-03
30	Management of change (MoC)	9.00E-02
31	Management oversight	2.00E-02
32	Ineffective engineering design review and PHA	8.00E-03
33	Fire & explosion	OR gate

Firstly, a pair-wise comparison is performed to analyze the influence of each factor on the other. The contextual relationship is developed for each factor in the form of yes or no and is transformed into SSIM, as shown in Appendix Table 5-6. For example, organizational culture affects human factors, substandard equipment maintenance, lack of safety culture, no vehicle traffic policy and lack of procedure and training. Similarly, other factors' influences can be established and seen in the developed SSIM (Appendix Table 5-6). To establish a contextual relationship among factors, expert opinion is taken into consideration due to the data unavailability. The SSIM is converted into an RM using the rules in step 5.2.2.3. In RM the relationship of the factors is converted into binary 0 or 1 instead of V, A, X and O in the SSIM matrix. For the sake of brevity, instead of RM, FRM is shown in Appendix Table 5-7. It consists of the same elements as in RM; transitivity links (indirect) are introduced in FRM and represented as 1* to differentiate them from RM elements. Further, FRM consists of two more

elements, driving power and dependence power. The former is the total number of interactions of each factor in a row. The latter is the total number of interactions of each factor in a column. In other words, the former is the number of factors it affects, whereas the latter is the number of factors which get affected.

FRM facilitates the development of CF levels by partitioning FRM into levels. The output from FRM helps to derive the reachability, R(Mi), and antecedent, A(Mi), sets shown in Appendix Table 5-8. This is an iterative process of levels allocated to each factor. A total of nine iterations were performed to develop the partitioning of FRM into levels. R(Mi) consists of the factor itself and other factors i that are influenced (factor i in the row of FRM). A(Mi) consists of the factor itself and other factors that influenced it (factors i in the column of FRM). The intersection of a reachability set, R(Mi), and antecedent set, A(Mi), i.e., R(Mi)∩A(Mi), is derived for all the factors. The factor for which the intersection of R(Mi) and R(Mi)∩A(Mi) is the same occupies the first level (top-level). In this case, factor 33 is assigned to the top level in the hierarchy and has no factor above it. A conical matrix is developed to better illustrate each factor and its respective level. Table 5-9 clearly shows each factor and its hierarchy to visualize its structure. All null elements are in the upper diagonal of the matrix, whereas the other half consists of unitary elements. All these steps are performed to develop a final digraph.

The conical matrix is used to create a digraph based on the direct relationship of each factor to another. Factor 33 (fire & explosion) occupies the top level in the hierarchy in the digraph, followed by five factors at level II, as shwn in Figure 5-5. In Table 5-9, the 2nd row shows the influence of factor 4 on 33. Therefore, a directed arc is drawn from the former to the latter to represent this relationship. Similarly, factors 5, 8, 10 and 22 also have the same relationship. Fire & explosion has the highest dependence power, i.e., 33, which means every factor in a digraph is, directly and indirectly, leading to fire & explosion. If transitivity is removed, then

only five factors, 4, 5, 8, 10 and 22, directly influence fire & explosion, and other factors indirectly lead to fire & explosion. Factors 21 (regulatory oversight) and 31 (management oversight) have the highest driving power, meaning these have the highest contribution towards affecting other factors, followed by organizational culture, resulting in fire & explosion.

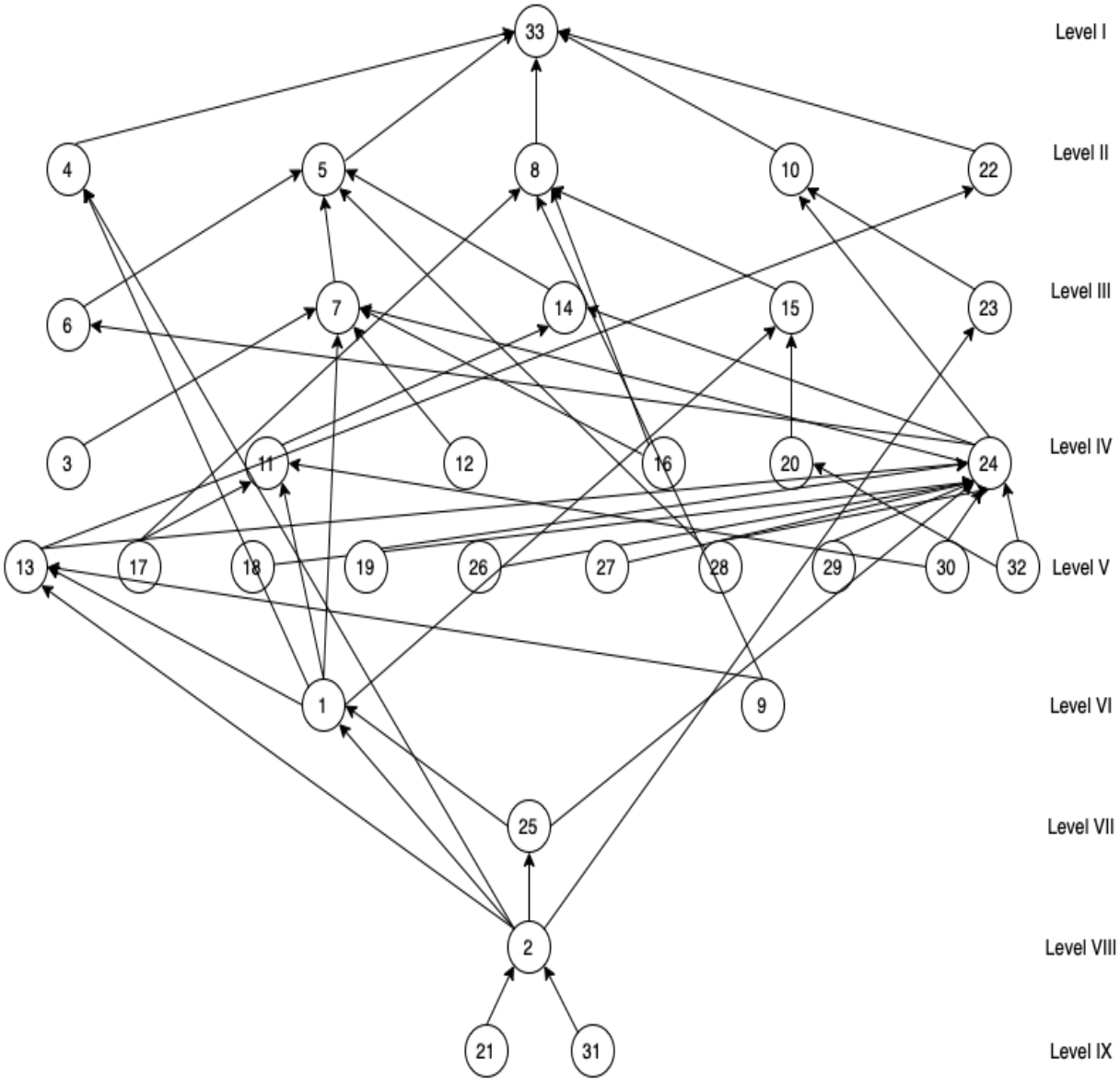


Figure 5-5 Digraph developed from the ISM method

5.3.3 Generalized causation likelihood model

The ISM method provides a generalized causation digraph that creates structures using a systematic approach. However, a qualitative approach can only establish interrelationships among factors. BN is a better fit when quantitative reasoning needs to be performed. Based on the mapping algorithm in Figure 5-3, an equivalent BN is constructed from the ISM digraph. Step 5.2.3.1 of the approach states that single-parent arcs from factors 6, 20, 22, and 25 are eliminated in the resulting BN model. In addition, there are no cyclic relations in the digraph (Figure 5-5); it is compatible with mapping into BN. The resulting BN is shown in Figure 5-6. Two components of BN required to calculate fire & explosion likelihood are prior probabilities and CPTs.

Table 5-2 states prior probabilities of factors and logical gates to model and defines factor's interrelationships.

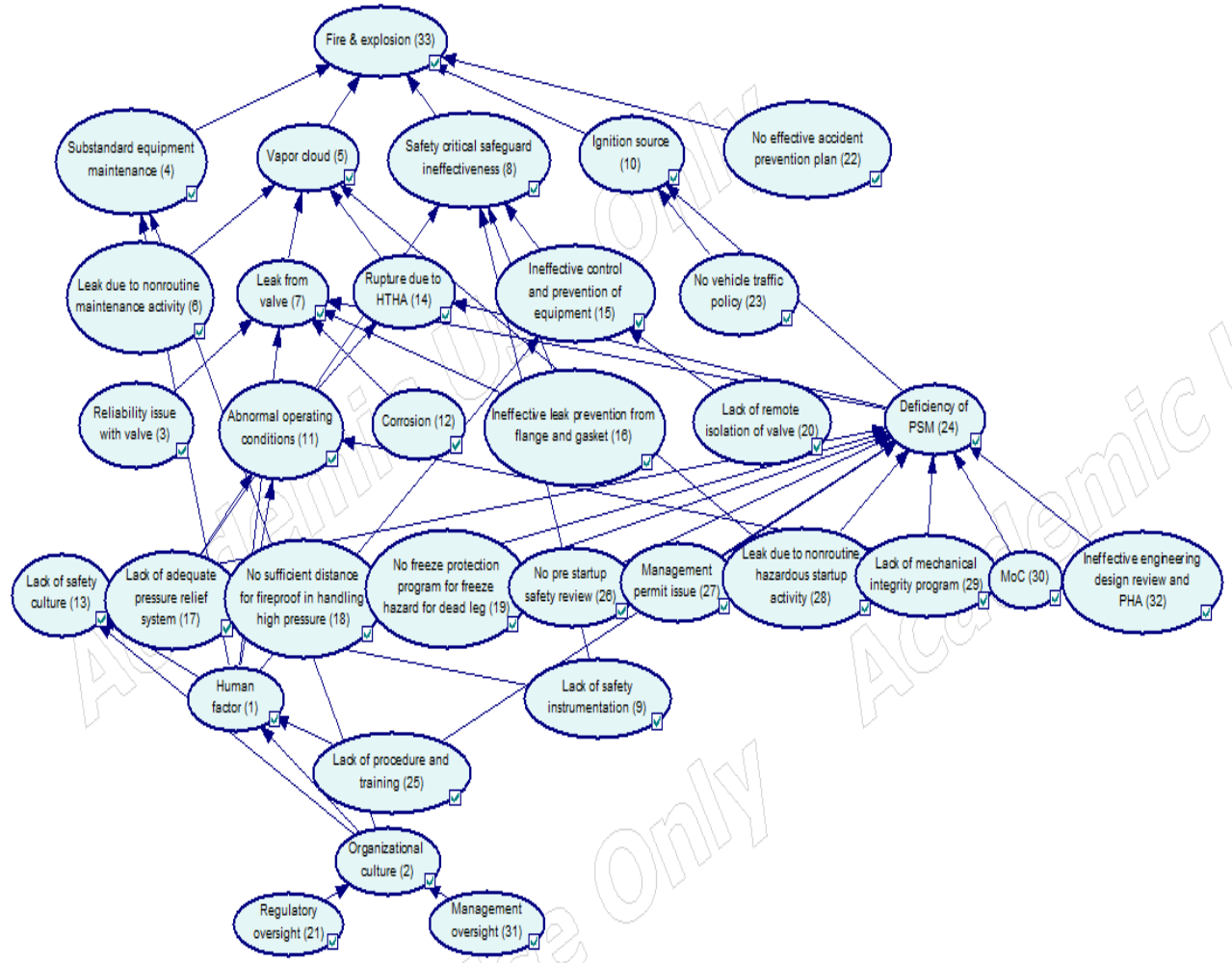


Figure 5-6 BN mapped from the ISM digraph

5.4 Results and Discussion

The integrated approach provides a means to extract causal factors of fire and explosion relevant to oil and refining (downstream) incidents. The extracted factors from all ten incidents are used for developing a hierarchical structure to establish the interrelationship of factors by the ISM process. BN transforms the qualitative hierarchical structure into a quantitative generalized causation model. The developed BN model shown in Figure 5-6 provides a likelihood of 4.72E-01. It shows that when combined through a systematically integrated approach (Figure 5-1), all factors are most likely to cause fire & explosion. Note that factors considered in the present study are from past incidents (oil and refining - downstream) that CSB investigates, coupled with expert opinion, called past experiences. The methodology

depicted in Figure 5-2 provides a good result for the adverse event. It is a likelihood of an event that has already occurred, valid through estimated value. The next step is to test the developed model on seen data and verify unseen data to analyze their prediction of adverse events.

5.4.1 Model testing and verification

The generalized causation model has the advantage of modeling different accident scenarios and understanding potential pathways. Model testing and verification aim to demonstrate its capability and efficacy in predicting adverse events. Table 5-1 is used as a scenario-based testing dataset to see if the approach proposed in this work can model individual accidents. In each scenario, only a few factors will be considered that will cause an adverse event. Therefore, accidents used to develop a generalized model are suitable for testing model performance. In scenario 1, the factors identified using the NER model in Table 5-1 are identified parameters. This work used an existing BN that uses bidirectional propagation introduced by (Pearl, 1988a). When these parameters are given to the BN model in Figure 5-6, the state of mentioned events is 1, reflecting that they have already occurred to estimate fire and explosion likelihood. The model gives a likelihood of 100% (likelihood =1.0), which means that these are sure to occur based on the presented evidence.

Similarly, scenarios 2-10 from Table 5-1 are considered via the developed model to estimate the likelihood of the pivotal node, i.e., 33 (fire and explosion). The result is summarized in Table 5-3, which shows that all the scenarios from the seen dataset are sure to occur based on model prediction. It shows that the methodology used in this study in eliciting data from past experiences (incident reports and expert opinion) can predict individual scenarios used in the development of the generalized causation model. In other words, the methodology is applicable in establishing a generalized causation model.

Table 5-3 Model testing results in estimating fire and explosion likelihood

Scenarios	Model results (%)
1	100%
2	
3	
4	
5	
6	
7	
8	
9	
10	

It is vital to verify the model by analyzing those incident scenarios that are not considered in the development. The model verification is based on incidents from the lessons learned database from IChemE. A book of fifty-two incidents has been published this year, including incidents related to energy, power generation, chemicals, water, food and drink sectors ("Learning Lessons from Major Incidents," 2022). However, our interest is those related to oil and gas (downstream) incidents, since the developed model is based on these kinds of incidents. The book contains fourteen oil and gas-related incidents, selected to find root causes. However, except for six incidents, the rest are considered for the model development from the CSB database. We are considering incidents which do not take part in the development of the generalized causation model.

Table 5-4 Accidents related to oil and gas (downstream) from ("Learning Lessons from Major Incidents," 2022)

Date	Incident name	Country	Type	Root causes
Jan 04 1966	Feyzin	France	BLEVE	Process design, equipment/piping design, protective systems (pressure), hazard awareness, procedures, training, emergency preparedness and design standards
Mar 22 1987	Grangemouth	UK	Explosion	Abnormal operations, escalation potential, hazard identification, equipment/piping design, instrumentation, safety instrumented systems, protective systems, hazard awareness, control of work, procedures, training, production over safety, MoC, failure to learn, PSM
Jul 24 1994	Milford Haven	UK	Explosion	Abnormal operations, escalation potential, equipment/piping design, instrumentation, process monitoring, alarm management, operational risk assessment, preventative maintenance, inspection, material degradation, control of work, human factor, communication, training, MoC, emergency preparedness, PSM
Aug 17 1999	Izmit	Turkey	Fire	Escalation potential, process design, equipment/piping design, materials of construction, protective systems, plant layout, operational risk assessment, material

				degradation, human factors, communication, training, emergency preparedness, PSM, regulatory compliance audits
Apr 16 2001	Humber	UK	Explosion	Escalation potential, hazard identification, equipment/piping design, preventative maintenance, inspection, material degradation, communication, MoC, PSM
Mar 11 2011	Chiba	Japan	BLEVE	Equipment/piping design, protective design, plant layout, creeping change, operational risk assessment, inspection, work planning, training, emergency preparedness, PSM and land use planning

Table 5-4 shows all six incidents considered for model verification purposes. The fifth column of Figure 5-4 states all the root causes that IChemE has identified from respective incidents. These root causes served as evidence shown in Table 5-5 to estimate the model prediction for each scenario. Let us consider the Feyzin incident from Table 5-4. Factors like process design, equipment/piping design and hazard awareness are taken into consideration - Ineffective engineering design review and PHA (32), protective systems (pressure) - lack of adequate pressure relief system (17), procedures, training - lack of procedure and training (25), emergency preparedness - organizational culture (2), design standards - regulatory oversight (21). Hence, Table 5-5, row 2 shows all the factors (32,17,25,2,21) reflecting the Feyzin incident. When the same condition based on the IChemE book is given to the model, this results in a 100% likelihood of fire & explosion based on the provided evidence (factors probability =1). Similarly, evidence related to other incidents in Table 5-4 is input into the model, as shown

in the 2nd column of Table 5-5, to reflect each incident scenario. In all cases, model predictability remains at 100%. This result shows that an incident will happen in real life when a condition is certain to occur. Therefore, this exercise is called verification; the model is verified based on the IChemE database's observed evidence of six incidents used here. These preliminary model testing and verification results show the promising efficacy of using the three-step approach to learn from past experiences.

Table 5-5 Model verification through unseen incident evidence

Scenarios	Evidence	Model results (%)
1	32,17, 25,2,21	100
2	11,32,9,4,15,25,30,24,2	
3	11, 32,9,15,4,1,25,30,24 7,17,30	
4	32,15,1,25,24,21	
5	32,4,30,24	
6	32,15,25,24,21	

5.4.2 Sensitivity Analysis

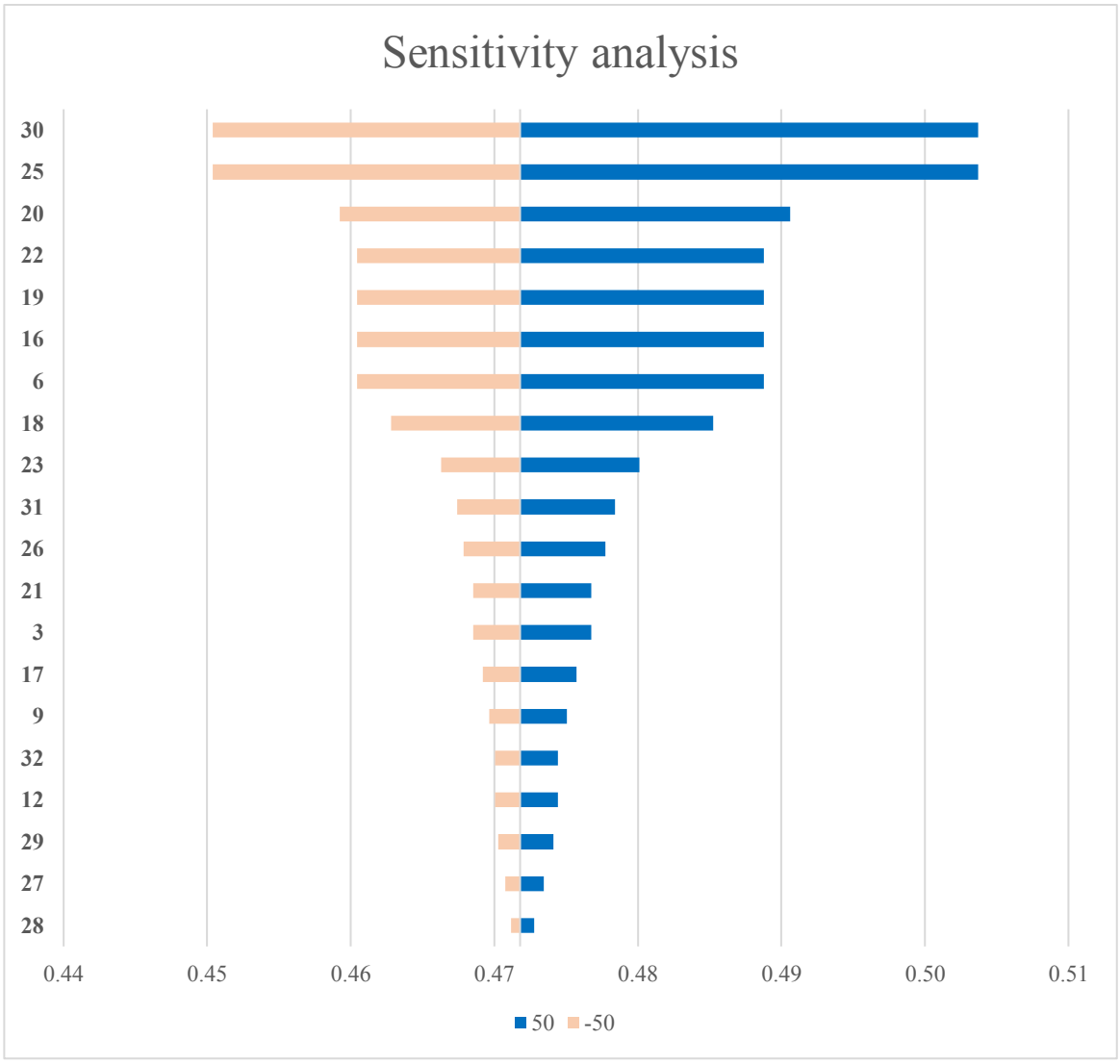


Figure 5-7 Tornado chart developed for sensitivity analysis of each factor

Sensitivity analysis is performed to investigate the sensitivity of factors towards fire & explosion. A percentage change ±50 of each root node is done one by one, and a shift in fire & likelihood is recorded. Based on the observation, a tornado is developed, as shown in Figure 5-7, to illustrate its effect on the pivotal node. Management of Change (MoC) and lack of procedure and training have the highest sensitivity towards the adverse event, followed by organizational culture. If we consider IChemE incidents related to oil and gas (downstream) ("Learning Lessons from Major Incidents," 2022), out of fourteen incidents, seven were due to

MoC as one of the root causes, whereas lack of procedure and training was involved in eleven incidents. Hence, it shows that the most sensitive factors based on sensitivity analysis are valid through the IChemE incidents root cause map.

This study introduces a unique, systematically integrated approach to modelling incidents from the CSB database. This study is not free from assumptions. Each step consists of a method: using NER to extract and characterize factors such as caution and CF to develop phrases. The annotation step in NER requires domain expertise to tag or label a corpus. Secondly, using the ISM method to combine all incident factors into a generalized digraph, the contextual relationship among factors requires domain expertise. Lastly, BN for quantitative reasoning requires domain expertise for probabilities of root nodes and modeling of interrelationships.

Past experiences refer to past incidents' information and domain expertise knowledge to learn lessons and predict or foresee adverse events. This way, information from a database and people's memory in terms of domain expertise can be used to develop a generalized causation model. It will serve as a way for memory retention in an organization. This study describes how lessons learned can develop a generalized causation model. It helps to understand that factors like management oversight and regulatory oversight have the highest driving power among considered incidents. This means they are influencing factors that result in fire or explosion. It provides insight to determine the most influential factor by modelling the same type of accidents. Often in the investigation of incidents, the human factor is identified as a root cause. However, further action is required. Humans will make mistakes; the system should be designed to handle those errors (Halim & Mannan, 2018). There are factors influencing human factors that are often not considered. As Kletz said, "For a long time, people were saying that most accidents were due to human error, and this is true in a sense, but it's not very helpful. It's a bit like saying that falls are due to gravity" (U.S. Chemical Safety and Hazards

Investigation Board, 2013). Also observed by Sydney Dekker, "Underneath every simple, obvious study about 'human error,' there is a deep and more complex story" (Dekker, 2017). This study unfolds a way to analyze human error influenced by organizational culture and lack of procedure and training (Figure 5-6). Organizational culture plays a vital role in human performance and behaviour in a workplace. Poor safety culture (a subset of organizational culture) led to many incidents (U.K. Health and Safety Executive, 2022). This study found two factors influencing organizational culture: management oversight and regulatory oversight. These include their interrelationships with factors like organizational culture, process safety management, human factor, and their hierarchy in a structure. More emphasis must be given to MoC and inadequate procedure and training in an organization based on sensitivity analysis findings and the root cause map from IChemE ("Learning Lessons from Major Incidents," 2022).

Combining three methods provides a unique avenue for analyzing similar types of oil and refining accidents together. When accidents are analyzed, many similarities are found that were not considered in the past. Many factors, such as human error, safety culture, ineffective prevention and control, and ineffective process hazard analysis, are common and need proper attention.

The model performance has been demonstrated using the IChemE incidents related to oil and refining. The results (see Table 5-5) suggest that the model can predict similar incidents, which proves the need for considering causation factors from historical accident databases while developing an efficient accident causation algorithm.

The sensitivity analysis provides two important factors, management of change and lack of procedure and training, which are found as CF in many accidents. The root cause map of IChemE ("Learning Lessons from Major Incidents," 2022) also highlighted these two as

common causes of accidents related to oil and refining; special attention must be given to these two factors in the process industries.

5.5 Conclusions

This chapter introduces a methodology for learning from past experiences to predict adverse events. Repetition of incidents and causal/contributing similarities depicts our failure to learn from them and implement those learnings (Halim & Mannan, 2018). Failure to learn from incidents is due to the unavailability of an efficient methodology that can systematically analyze incident reports in the CSB and other databases. The current work addresses this issue by using NER, ISM and BN to develop a generalized causation model. This model assists in unfolding critical common factors that influence similar incidents and their potential pathways. It also highlights the importance of previously developed resources that can be better used to manage the risk posed by hazards in chemical processing industries. The unique aspects of the study are as follows:

1. Provision of an easy-to-implement three-step approach to elicit information from the CSB database and memory from domain expertise to predict adverse events.
2. Insight into the complex interrelationships of extracted factors and accident pathways in a hierarchical structure.
3. Identification of MoC and lack of procedure and training as having the highest sensitivity towards fire and explosion causation.
4. Identification of management and regulatory oversights as having the highest driving power to influence other factors.
5. Strategies to manage and reduce the likelihood of accidents, which are developed by resource allocation based on the hierarchy of factors.
6. A requirement for minimal intervention in terms of domain expertise to adopt the approach in other incident types or domain-specific work.

The study attempts to develop a generalized causation model for incidents related to oil and gas refining (downstream) operations. The model is tested on ten incidents, verified on six incidents, and able to predict all incidents with 100% likelihood, i.e., coherent with actual conditions. It provides insights into developing strategies and policymaking to avoid future incidents. Furthermore, it highlights that existing resources in terms of the database are an excellent source of learning if appropriately utilized with domain expertise to develop a causation model to foresee future incidents. There are many similarities in causal/contributing factors which cause similar types of incidents, as seen in the CSB case studies. The methodology can be applied to different types of incidents available in the CSB database to develop generalized causation for each incident type and to compare each model with another to comprehend similarities between them. The model also has limitations in terms of uncertainty handling. Uncertainties arise from a lack of data, with expert opinions being a primary source of data uncertainty. Additionally, using OR/AND logic gates also introduces model uncertainty. Therefore, handling uncertainties in a generalized causation model and fusion of historical numerical data with an accident database can be a direction for future work.

5.6 Acknowledgements

The authors acknowledge the financial support provided by Genome Canada and their supporting partners through the Large Scale Applied Research Project; the Canada Research Chair (CRC) Tier I Program in Offshore Safety and Risk Engineering; and the Mary Kay O' Connor Process Safety Center at Texas A&M University, TX, USA.

5.7 References

1. Amin, M. T., Imtiaz, S., & Khan, F. (2018). Process system fault detection and diagnosis using a hybrid technique. *Chemical Engineering Science*, *189*, 191–211. https://doi.org/10.1016/j.ces.2018.05.045
2. Amyotte, P., Irvine, Y., & Khan, F. (2018). Chemical safety board investigation reports and the hierarchy of controls: Round 2. *Process Safety Progress*, *37*(4), 459–466. https://doi.org/10.1002/prs.12009
3. Amyotte, P. R., Berger, S., Edwards, D. W., Gupta, J. P., Hendershot, D. C., Khan, F. I., Mannan, M. S., & Willey, R. J. (2016). Why major accidents are still occurring. *Current Opinion in Chemical Engineering*, *14*, 1–8. https://doi.org/10.1016/j.coche.2016.07.003
4. Amyotte, P. R., & Khan, F. I. (2021). The role of inherently safer design in process safety. *Canadian Journal of Chemical Engineering*, *99*(4), 853–871. https://doi.org/10.1002/cjce.23987
5. Amyotte, P. R., Khan, F. I., & Kletz, T. A. (2009). INHERENTLY SAFER DESIGN ACTIVITIES OVER THE PAST DECADE. *Proceedings of the IChemE Symposium Series No.155*, 736–743.
6. Amyotte, P. R., Macdonald, D. K., & Khan, F. I. (2011). An analysis of CSB investigation reports concerning the hierarchy of controls. *Process Safety Progress*, *30*(3), 261–265. https://doi.org/10.1002/prs.10461
7. Attri, R., Dev, N., & Sharma, V. (2013). Interpretive Structural Modelling (ISM) approach: An Overview. In *Research Journal of Management Sciences* (Vol. 2, Issue 2). www.isca.in

8. Baybutt, P. (2016). Insights into process safety incidents from an analysis of CSB investigations. *Journal of Loss Prevention in the Process Industries*, *43*, 537–548. https://doi.org/10.1016/j.jlp.2016.07.002
9. Chemical Safety and Hazard Investigation Board. (2022, March 10). https://www.csb.gov/investigations/completed-investigations/?Type=2
10. Dekker, S. (2017). *The field guide to understanding 'human error'''* (3rd ed.). CRC press.
11. Fu, G., Xie, X., Jia, Q., Li, Z., Chen, P., & Ge, Y. (2020). The development history of accident causation models in the past 100 years: 24Model, a more modern accident causation model. *Process Safety and Environmental Protection*, *134*, 47–82. https://doi.org/10.1016/j.psep.2019.11.027
12. Grosman, J. S., Furtado, P. H. T., Rodrigues, A. M. B., Schardong, G. G., Barbosa, S. D. J., & Lopes, H. C. V. (2020). Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*. https://doi.org/10.1016/j.is.2020.101553
13. Haghighattalab, S., Chen, A., Fan, Y., & Mohammadi, R. (2019). Engineering ethics within accident analysis models. *Accident Analysis & Prevention*, *129*, 119–125. https://doi.org/https://doi.org/10.1016/j.aap.2019.05.013
14. Halim, S. Z., & Mannan, M. S. (2018). A journey to excellence in process safety management. *Journal of Loss Prevention in the Process Industries*, *55*, 71–79. https://doi.org/10.1016/j.jlp.2018.06.002
15. Honnibal, M., & Montani, I. (2021a). *Prodigy*. https://prodi.gy/docs/recipes#ner-manual
16. Honnibal, M., & Montani, I. (2021b). *Prodigy · An annotation tool for AI, Machine Learning & NLP*. https://prodi.gy/

17. Honnibal, M., & Montani, I. (2021c). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. In *To appear* (3.0). MIT. https://spacy.io/
18. Huang, W., Zhang, Y., Kou, X., Yin, D., Mi, R., & Li, L. (2020). Railway dangerous goods transportation system risk analysis: An Interpretive Structural Modeling and Bayesian Network combining approach. *Reliability Engineering and System Safety*, *204*. https://doi.org/10.1016/j.ress.2020.107220
19. IChemE *Safety and Loss Prevention*. (2022). https://www.icheme.org/membership/communities/special-interest-groups/safety-and-loss-prevention/resources/lessons-learned-database/
20. Kamil, M. Z., Khan, F., Song, G., & Ahmed, S. (2019). Dynamic Risk Analysis Using Imprecise and Incomplete Information. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, *5*(4). https://doi.org/10.1115/1.4044042
21. Kaszniak, M. (2010). Oversights and omissions in process hazard analyses: Lessons learned from CSB Investigations. *Process Safety Progress*, *29*(3), 264–269. https://doi.org/10.1002/prs.10373
22. Khakzad, N., Khan, F., & Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. *Process Safety and Environmental Protection*, *91*(1–2), 46–53. https://doi.org/10.1016/j.psep.2012.01.005
23. Khan, F. I., & Amyotte, P. R. (2007). Modeling of BP Texas City refinery incident. *Journal of Loss Prevention in the Process Industries*, *20*(4–6), 387–395. https://doi.org/10.1016/j.jlp.2007.04.037

24. Knegtering, B., & Pasman, H. J. (2009). Safety of the process industries in the 21st century: A changing need of process safety management for a changing industry. *Journal of Loss Prevention in the Process Industries*, *22*(2), 162–168. https://doi.org/10.1016/j.jlp.2008.11.005

25. la Torre, F., Meocci, M., Domenichini, L., Branzi, V., Tanzi, N., & Paliotto, A. (2019). Development of an accident prediction model for Italian freeways. *Accident Analysis & Prevention*, *124*, 1–11. https://doi.org/https://doi.org/10.1016/j.aap.2018.12.023

26. Learning lessons from major incidents. (2022). In *IChemE Safety & Loss Prevention Special Interest Group* (Centenary edition, Issue Centenary edition). https://www.icheme.org/media/18415/learning-lessons-from-major-incidents-v10.pdf

27. Li, F., Wang, W., Dubljevic, S., Khan, F., Xu, J., & Yi, J. (2019). Analysis on accident-causing factors of urban buried gas pipeline network by combining DEMATEL, ISM and BN methods. *Journal of Loss Prevention in the Process Industries*, *61*, 49–57. https://doi.org/10.1016/j.jlp.2019.06.001

28. 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. https://doi.org/https://doi.org/10.1016/j.psep.2021.05.036

29. Mannan, M. S., & Waldram, S. P. (2014). Learning lessons from incidents: A paradigm shift is overdue. *Process Safety and Environmental Protection*, *92*(6), 760–765. https://doi.org/10.1016/j.psep.2014.02.001

30. Olive, C., O'Connor, T. M., & Mannan, M. S. (2006). Relationship of safety culture and process safety. *Journal of Hazardous Materials*, *130*(1-2 SPEC. ISS.), 133–140. https://doi.org/10.1016/j.jhazmat.2005.07.043
31. Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., & Diamantaras, K. I. (2019). Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy. *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*. https://doi.org/10.1145/3350546.3352543
32. Pasman, H. J., Rogers, W. J., & Mannan, M. S. (2018). How can we improve process hazard identification? What can accident investigation methods contribute and what other recent developments? A brief historical survey and a sketch of how to advance. In *Journal of Loss Prevention in the Process Industries* (Vol. 55, pp. 80–106). Elsevier Ltd. https://doi.org/10.1016/j.jlp.2018.05.018
33. Pearl, J. (1988a). Chapter 4 - BELIEF UPDATING BY NETWORK PROPAGATION. In J. Pearl (Ed.), *Probabilistic Reasoning in Intelligent Systems* (pp. 143–237). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-08-051489-5.50010-2
34. Pearl, J. (1988b). Probabilistic Reasoning in Intelligent Systems. In *Morgan Kauffmann San Mateo* (Vol. 88, p. 552). https://doi.org/10.2307/2026705
35. Pipeline and Hazardous Materials Safety Administration. (2022). https://www.phmsa.dot.gov/incident-reporting
36. Rathnayaka, S., Khan, F., & Amyotte, P. (2011). SHIPP methodology: Predictive accident modeling approach. Part I: Methodology and model description. *Process Safety and Environmental Protection*, *89*(3), 151–164. https://doi.org/10.1016/j.psep.2011.01.002

37. Reuters. (2022a). At Least 10 People Killed in India Factory Explosion. https://www.reuters.com/world/india/least-six-killed-india-chemical-factory-explosion-2022-06-04/
38. Reuters. (2022b, May 13). Six People Confirmed Dead in Slovenia Chemical Plant Blast. https://www.reuters.com/world/europe/six-people-confirmed-dead-slovenia-chemical-plant-blast-2022-05-13/
39. Saeed, M. S., Halim, S. Z., Fahd, F., Khan, F., Sadiq, R., & Chen, B. (2022). An ecotoxicological risk model for the microplastics in arctic waters. *Environmental Pollution*, *315*, 120417. https://doi.org/https://doi.org/10.1016/j.envpol.2022.120417
40. Sajid, Z., Khan, F., & Zhang, Y. (2017). Integration of interpretive structural modelling with Bayesian network for biodiesel performance analysis. *Renewable Energy*, *107*, 194–203. https://doi.org/10.1016/j.renene.2017.01.058
41. Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Model. *Science and Technology Libraries*, *39*(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479
42. The Guardian. (2022, June 2). *Thousands Evacuated after US Fertilizer Plant Fire Sparked Fears of Explosion*. https://www.theguardian.com/us-news/2022/feb/04/north-carolina-fertilizer-plant-fire-explosion-evacuation
43. The New York Times. (2022, June 7). *Firefighters Unaware of Chemicals at Bangladesh Depot, Official Says*. https://www.nytimes.com/2022/06/07/world/asia/bangladesh-fire-depot.html
44. U.K. Health and Safety Executive. (2022). Organisational Culture. https://www.hse.gov.uk/humanfactors/topics/culture.html

45. U.K. HSE, *Investigating accidents and incidents*. (2004). https://www.hse.gov.uk/pubns/hsg245.pdf
46. U.S. Chemical Safety and Hazard Investigation Board. (2022). Incident Reporting Rule Submission Information and Data. https://www.csb.gov/news/incident-report-rule-form-/
47. U.S. Chemical Safety and Hazards Investigation Board. (2013). Video Excerpts from Dr. Trevor Kletz. http://www.csb.gov/videos/csb-video-excerpts-from-drtrevor- kletz/
48. U.S. News. (2022a, January 26). *6 Injured in Explosion at Louisiana Chemical Plant*. https://www.usnews.com/news/us/articles/2022-01-26/6-injured-in-explosion-at-louisiana-chemical-plant
49. U.S. News. (2022b, March 25). *4 Companies Cited for Louisiana Chemical Plant Explosion*. https://www.usnews.com/news/best-states/louisiana/articles/2022-03-25/4-companies-cited-for-louisiana-chemical-plant-explosion
50. Warfield, J. N. (1974). Developing Interconnection Matrices in Structural Modeling. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-4*(1), 81–87. https://doi.org/10.1109/TSMC.1974.5408524
51. Wu, W. S., Yang, C. F., Chang, J. C., Château, P. A., & Chang, Y. C. (2015). Risk assessment by integrating interpretive structural modeling and Bayesian network, case of offshore pipeline project. *Reliability Engineering and System Safety*, *142*, 515–524. https://doi.org/10.1016/j.ress.2015.06.013
52. Yu, J., & Rashid, M. M. (2013). A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, *59*(7), 2348–2365. https://doi.org/10.1002/aic.14013

53. Yuan, C., Cui, H., Ma, S., Zhang, Y., Hu, Y., & Zuo, T. (2019). Analysis method for causal factors in emergency processes of fire accidents for oil-gas storage and transportation based on ISM and MBN. *Journal of Loss Prevention in the Process Industries*, *62*. https://doi.org/10.1016/j.jlp.2019.103964

Appendix

Table 5-6 Structural self-interaction matrix (SSIM) developed by performing pair-wise comparison

Factors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	X	A	O	V	O	O	V	O	O	O	V	O	V	O	V	O	O	O	O	O	O	O	O	O	A	O	O	O	O	O	O	O	O
2		X	O	V	O	O	O	O	O	O	O	O	V	O	O	O	O	O	O	O	A	O	V	O	V	O	O	O	O	O	A	O	O
3			X	O	O	O	V	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
4				X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	V
5					X	A	A	O	O	O	O	O	O	A	O	O	O	O	O	O	O	O	O	O	O	O	O	O	A	O	O	O	O	V
6						X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	A	O	O	O	O	O	O	O	O	O	O
7							X	O	O	O	O	A	O	O	O	A	O	O	O	O	O	O	O	A	O	O	O	O	O	O	O	O	O	O
8								X	A	O	O	O	O	O	A	A	A	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	V
9									X	O	O	O	V	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
10										X	O	O	O	O	O	O	O	O	O	O	O	O	A	A	O	O	O	O	O	O	O	O	O	V
11											X	O	O	V	O	O	A	O	O	O	O	O	O	O	O	O	O	O	O	O	A	O	O	O

12											X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
13												X	O	O	O	O	O	O	O	O	V	O	V	O	O	O	O	O	O	O	O
14													X	O	O	O	O	O	O	O	O	O	A	O	O	O	O	O	O	O	O	O
15														X	O	O	O	O	A	O	O	O	O	O	O	O	O	O	O	O	O	O
16															X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
17																X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
18																	X	O	O	O	O	O	V	O	O	O	O	O	O	O	O	O
19																		X	O	O	O	O	V	O	O	O	O	O	O	O	O	O
20																			X	O	O	O	O	O	O	O	O	O	O	O	A	O
21																				X	O	O	O	O	O	O	O	O	O	O	O	O
22																					X	O	O	O	O	O	O	O	O	O	O	V
23																						X	O	O	O	O	O	O	O	O	O	O
24																							X	A	A	A	A	A	A	O	A	O
25																								X	O	O	O	O	O	O	O	O
26																									X	O	O	O	O	O	O	O
27																										X	O	O	O	O	O	O

28																											X	O	O	O	O	O
29																												X	O	O	O	O
30																													X	O	O	O
31																														X	O	O
32																															X	O
33																																X

Table 5-7 Final reachability matrix (FRM)

Factors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	Driving Power
1	1	0	0	1	1*	1*	1	1*	0	1*	1	0	1	1*	1	0	0	0	0	0	0	1*	0	1*	0	0	0	0	0	0	0	0	1*	14

2	1	1	0	1	1*	1*	1*	1*	0	1*	1*	0	1	1*	1*	0	0	0	0	0	0	1*	1	1*	1	0	0	0	0	0	0	0	1*	17
3	0	0	1	0	1*	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	4
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
6	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	3
7	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	3
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
9	0	0	0	0	1*	1*	1*	1	1	1*	0	0	1	1*	0	0	0	0	0	0	0	1*	0	1*	0	0	0	0	0	0	0	0	1*	11
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
11	0	0	0	0	1*	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	4

12	0	0	0	0	1*	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	4
13	0	0	0	0	1*	1*	1*	0	0	1*	0	0	1	1*	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1*	9
14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	3
15	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	3
16	0	0	0	0	1*	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	5
17	0	0	0	0	1*	0	0	1	0	0	1	0	0	1*	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1*	6
18	0	0	0	0	1*	1*	1*	0	0	1*	0	0	0	1*	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1*	8
19	0	0	0	0	1*	1*	1*	0	0	1*	0	0	0	1*	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1*	8

20	0	0	0	0	0	0	0	1*	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1*	4
21	1*	1	0	1*	1*	1*	1*	1*	0	1*	1*	0	1*	1*	1*	0	0	0	0	0	1	1*	1*	1*	1*	0	0	0	0	0	0	0	1*	18
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	2
23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1*	3
24	0	0	0	0	1*	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1*	7
25	1	0	0	1*	1*	1*	1*	1*	0	1*	1*	0	1*	1*	1*	0	0	0	0	0	0	1*	0	1	1	0	0	0	0	0	0	0	1*	15
26	0	0	0	0	1*	1*	1*	0	0	1*	0	0	0	1*	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1*	8
27	0	0	0	0	1*	1*	1*	0	0	1*	0	0	0	1*	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1*	8

28	0	0	0	0	1	1*	1*	0	0	1*	0	0	0	1*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1*	8
29	0	0	0	0	1*	1*	1*	0	0	1*	0	0	0	1*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1*	8
30	0	0	0	0	1*	1*	1*	0	0	1*	1	0	0	1*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1*	9
31	1*	1	0	1*	1*	1*	1*	1*	0	1*	1*	0	1*	1*	1*	0	0	0	0	0	0	1*	1*	1*	1*	0	0	0	0	0	1	0	1*	18
32	0	0	0	0	1*	1*	1*	1*	0	1*	0	0	0	1*	1*	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1*	11
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Dependence Power	5	3	1	6	25	17	20	12	1	18	8	1	7	19	8	1	1	1	1	2	1	8	4	16	4	1	1	1	1	1	1	33		

Table 5-8 Partitioning of FRM

Elements(Mi)	Reachability Set R(Mi)	Antecedent Set A(Mi)	Intersection Set R(Mi)∩A(Mi)	Level
1	1,	1, 2, 21, 25, 31,	1,	6
2	2,	2, 21, 31,	2,	8
3	3,	3,	3,	4
4	4,	1, 2, 4, 21, 25, 31,	4,	2
5	5,	1, 2, 3, 5, 6, 7, 9, 11, 12, 13, 14, 16, 17, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32,	5,	2
6	6,	1, 2, 6, 9, 13, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32,	6,	3
7	7,	1, 2, 3, 7, 9, 12, 13, 16, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32,	7,	3
8	8,	1, 2, 8, 9, 15, 16, 17, 20, 21, 25, 31, 32,	8,	2
9	9,	9,	9,	6
10	10,	1, 2, 9, 10, 13, 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32,	10,	2
11	11,	1, 2, 11, 17, 21, 25, 30, 31,	11,	4
12	12,	12,	12,	4
13	13,	1, 2, 9, 13, 21, 25, 31,	13,	5

14	14,	1, 2, 9, 11, 13, 14, 17, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32,	14,	3
15	15,	1, 2, 15, 20, 21, 25, 31, 32,	15,	3
16	16,	16,	16,	4
17	17,	17,	17,	5
18	18,	18,	18,	5
19	19,	19,	19,	5
20	20,	20, 32,	20,	4
21	21,	21,	21,	9
22	22,	1, 2, 9, 13, 21, 22, 25, 31,	22,	2
23	23,	2, 21, 23, 31,	23,	3
24	24,	1, 2, 9, 13, 18, 19, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32,	24,	4
25	25,	2, 21, 25, 31,	25,	7
26	26,	26,	26,	5
27	27,	27,	27,	5
28	28,	28,	28,	5
29	29,	29,	29,	5

30	30,	30,	30,	5
31	31,	31,	31,	9
32	32,	32,	32,	5
33	33,	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,	33,	1

Table 5-9 Conical matrix

Factors	33	4	5	8	10	22	6	7	14	15	23	3	11	12	16	20	24	13	17	18	19	26	27	28	29	30	32	1	9	25	2	21	31	Driving Power	Level
33	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
8	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2

10	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
22	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
6	1*	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
7	1*	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
14	1*	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
15	1*	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
23	1*	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
3	1*	0	1*	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
11	1*	0	1*	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4

12	1 *	0	1 *	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
16	1 *	0	1 *	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4
20	1 *	0	0	1 *	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
24	1 *	0	1 *	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	4
13	1 *	0	1 *	0	1 *	1	1 *	1 *	1 *	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	5
17	1 *	0	1 *	1	0	0	0	0	1 *	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	5
18	1 *	0	1 *	0	1 *	0	1 *	1 *	1 *	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	8	5
19	1 *	0	1 *	0	1 *	0	1 *	1 *	1 *	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	8	5

26	1*	0	1*	0	1*	0	1*	1*	1*	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8	5
27	1*	0	1*	0	1*	0	1*	1*	1*	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8	5
28	1*	0	1	0	1*	0	1*	1*	1*	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	5
29	1*	0	1*	0	1*	0	1*	1*	1*	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	8	5
30	1*	0	1*	0	1*	0	1*	1*	1*	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	9	5
32	1*	0	1*	1*	1*	0	1*	1*	1*	1*	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	11	5
1	1*	1	1*	1*	1*	1*	1*	1	1*	1	0	0	1	0	0	0	1*	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	14	6
9	1*	0	1*	1	1*	1*	1*	1*	1*	0	0	0	0	0	0	0	1*	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	11	6

25	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	0	0	1*	0	0	0	1	1*	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	15	7
2	1*	1	1*	1*	1*	1*	1*	1*	1*	1*	1	0	1*	0	0	0	1*	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	17	8
21	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	0	1*	0	0	0	1*	1*	0	0	0	0	0	0	0	0	0	1*	0	1*	1	1	0	18	9
31	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	0	1*	0	0	0	1*	1*	0	0	0	0	0	0	0	0	0	1*	0	1*	1	0	1	18	9
Dependence Power	33	6	25	12	18	8	17	20	19	8	4	1	8	1	1	2	16	7	1	1	1	1	1	1	1	1	1	5	1	4	3	1	1		
Level	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	6	6	7	8	9	9		

6 Multi-source heterogeneous data integration for incident likelihood analysis in the processing systems

Preface

This chapter has been submitted to the ***Computers and Chemical Engineering*** Journal. I am the primary author of this manuscript, along with co-authors Drs. Faisal Khan, Paul Amyotte and Salim Ahmed. I developed the multi-source heterogeneous data framework for accident likelihood analysis and its application in developing the model. I prepared the first draft of the manuscript and revised it based on the co-authors' and peer review feedback. The co-author Dr. Faisal Khan proposed the conceptual framework and helped develop the framework, testing and revising the model. The co-authors, Drs. Paul Amyotte and Salim Ahmed provided constructive feedback to improve the readability, review and revision based on peer review feedback and finalizing the manuscript.

Reference: Kamil, M. Z., Khan, F., Amyotte, P., & Ahmed, S. (2023). Multi-source heterogeneous data integration for incident likelihood analysis in the processing systems. *Computers and Chemical Engineering* - submitted.

Abstract

Structured data, such as sensor data, can provide valuable insights to safety practitioners for developing prevention and mitigation strategies. However, relying on a single data source can introduce biases. In this era of safety 4.0, a methodology that can leverage insights from multiple sources (incident databases and physical observations) is required. This study proposes an approach based on natural language processing (NLP) to learn lessons from past incidents and combine them with contemporary data to predict adverse events. The model is based on feature extraction using a co-occurrence network on the loss of containment (LOC)/release of hazardous substance accidents from 2002 to 2021, sourced from the Chemical

Safety and Hazard Investigation Board (CSB) database. Coupled with the operational parameters, it provides a robust likelihood model. Scenario-based model verification is performed by simulated scenarios based on past incidents of LOC to assess model efficacy in predicting similar incidents. Sensitivity analysis shows inadequate written procedures resulting from management and organizational failure have the highest sensitivity towards LOC incidents. This work assists practitioners in monitoring sensor data and lessons learned from past incidents by utilizing multi-source heterogeneous data sources. Thus, the current research work serves as an important tool to enhance data-driven prediction as part of safety 4.0.

Keywords: Safety 4.0, Natural language processing (NLP) Chemical Safety and Hazard Investigation Board (CSB), Data and insight-driven approach

6.1 Introduction

Over many decades, the world has encountered major process incidents due to several factors, including disregarding safety norms due to lax regulatory and management inspections. These incidents resulted in loss of human life, economic losses, and environmental degradation. When a catastrophic incident happens, the universal phrase we all know is "lessons will be learned" (Mannan & Waldram, 2014). Although, if lessons were learned, similar incidents would not have occurred. The critical challenge is continuous learning from process incidents that keep happening. Indeed, the late Dr. Sam Mannan reminded us of the importance of a paradigm shift to learn from past incidents and develop a multi-national and multilingual database (Mannan & Waldram, 2014). Earlier this year, an Ohio train derailment posed serious health risks to the community of East Palestine, Ohio (NRDC, 2023). From March 2020 until December 2022, the U.S. Chemical Safety and Hazard Investigation Board (CSB) has received reports of 224 process incidents, of which 31 resulted in fatalities, 126 caused severe injury, and 101 led to substantial damage in the U.S.A. alone (U.S. Chemical Safety and Hazard Investigation Board, 2022). The latest reporting period from October 1, 2022, to December 26,

2022, accounts for 36 incidents. Out of those, eight were during the past holiday season. In contrast, October-December of 2021 data shows 16 events, whereas in 2020, 14 events. These numbers depict a significant increase in incidents that coincide with cold temperatures across the USA (CSB News Release, 2022). Another chemical incident occurred in an electronics factory in Hapur, India in which a chemical explosion resulted in ten fatalities and 22 injuries (Reuters, 2022). Citing another example from India, an accident occurred in a meat export plant in Aligarh, India, where ammonia was released, resulting in 59 workers falling ill. The preliminary investigation shows that the leakage was due to two main factors, inadequate supervision and maintenance of gas infrastructure (The Times of India, 2022). Another ammonia leak happened in Massachusetts that resulted in a fatality. The reason was unknown, but according to a US Environmental Protection Agency representative, the system should be resilient enough to prevent loss of containment or minimize the impact of chemical release (NBC Boston, 2022). A fire happened in the BP Husky Toledo refinery's most significant crude unit resulting from the release of flammable chemicals causing two fatalities as well as substantial property damage (Bloomberg, 2022). Many factors are responsible for process incidents, for instance, avoiding proper procedures because of lack of training, no written procedures, negligence, and/or oversight. Poor safety culture, inadequate emergency preparedness, and compromised mechanical integrity are other contributory factors responsible for process incidents (Bhusari et al., 2021).

It is essential to learn from failures by asking two critical questions. What circumstances led to the failure? And why? These questions can be answered by developing an accident causation model (Kamil et al., 2023a). An accident causation model consists of linear and non-linear models. Detailed list of accident models and their advancement is discussed in the literature to highlight their pros and cons (Fu et al., 2020). These models aim to answer the above two questions. However, due to increased process complexity, the approach to developing an

accident causation model needs to be evolved. A model can be generalized for the same incident type while capturing input data from multiple heterogeneous sources. In other words, a generalized hybrid model can be developed that captures insights from multiple data sources. One of the best ways to learn from an incident is to leverage data available in databases. In the past, databases have been utilized for different learning purposes. One important study analyzed 88 CSB investigation reports determining whether the inherent safety concept was followed during the design and operation levels. Initially, 63 reports were analyzed (Amyotte et al., 2011), whereas 25 incidents were considered in the second round (Amyotte et al., 2018). According to their findings, safety measures were not followed at each level in the hierarchy of control. The breakdown of hierachy of controls safety measures by specific levels is as follows: 26% inherent safer design, 10% passive, 16% active, and 48% procedural (Amyotte et al., 2018). This study highlights the importance of the hierarchy of controls in process facilities to prevent adverse events. Another study examined 60 reports to determine the factors responsible for most cases by finding the commonalities (Baybutt, 2016). Manually analyzing accident investigation reports is a time-consuming and labor-intensive task. An enriching solution was provided by introducing NLP applications to CSB reports and introducing a systematic approach to analyzing commonalities (Kamil et al., 2023a). According to the findings, attention must be given to procedures, training, and management of change (MoC) due to their high sensitivity toward oil and refining process incidents (Kamil et al., 2023a). The CSB database was also used to analyze 21 cases to evaluate omissions and oversights in process hazard analysis (PHA). The study (Kaszniak, 2010) concluded that the PHA teams failed to evaluate the control measures and/or safeguards in 19% of cases. There was insufficient layers of protection which enhanced the severity of the hazard (Kaszniak, 2010). An objective risk assessment was performed on microbiologically influenced corrosion (MIC) case sources from Pipeline and Hazardous Materials Safety Administration (PHMSA) database using NLP to

develop risk models (Kamil, et al., 2023b). This study provides a new method of assessing risk using textual data and demonstrates the application of the named entity recognition model in evaluating objective risk (Kamil, et al., 2023b). A semi-supervised method (Ahadh et al., 2021) and a co-occurrence network (Liu et al., 2021) were developed using the PHMSA database. Recently, NLP was used to analyze subject and action words from their co-occurrences for accident consequence prediction (Wang et al., 2023).

Past studies show the importance of database resources. They provide the necessary data for analyzing trends and identifying underlying causes of accidents. This data can then be used to develop best practices and implement safety protocols that can reduce the likelihood of future accidents. Moreover, common risk factors can be identified by organizations with the help of a database, leading them to prioritize safety measures and initiatives more effectively. However, relying on a single source for information introduces biases and does not capture all the factors, such as unsafe acts, conditions, and management and organizational failures that result in an accident. Learning from past experiences requires feature extraction from textual data. Equally important is to know the operating conditions via sensor data and its usage in monitoring an adverse event. Feature extraction from textual data (textual information) and integration with sensor-based operating conditions (numerical data) demand a robust model that can accommodate data from both sources. Conventional multi-source data integration approaches aim to fuse structured data such as sensor data. Data integration from multi-source homogeneous data sources is well studied (Goodman et al., 2013), but multi-source heterogeneous data remains a topic of interest, and comparatively less studied.

The present study attempts to fill the knowledge gap well documented in the literature regarding fusing data from multiple sources (Liu & El-Gohary, 2020). The scope of this study is confined to two types of heterogeneous data; (a) Textual data sources such as accident databases and (b) Structured data sources, such as operational parameters from sensors for real-

time monitoring. The former was leveraged by employing NLP and text-mining techniques to extract features and evaluate accident likelihood (Kamil et al., 2023a; 2023b). On the other hand, the latter was used to develop a learning-based BN model to drive meaningful information for decision-making (Kamil et al., 2021). The objective is to develop an approach to fuse past events' textual data with real-time monitoring numerical data to assess accident likelihood. The objective aims to answer the following research questions:

1. How can a robust accident likelihood model be developed for multi-source heterogeneous data?
2. How to establish and model interrelationships among textual and numerical data?
3. How to assess accident likelihood by learning from past experiences and present conditions?

The present study aims to develop a novel hybrid generalized causation model for loss of containment accidents to serve as a tool for Safety 4.0. Safety 4.0 demands a data-driven approach integrating artificial intelligence techniques (i.e., NLP) to gain insights for better safety management. Collecting data from the database and real-time data from sensors provides a comprehensive view of accident patterns and potential hazards that would otherwise be difficult to detect. The data-driven approach of Safety 4.0 also helps to assess the performance of safety measures for reducing risk. Therefore, a data-driven approach is introduced for Safety 4.0 to assess accident likelihood from multi-source heterogeneous data integration. Firstly, a co-occurrence network is constructed to extract features from a database. The network provides insights into what went wrong and depicts causation. Secondly, real-time monitoring data are captured based on operational parameters. Interpretive structural modeling (ISM) combines these two data sources. ISM establishes interrelationships among past event factors and monitored parameters. The outcome is a hierarchical structure consisting of multi-source heterogeneous data. ISM digraph is mapped into a Bayesian network (BN) consisting of fuzzy

and monitored nodes. Fuzzy logic is used to define linguistic variables from a co-occurrence network. On the other hand, real-time data governs the probability of monitored nodes. Therefore, the novel approach develops a hybrid BN that can accommodate heterogeneous data sources (i.e., textual and numerical) that improve prediction.

Section 6.2 consists of details and steps of the novel approach to assess accident likelihood based on heterogeneous data sources. Section 6.3 contains an application section that develops a generalized hybrid causation model for LOC accidents. Section 6.4 deals with the result and discussion of the study comprising scenario-based model verification exercise and sensitivity analysis to find out the most sensitive parameters. Conclusion of this work is discussed in section 6.5, along with the limitations.

6.2 Research Methodology

A novel methodology of integrating textual data and numerical data are introduced in this study. There are four steps involved in the methodology. Steps 6.11 and 6.22 are related to organizing, analyzing, and interpreting data. Step 6.33 focuses on developing the interrelationship of factors from extracted data into a hierarchical structure. Step 6.4 relies on handling uncertainties with textual data using fuzzy logic and developing a hybrid BN model from a fusion of fuzzy and monitored nodes. Figure 6-1 illustrates the step-by-step approach to data integration of textual and numerical data. The details of each step are as follows:

6.2.1 Employing Natural Language Processing (NLP)

NLP has recently been prominent due to its ability to analyze and determine underlying causes from accident databases. This work utilizes a co-occurrence network diagram to text-mine textual data from the database. Co-occurrence network consists of nodes and edges. Nodes represent words, whereas edges depict co-occurrence between words in the corpus (Zhang et al., 2018). The application of co-occurrence networks in NLP has been widely seen, such as

determining causal relation (Liu et al., 2021), key object extraction (Mihalcea & Tarau, 2004), word sense discrimination (Ferret, 2004) and accident consequence prediction (Wang et al., 2023). This work uses an open-source KH Coder software (Higuchi, 2016) to develop the co-occurrence network. The present study employs the co-occurrence network method to text-mine a database and develop a generalized hybrid causation model qualitatively and quantitatively in section 6.3.

6.2.1.1 Report section selection

Accidents in process industries are seen as a failure; this failure could be in the form of property damage, business interruption, loss of material, environmental degradation, loss of human life, and reputational damage. There is an opportunity to learn from the accidents and avoid future mishaps. It demands an approach that can assist in understanding what went wrong in the past and the underlying causes. An incident database is a key to learning from mistakes that resulted in catastrophic incidents. Every database has its way of storing and categorizing data. It is important to consider the important sections of an incident report, such as the comment section, root cause, contributing cause, or key findings.

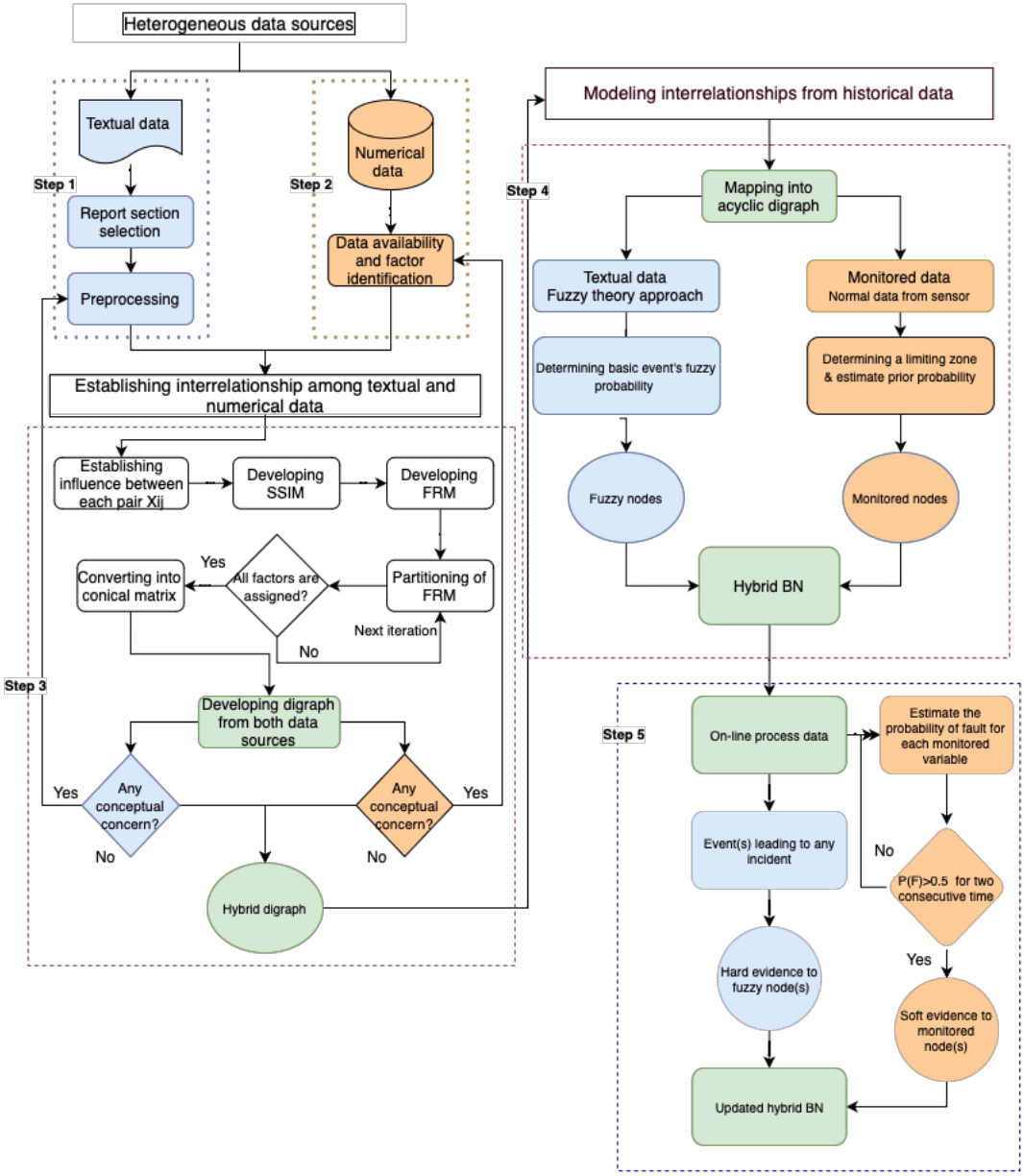


Figure 6-1 The methodology of creating a hybrid causation model from multi-source heterogeneous data

6.2.1.2 Preprocessing

After selecting the desired report section, the next step is to initiate pre-processing of natural language to make it more compatible with NLP tasks. This study advocates four steps: tokenization, stopwords removal, lemmatization, and filtration. Each word is assigned as a token in a sentence in the tokenization step. For instance, *failure in a level controller leads to*

a release of hydrocarbon. This sentence consists of 11 tokens. After tokenization, the next step is stopwords removal. This step requires the removal of those tokens that are less relevant and occur multiple times. These include punctuations, numbers, dates, and stopwords (an, the) that do not provide value in the NLP task. The next step in pre-processing is lemmatization, which converts each word into its base form. Unlike stemming, lemmatization considers the word's context in a sentence before its base form. Due to this reason, this study advocates lemmatization uses as opposed to stemming. The last step in pre-processing is filtration. It filters out those words not included in the stopwords list but does not provide value in the NLP task. These words are domain-specific and can be decided based on domain expertise.

6.2.2 Numerical data

Sensor data are abundant in organizations and processed to convert into meaningful information. The present work aims to integrate textual and sensor data features into a likelihood model. The former provides past information regarding what went wrong, whereas the latter depicts contemporary data useful for monitoring process operations. These two sources together assist in developing a robust likelihood model of combining multi-source heterogeneous data and improving prediction.

6.2.2.1 Data Availability and Factor Identification

Data availability of plants where accident investigations occurred can be obtained using investigation reports. The factors leading to an abnormal situation can be identified from each accident for which methodology is applied. The data are sometimes unavailable or insufficient for analysis. Data availability remains challenging, and more details are provided in section 6.3.

6.2.3 Interpretive Structure Modelling (ISM)

Warfield proposed the ISM method to establish a visual hierarchical structure from unstructured data (Attri et al., 2013; Warfield, 1974). The intent was to use it for decision-making purposes for complex issues. Data from steps 6.2.1 and 6.2.2 serve as input into the ISM method. Unlike previous studies, the literature review serves as an input for the ISM method (Huang et al., 2020; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019). However, the present study advocates a different route of coupling textual and numerical data integration via ISM. The outcome is a well-defined informative digraph that can be used for further analysis. The steps of the ISM process are illustrated in Figure 6-2.

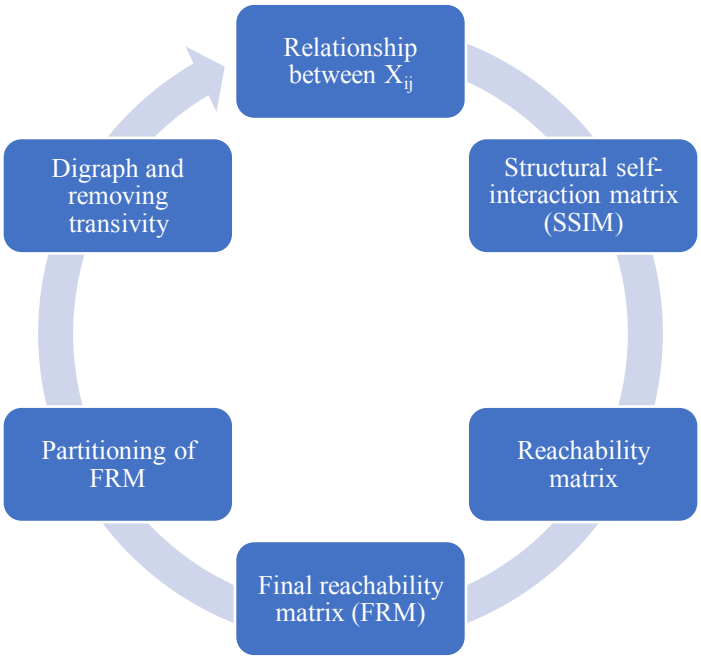


Figure 6-2 Steps of ISM process

6.2.3.1 Establishing interrelationships among heterogeneous factors

The factors identified for the study are derived from the accident database and simulated sensor data. Both serve as inputs and establish interrelationships among identified factors. A pair-wise relation is considered to develop a contextual relationship between both factors or factors within each source. When there is a relation between two factors X_{ij}, if factor i influence factor j but vice versa is not true, this contextual relationship is termed as "yes" from i to j whereas, for j to i is no. Similarly, the relation between each pair is determined to establish the contextual relationship. Domain expertise plays a vital role in deciding the influence of one factor over another.

6.2.3.2 Developing Structural Self-interaction matrix (SSIM)

After establishing the interrelationship among factors, the next step is to develop an SSIM. The SSIM depicts the directed influence of one factor on another through the pair-wise comparison in the previous step. In the ISM process, variables to represent interrelationships are predefined as V, A, X, and O. These relationships are as follows (Kamil et al., 2023a; Sajid et al., 2017):

V- denotes when i influences j, but vice versa is not true

A - denotes when j influences i, but vice versa is not true

X- denotes both i and j influence each other

O- denotes when there is no relation between i and j

6.2.3.3 Converting SSIM into Final Reachability matrix (FRM)

The next step consists of converting SSIM into FRM. First, SSIM is transformed into a reachability matrix (RM) and FRM. Predefined variables V, A, X, and O are used to convert them into binary, 0 or 1. The RM entry can be formulated as follows:

When SSIM entry is V- denotes i to j entry becomes 1 and j to 1 entry becomes 0

When SSIM entry is A- denotes i to j entry becomes 0 and j to 1 entry becomes 1

When SSIM entry is X- denotes i to j entry becomes 1 and j to 1 entry becomes 1

When SSIM entry is O- denotes i to j entry becomes 0 and j to 1 entry becomes 0

According to the predefined variables V, A, X and O initial reachability matrix can be developed. In the ISM method, one assumption is the incorporation of transitivity. Transitivity means that if there are three factors, 1, 2, and 3., Factor 1 influences factor 2, and if 2 influences 3, then 1 is indirectly related to 3 through 2. This relationship in RM is called transitivity and incorporated by 1^{*}. Introducing transitivity to RM resulted in FRM (Attri et al., 2013).

6.2.3.4 Partitioning of FRM and converting into a conical matrix

Partitioning of FRM is essential in establishing a hierarchical level of factors. Two sets are derived from FRM; reachability set $R(X_i)$, and antecedent set $A(X_i)$. The former consists of factor i itself and other factors that i influence, whereas the latter consists of all the factors that influence factor i and factor i itself (Attri et al., 2013). Further, an intersection of the former and the latter $R(X_i) \cap A(X_i)$ is also derived. The factor in which $R(X_i)$ and $R(X_i) \cap A(X_i)$ intersection are the same obtained the top or highest level in the hierarchy as level I. The top-level or highest-level factor does not influence any other factor; in other words, it has 0 driving power. The exact process is repeated by omitting the factor for which the level is assigned until all factor's levels are determined. It is noted that more than one factor can be assigned at the same level. These levels determine their visual hierarchical structure. The following process is converting partitioned FRM into a conical matrix. A rearrangement of all the factors takes place in which all factors assigned to the same level are pooled together in such a way that most zero (0) factors are in the upper half of the matrix while the lower half is unitary (1) (Kamil et al., 2023a).

6.2.3.5 Developing hybrid digraph

The final step of the ISM process is developing a directed graph or digraph from the conical matrix data. A digraph is drawn from the conical matrix relations. If a factor Y affects another factor Z and is represented by unity in the matrix, a directed arc must be drawn from the former to the latter. Likewise, other factors affecting Z are shown by drawing a directed arc from those factors to Z. If the conical matrix entry is 0, factor Y does not affect Z; therefore, no directed arc is drawn. This process continues until all the factors arcs are drawn and their interrelationship is established. In addition, the transitivity links shown in step 6.2.3.3 are removed. The resulting graph is called an ISM digraph, a complex hierarchical structure comprising complex interrelationships among factors and their levels in the hierarchy (Kamil et al., 2023a). In case of any conceptual concern, readers are referred to both sources' data preprocessing steps.

6.2.4 Quantitative reasoning

The digraph developed using the ISM method is qualitative. The last step of the methodology is to model interrelationships quantitatively and estimate causation likelihood. The following substeps provide details on mapping the ISM digraph into quantitative analysis and handling uncertainty in the assessment.

6.2.4.1 Mapping hybrid digraph into an acyclic digraph

This study methodology introduces a probabilistic technique incorporating natural language textual and sensor data. The textual data are dealt with using fuzzy logic as a bridge to quantify and manage qualitative data, whereas the latter can be converted to probabilities based on the three-sigma rule. One popular probabilistic technique that can model fuzzy and monitored nodes is the BN. BN represents a failure scenario from causation to consequences, making

modeling of failure scenarios easy to follow. The influence of causes among each other can be modeled through conditional probability. Based on the quantitative relationship, BN estimates the posterior probability of the pivotal node (Saeed et al., 2022). Many studies in the past (Huang et al., 2020; Kamil et al., 2023a; Li et al., 2019; Sajid et al., 2017; Wu et al., 2015; Yuan et al., 2019) chose BN as a prominent option for quantitative analysis combined with the ISM method. However, one core difference between both approaches is that BN is acyclic, whereas ISM can be cyclic or acyclic. As proposed by (Kamil et al., 2023a), two main rules must be considered while mapping the ISM digraph into BN. Firstly, eliminate any single-parent arc to a child node. Secondly, check for cyclic structure. The mapping of the ISM digraph can be made by following the mapping algorithm illustrated by (Kamil et al., 2023a). If the mapped BN consists of a cyclic structure, then modification can be made by introducing a dummy node. The concept of the dummy node has been introduced and leveraged in the past (Amin et al., 2018; Yu & Rashid, 2013).

6.2.4.2 Estimation of fuzzy probabilities

Fuzzy logic is leveraged to handle vagueness in natural language and quantify subjective qualifications. It provides a bridge between qualitative data and quantitative data. Factors used in the ISM process are derived from the co-occurrence network and require exploiting the numerical relationship between vague quantities.

Table 6-1 Linguistic variables and associated fuzzy numbers to describe fuzzy event, adopted from (Chen Shu-Jen and Hwang, 1992; Zarei et al., 2019)

Linguistic variable	Definition	Fuzzy set
Very High	Occurrence is monthly	(0.8,1,1,1)
High-very High	Occurrence in 1-3 months	(0.7,0.9,1,1)

High	Occurrence in 3-6 months	(0.6,0.8,0.8,1)
Fairly High	Occurrence in 6-12 months	(0.5,0.65,0.65,0.8)
Medium	Occurrence in 1-5 years	(0.3,0.5,0.5,0.7)
Fairly Low	Occurrence in 5-10 years	(0.2,0.35,0.35,0.5)
Low	Occurrence in 10-15 years	(0,0.2,0.2,0.4)
Low-Very Low	Occurrence in 15-20 years	(0,0,0.1,0.3)
Very Low	No occurrence during life cycle	(0,0,0,0.2)

Assigning a probability of failure to a vague event, i.e., naturally spoken/written events, is challenging. Expert elicitation is a consensus scientific way of estimating the probability of such events. The linguistic variable is a potential and effective way of dealing with naturally spoken/written events (Zadeh, 1965). A variable that is defined by words or sentences in a natural or artificial language is called a linguistic variable (Chen Shu-Jen and Hwang, 1992). This work advocates for selecting a scale of 7 fuzzy numbers that consists of 9 linguistic terms, as shown in, Table 6-1 for estimating the likelihood of an event and trapezoidal fuzzy numbers (Zarei et al., 2019). Figure 6-3 depicts a conversion scale of a linguistic variable that can be used to estimate the likelihood of an event. Expert opinions can be aggregated using arithmetic averaging, voting, and fuzzy preference relations (Nurmi, 1981). An appealing technique is a linear opinion pooling as shown in equation (1) (Clemen & Winkler, 1999; Zarei et al., 2019):

$$P_i = \sum_{i=1}^{p} E_i L_{ij} \quad , \quad j = 1,2 \dots, n. \qquad (1)$$

Where Pi is the fuzzy likelihood of an event, E_j is the weighting score of expert j, and L_{ij} is the linguistic variable value from expert j about the event i. The weighting factor and trapezoidal function defuzzification are estimated according to recent studies (Ramzali et al., 2015; Zarei

et al., 2019). The center-of-area technique is considered for defuzzification (Sugeno & Kang, 1986).

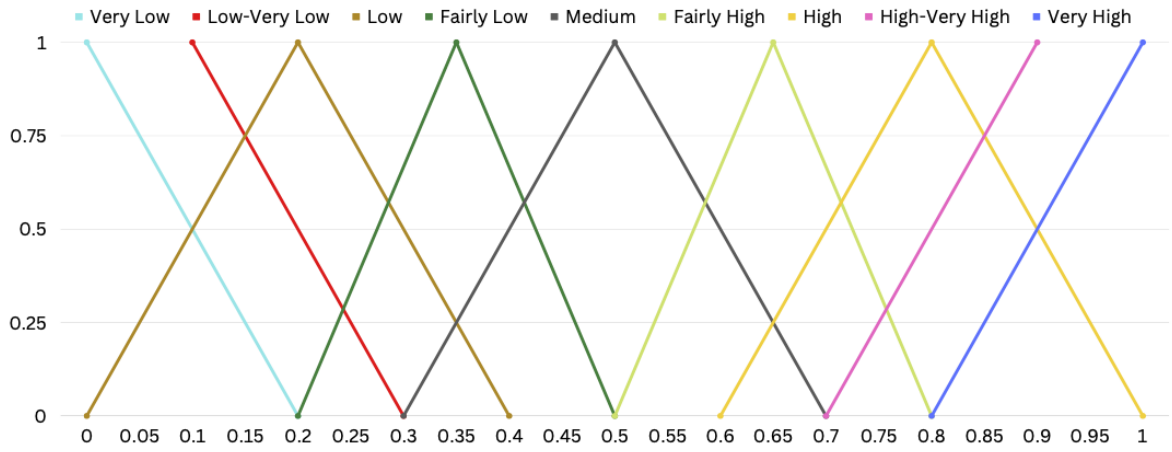


Figure 6-3 Estimation scale of a linguistic variable into fuzzy likelihood

The final step is to estimate fuzzy probability from the fuzzy possibility. Fuzzy probability can be obtained using a defined function, as shown in equations (2) and (3), developed by (Onisawa, 1988):

$$FPr = \begin{cases} \frac{1}{10^K} \; if \; FPs \neq 0 \\ 0 \; if \; FPs = 0 \end{cases} \qquad (2)$$

$$K = \left[\left(\frac{1 - FPs}{FPs} \right)^{\left(\frac{1}{3}\right)} \right] \times 2.301. \qquad (3)$$

Where FPr is a fuzzy probability, FPs is a fuzzy possibility, and K is a constant value for each event.

6.2.4.3 Estimation of monitored nodes

The next aspect of modeling interrelationships and estimating failure likelihood is the estimation of probabilities from normal and faulty data of monitored variables. The previous

step employed fuzzy nodes for handling uncertainty from natural language in the accident database. In contrast, this step focuses on estimating probabilities from sensor data.

$$\Pr(fault) = \varphi\left(\frac{Y \pm \mu}{\sigma}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

Where Y is an arbitrary value, μ is the mean and σ is the standard deviation of Gaussian cumulative distribution.

Using the three-sigma rule, a limiting zone is defined to estimate the probability of fault from sensor data. This limit consideration is recommended due to noise in process data. In a Gaussian distribution, mostly all (99.7%) values lie between the upper and lower thresholds within 3 standard deviations of the mean, i.e., $\mu + 3\sigma$ (upper control limit) and $\mu - 3\sigma$ (lower control limit), respectively. At the mean, the probability of fault is 0, whereas at the lower and upper thresholds is 0.5 (Amin et al., 2021; Bao et al., 2011).

$y_{ij} > \mu_j$,

$$\Pr(Fault) = \varphi\left(\frac{y_{ij} - (\mu_j + 3\sigma_j)}{\sigma_j}\right)$$

$$= \int_{-\infty}^{y_{ij}} \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{\{y_{ij}-(\mu_j+3\sigma_j)\}^2}{2\sigma_j^2}} dx \dots\dots\dots\dots (5)$$

$y_{ij} < \mu_j$,

$$\Pr(Fault) = 1 - \varphi\left(\frac{y_{ij}-(\mu_j-3\sigma_j)}{\sigma_j}\right)$$

$$= 1 - \int_{-\infty}^{y_{ij}} \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{\{y_{ij}-(\mu_j-3\sigma_j)\}^2}{2\sigma_j^2}} dx \dots\dots.. (6)$$

Where i=1,2,..n and j=1,2,..m

The prior probability of monitored variables is estimated by averaging the probability obtained from normal data using equations (5) and (6) for each corresponding variable (Amin et al., 2021).

6.2.5 Generalized Hybrid Causation Model

The textual and numerical data interrelationships can be modeled by estimating fuzzy and monitored nodes. Another parameter is the conditional probability table (CPT) that can be defined based on OR/AND gate. The resulting causation model consists of two nodes: fuzzy nodes from the textual data and monitored nodes from the numerical data when these nodes are combined in BN results in a generalized hybrid causation model.

6.2.6 Updated Hybrid Causation Model

Online process data plays an important role in risk monitoring. The steps included in the methodology for online process data are adapted from a recent study (Amin et al., 2019). Firstly, the fault probability is estimated using equations (5) and (6); if the probability is more than 0.5 for two consecutive samples, there is a fault. Therefore, the corresponding monitored variable node is updated by providing soft evidence (i.e., likelihood evidence).

6.3 Application to CSB Database

The approach developed in the previous section aims to use our knowledge by utilizing historical data from an accident database. The lessons learned from historical data give insights that would be assisted with contemporary data to monitor risk and assess the process operating condition. This section applies the methodology to the CSB database of loss of containment/release incidents between 2002 to 2021, accounting for 18 such incidents. LOC occurs due to the escape of hazardous substances such as gas, fuel, or chemicals from a storage vessel (U.K. HSE discovering safety, 2021).

6.3.1 Heterogeneous Data Sources

Textual data use is challenging, as well as driving interest among risk analysts to leverage for risk estimation. A co-occurrence matrix of dimension C*C consists of entities in rows and

columns based on a unique word in the database, where C is the sum of unique words in input data. Therefore, results in a word-to-word matrix to identify their linkage. Each word is vectorized as a co-occurrence frequency with other words, leading to a co-occurrence network (Liu et al., 2021). The network is a qualitative analysis of words and their interaction, representing their occurrences in a dataset (Zhang et al., 2018). The co-occurrence network is a popular NLP technique widely used for graphical representation.

Structured data from a sensor provides information about process operation. Real-time monitoring data are a useful resource that can be utilized in integration with unstructured data from accident investigation reports. The structured and unstructured data are used to develop a robust likelihood model.

6.3.1.1 Selecting textual data and Preprocessing

This step comprises selecting a section from the accident investigation and processing the data. The accident report provides sufficient information about the incident. The sections explored in CSB reports are executive summary, key findings, and root/contributory causes. The corpus consists of data from the sections about each incident of LOC. After, developing corpus, preprocessing is applied to remove noise, convert texts into base form and filtering of words that adds less value. In this case, company names where the accident occurred are repeated in the accident descriptions. Therefore, such names are omitted in the filtration step.

6.3.1.2 Developing co-occurrence network

An open-source KH coder (Higuchi, 2016) is used to establish a co-occurrence network. The nodes in the network denote the target words with their sizes representing their occurrences in the dataset. The strength of the edges shows the value of a Jaccard coefficient. The Jaccard coefficient measures similarities between different data sets by estimating shared and distinct

elements to evaluate all possible relations between two words (Liu et al., 2021; Romesburg, 2004). A threshold value of 0.2 is set to consider strong co-occurrences among words in the graph.

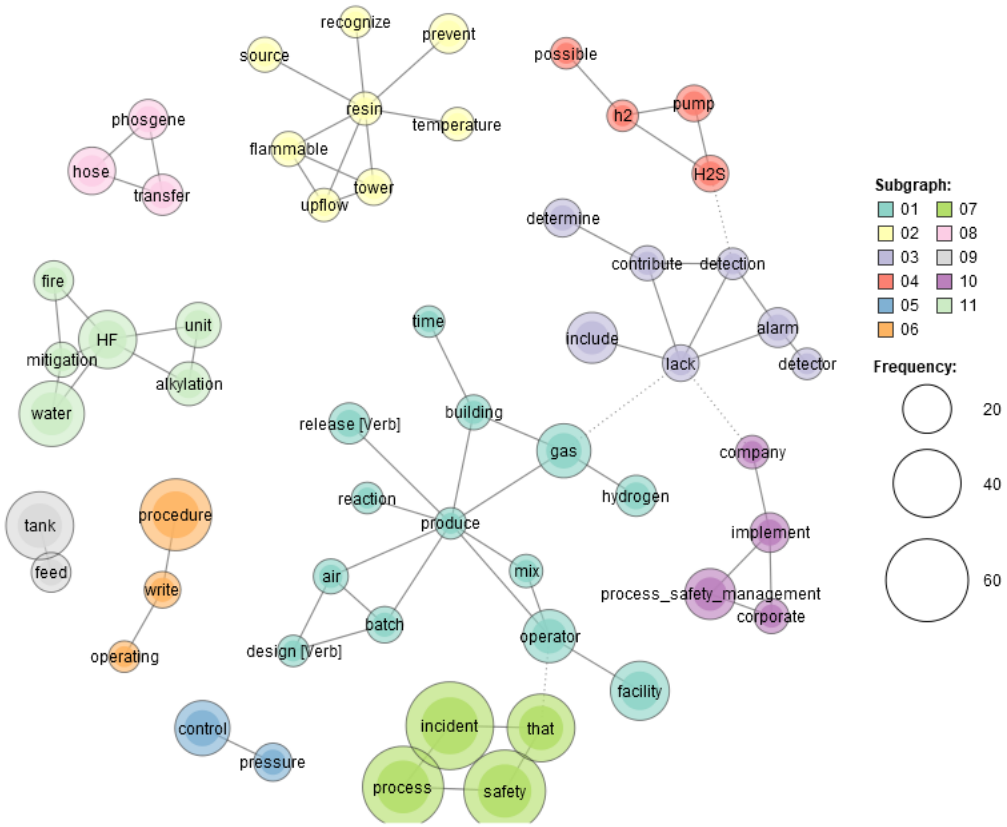


Figure 6-4 Co-occurrence network of release incidents from the CSB database of LOC incidents

The co-occurrence network developed from CSB database incidents related to the release scenario is illustrated in Figure 6-4. There are 11 color-coded subgraphs. The network represents that certain words are closely associated, forming a subgraph in the network, and each subgraph can demonstrate an incident's causation. Based on the initial network, the filtering step can omit words that do not add value in causation. The filtering step is a hit-and-trial method that requires domain expertise while maintaining clarity and information in the graph. Causations from subgraphs related to each other through dash edges (co-occurrences exist in different communities). Subgraphs 01, 03, 04, 07, and 10 have dash edges and discuss

them. Subgraph 03 denotes, it is "determined" that due to "lack" of "alarm" "detector" contributes to the severity of an incident. Dash edges from "lack" lead to "companies", "implementation" of "process safety management" elements is due to weak safety culture by the "corporate" as shown in subgraph 10. One of the "company" policies does not even identify hot work in activities that can be a source of ignition.

Subgraph 04 indicates that toxic H_2S gas is accumulated in a facility's "pump" room and is unable to be vented due to the "lack" of a "detection" device or "detector" fail to trigger "alarm." Subgraph 01 suggests that a chemical "reaction" "produces" "Hydrogen" "gas." The reaction occurred due to the "operator" "mixing" incompatible chemicals. It is linked to subgraph 10, i.e., "lack" of "process safety management." The "air" mover is designed to bring fresh air into a "building" close to a "batch" operation carried out that "produces" "hydrogen" "gas" by the "operator." When "hydrogen" "mixed" with "air," “that” resulted in "process," "safety," "incident," shown in subgraph 07. In addition, the building was "lacking" a gas "detection" system. This could be due to a lack of sensors or sensor failure in detecting the gas.

Subgraph 02 suggests that during a maintenance procedure in "upflow," "tower," "flammable," "resin" came into contact with the ignition "source" (heat gun), leading to high "resin" temperature resulting in fire. Maintenance workers failed to "recognize" heat guns as hot work that could ignite flammable material. Subgraph 05 suggests two words, "pressure" and "control." It may reflect that an inadequate "pressure" "control" system leads to an incident. Subgraph 06 shows inadequate "written," "operating," "procedure" cause behind improper action of an operator that could be catastrophic, as seen in subgraph 01, due to the operator's inadequate experience leading to producing hydrogen gas that released and ignited.

Subgraph 08 shows causation due to "phosgene," "transfer," "hose" that led to the release of "phosgene". The highly toxic "phosgene" is released due to "hose" failure. Subgraph 09 shows an exceeding operating temperature related to the "feed" "tank" due to the lack of operating

procedure and hazard analysis. The last subgraph 11 illustrates an accident scenario of "HF" (Hydrogen Floride) release from an "alkylation" "unit" resulting in "fire." The "water" "mitigation" system unable to performs its intended job.

Step 6.2.1 of the methodology investigates an automated identification of underlying causal factors of incidents using textual data. There are limited available solutions for text-mining incident reports. Recent works introduce a method of extracting causation from incident textual data (Kamil et al., 2023a; 2023b), but it requires corpus training, unlike a co-occurrence network method. Therefore, avoiding training datasets will be less time-consuming and require minimal labor.

6.3.1.3 Simulating real-time sensor data

Another important aspect of this step is obtaining numerical data from sensors. Collecting desired data are not easy; it is challenging due to the unavailability of data or insufficient data points for analysis (Kamil et al., 2021). The former is the concern in this work, leading to simulation of monitoring data for quantitative reasoning. Three monitored variables are important, particularly for LOC accidents based on CSB cases considered. These monitored variables are valve opening malfunction, pressure, and temperature. These factors vary from one process operation to another. However, to show the proposed methodology's efficacy, three factors from CSB cases are considered in the present work. Getting real monitoring data remains a challenge to the present work.

6.3.2 Establishing Interrelationship among Textual and Numerical Data

The next step of the developed approach is to establish interrelationships among factors from both data sources, textual and numerical. The ISM method is proven to be capable of modeling complex interrelationships among factors (Attri et al., 2013). This step establishes

interrelationships among factors from textual data and monitored variables identified from the CSB cases of LOC incidents. If any study only relies on numerical data, then the Kullback-Leibler divergence method can define the relationship between monitored variables (Amin et al., 2021). In the present case, an interrelationship is established among both data types. The outcome from the ISM digraph can be mapped to BN using the available mapping algorithm. The challenges arising from the former and latter method incompatibility are discussed in a recent study (Kamil et al., 2023a). Table 6-2 lists all the factors identified from both data sources.

Table 6-2 Identified factors for ISM process

Serial number	**Factors**
I	Loss of containment
II	Operator/Human factor
III	Ignition source
IV	Inadequate written operating procedure
V	Lack of PSM
VI	Fire
VII	Weak safety culture by corporate
VIII	Lack of alarm
IX	Lack of detection devices (gas detectors)
X	Inadequate pressure control
XI	Hose failure
XII	Feed tank
XIII	Failure of water mitigation system
XIV	Flammable resin due to hot work permit

XV	Temperature
XVI	Pressure
XVII	Level
XVIII	Valve opening malfunction
XIX	Sensor malfunction
XX	Lack of sensor

6.3.2.1 Developing SSIM, RM and FRM from heterogeneous data

Firstly, a pair-wise contextual relationship is established by analyzing each factor's influence on other factors. This relationship is developed in the form of yes or no. The pair-wise relationship is established among all the factors using the understanding of the accident causation and then converted into SSIM. The developed SSIM is shown in Appendix Table 6-5, consisting of V, A, X, and O to demonstrate the pair-wise relationship among factors. Based on SSIM, RM is developed comprised of 0 or 1. In addition, the indirect relation of identified factors is also incorporated in the RM, resulting in FRM, as Appendix Table 6-6 shows. Furthermore, FRM also includes two important aspects: driving power and dependence power. As the name suggests, the former denotes the total number of interactions by each factor in a row (i.e., factors it affects). In contrast, the latter is the total number of interactions for columns (Kamil et al., 2023).

6.3.2.2 Establishing hierarchy among identified factors

The next step of ISM process is very important because this will dictate the hierarchy of factors and decide the causal factors in the resulting ISM digraph. FRM of appendix Table 6-6 is partitioned into different levels. FRM data facilitate this partition by developing two sets: reachability set, R(Xi), and antecedent set, A(Xi). Appendix Table 6-7 shows level partitioning,

which is an iterative process. At each iteration, a level is assigned to the factors. In the present case, there are seven iterations to develop the level partitioning shown in Table 6-7. R(Xi) consists of factor i and other factors influenced by factor i, whereas A(Xi) comprises factor i itself and factors that influence factor i. Consequently, the intersection of R(Xi) and A(Xi) is derived for all the cases. The factor that is common between R(Xi) and R(Xi)∩A(Xi) column occupies the top level in the hierarchy (Kamil et al., 2023; Sajid et al., 2017). In the current work, Fire & explosion (F&E) due to LOC occupy level I, and no factor exists above the level I. Similarly, other iterations are performed after removing the factor already assigned a level until all factors have been assigned. All factors have been assigned a level in the hierarchical structure in seven iterations. Table 6-7 shows all the factors and their assigned levels.

6.3.2.3 Developing hybrid ISM digraph

The last step in the ISM process is developing a conical matrix from level partitioning that develops ISM digraph. A conical matrix is developed to visualize each factor hierarchy to create a digraph, as developed in appendix Table 6-8. The upper half of the conical matrix consists of null elements. In contrast, the other half consists of unitary elements to reflect dependence and driving power with their respective levels. For instance, F&E occupies level 1, according to the conical matrix, followed by two factors, LOC and ignition source, at level 2. In Table 6-8, rows 2 and 3 depict LOC and ignition source influence F&E. Therefore, a directed arc is drawn from both factors of level 2 to the F&E (level 1) to show the influence. In row 4th, the factor level directly influences LOC and indirectly influences (denoted by transitivity) F&E. Both relations are used at this stage to draw two arcs.

Similarly, each relationship of factors is depicted in a complex structure while maintaining their hierarchy in the overall process. Once all the relationships are depicted in the ISM process, the resulting diagram is known as a digraph. All the transitivity links are removed from the

final digraph, as shown in Figure 6-5. The digraph suggests that F&E has the highest dependence power of 20, meaning that all 20 factors in the ISM digraph directly or indirectly lead to F&E. Two factors have the highest driving power, i.e., sensor malfunction and lack of sensor, meaning that these factors contribute most towards the LOC and subsequently to F&E.

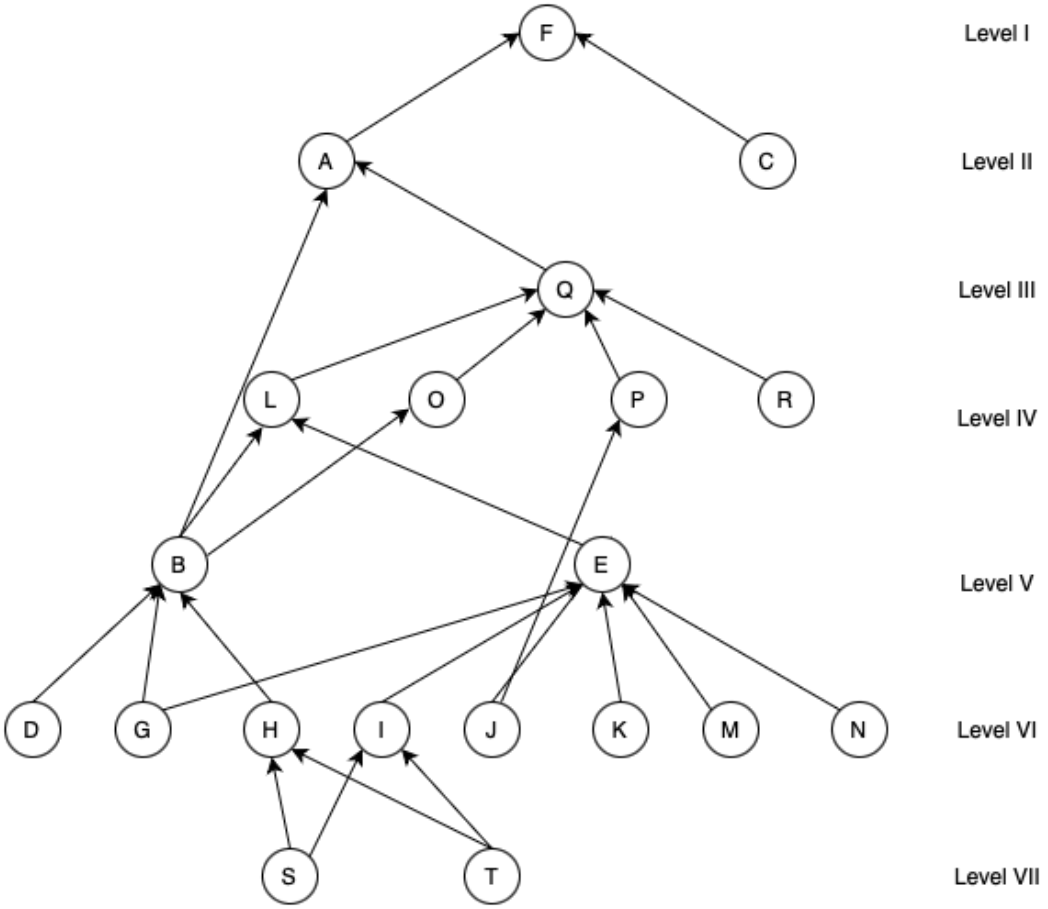


Figure 6-5 Developed digraph from heterogeneous data sources

6.3.3 Developing Hybrid Causation Model

The digraph is constructed from a complex relationship among factors from both data sources. ISM is a qualitative analysis that can be beneficial to understand the hierarchy of factors and factors with the highest driving and dependence power. In order to estimate the likelihood of LOC and, subsequently, F&E. ISM digraph needs to be evolved into a quantitative approach. BN has proven to be a useful technique for cases that can model causation from root causes to

consequences in one pictorial representation. A mapping algorithm is used to map the digraph into BN (Kamil et al., 2023a). The mapped BN consists of two monitored nodes, O & P has a single parent node. These single-parent arcs are removed from BN model resulting in mapped BN shown in Figure 6-7. No cyclic relations are encountered throughout the process.

Table 6-3 Fuzzy probability estimation

Fuzzy variable	**Expert 1**				**Expert 2**				**Expert 3**				**Fuzzy possibility**	**K**	**Fuzzy Probability**
Ignition source	0.80	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.93	0.96	1.11E-01
Inadequate written operating procedure	0.70	0.90	1.00	1.00	0.80	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.92	1.02	9.63E-02
Weak safety culture by corporate	0.20	0.35	0.35	0.50	0.30	0.50	0.50	0.70	0.20	0.35	0.35	0.50	0.40	2.63	2.32E-03
Inadequate pressure control	0.60	0.80	0.80	1.00	0.50	0.65	0.65	0.80	0.50	0.65	0.65	0.80	0.70	1.74	1.84E-02
Hose failure	0.30	0.50	0.50	0.70	0.50	0.65	0.65	0.80	0.30	0.50	0.50	0.70	0.55	2.15	7.04E-03
sensor failure	0.80	1.00	1.00	1.00	0.70	0.90	1.00	1.00	0.80	1.00	1.00	1.00	0.92	1.02	9.63E-02

Failure of water mitigation system	0.60	0.80	0.80	1.00	0.50	0.65	0.65	0.80	0.60	0.80	0.80	1.00	0.75	1.60	2.54E-02
Flammable resin due to hot work permit	0.50	0.65	0.65	0.80	0.60	0.80	0.80	1.00	0.80	1.00	1.00	1.00	0.79	1.47	3.41E-02
Lack sensor	0.70	0.90	1.00	1.00	0.80	1.00	1.00	1.00	0.70	0.90	1.00	1.00	0.91	1.08	8.33E-02

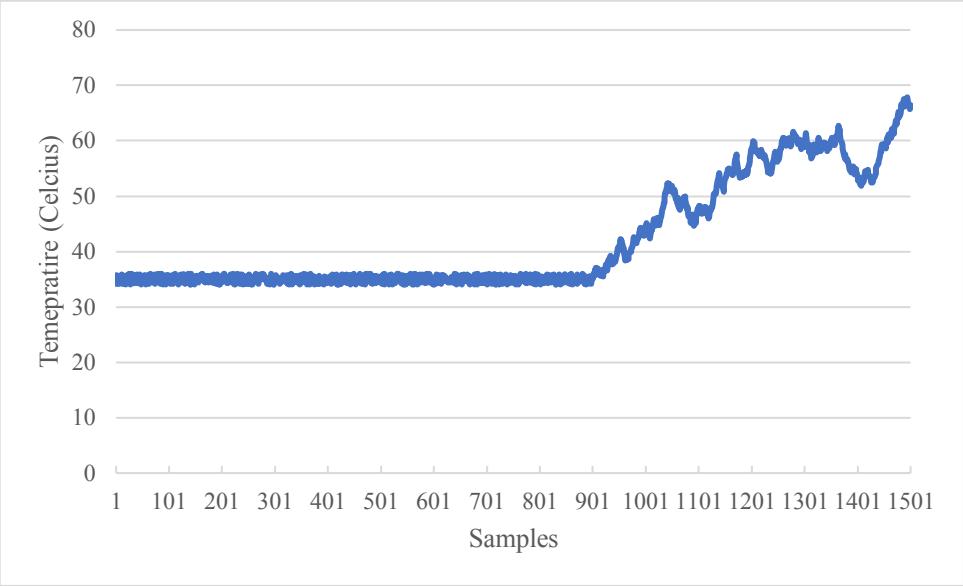


Figure 6-6 Simulated sensor data for temperature

Two parameters must be defined: prior probability of root nodes (fuzzy/monitored nodes) and conditional probability. The former is estimated using fuzzy logic for fuzzy nodes (shown in Table 6-3). For monitored nodes, the prior probability is estimated from normal data by averaging it using equations 5 & 6 using simulated sensor data. Three monitored parameters are temperature, pressure, and valve opening malfunction. For each parameter, 1500 data points are simulated in which fault is introduced at the 900th sample point. An example of temperature sensor data are shown in Figure 6-6 to visualize the simulated data. Similarly, the other two parameters are simulated, and their prior probabilities are estimated based on the three-sigma rule. Another important aspect of BN is CPT. The CPTs are defined using OR/AND gates to model the interrelationship among factors.

LOC accidents are mainly due to three causation factors: unsafe acts, unsafe conditions and organizational and management failures. These factors are illustrated and color-coded in Figure 6-7, along with monitored nodes. The methodology shown in Figure 6-1 is applied to develop a hybrid causation model consisting of monitored factors with causation factors from historical

data. The next step is to update the monitored nodes with a probability of fault. When Pr (fault) > 0.5 for two consecutive samples, it is considered to be a fault. Equations (5) & (6) are used to estimate fault probability for the hybrid causation model (Figure 6-7).

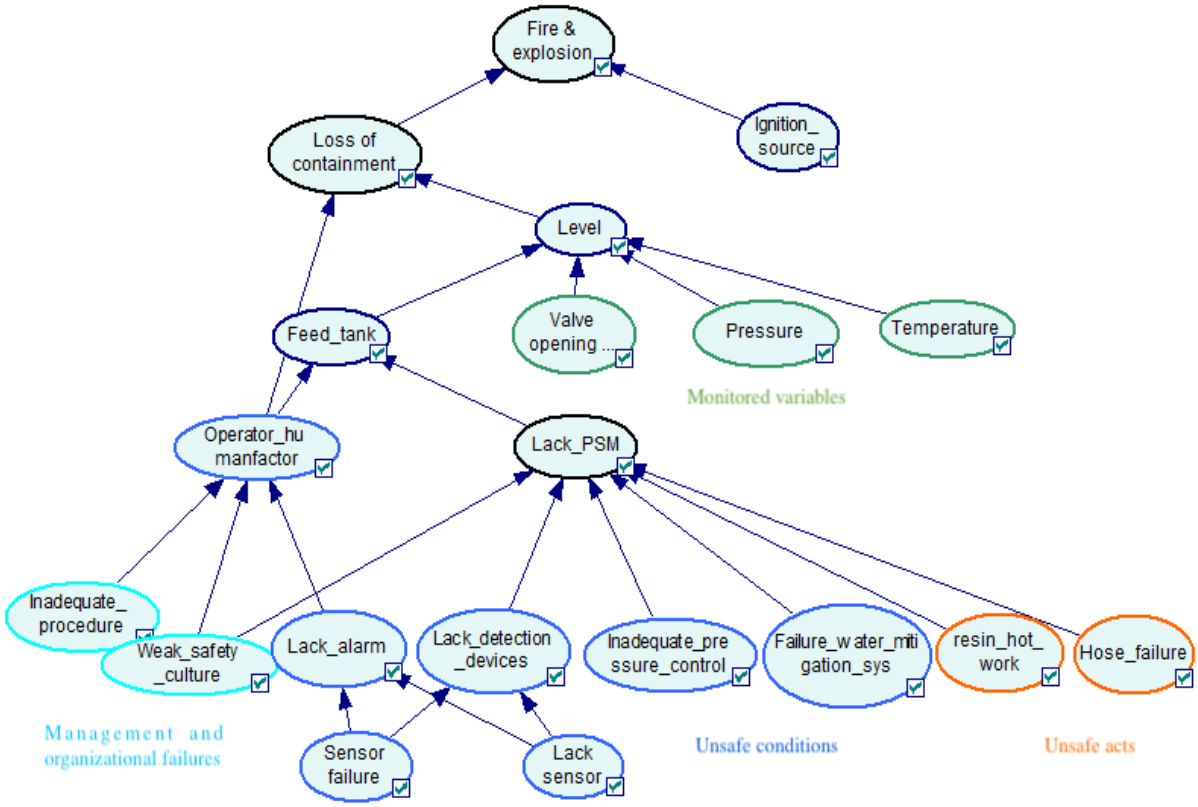


Figure 6-7 Mapped BN from ISM digraph

6.4 Results and discussion

The present study comprises of multi-source heterogeneous data in developing a qualitative model in the form of an ISM digraph to analyze the hierarchy of factors and establish their interrelationships. The ISM digraph is mapped into BN to quantify interrelationships and analyze the likelihood of LOC accidents of nearly two decades from the CSB database. The hybrid model mapped from the ISM digraph is shown in Figure 6-7. The result gives LOC likelihood of 3.70E-01. This prior likelihood is updated with soft evidence of each monitored node (Amin et al., 2019). The hybrid causation model is updated with Pr (fault) for all three nodes and provides a posterior likelihood of 5.88E-01 for LOC, which leads to the F&E

likelihood of 6.51E-02. The results show that integrating fault probabilities of monitored nodes with fuzzy nodes from the accident database led to high LOC likelihood. This way, causation factors are comprised of fuzzy and monitored nodes in the hybrid causation model. The result is coherent with the actual condition because the LOC accidents from which the hybrid BN model is developed are based on real accidents. Therefore, the hybrid BN model accommodates structured and unstructured data to analyze accident causation. Scenario-based verification is carried out to assess the efficacy of the model.

6.4.1 Scenario-based verification

The LOC model developed through a systematic approach depicted in Figure 6-1 consists of different accident scenarios. Each accident scenario has a unique accident pathway comprised of causation factors. The aim is to conduct scenario-based verification by simulating different accident scenarios to assess model prediction. This exercise dictates the efficacy of the model in predicting LOC accidents. The model consists of hard evidence for fuzzy nodes and soft evidence for monitored nodes. The advantage of developing a generalized model is understanding similarities among accidents of similar types and being able to model single accident causation. The accident causation modeling approach is evolved to consider a generalized model for a particular accident type (Kamil et al., 2023).

Scenarios 1-5 are generated from LOC incidents from the CSB database used to develop the BN model in Figure 6-7. Scenario 1 identified factors: lack of sensor (gas detection), weak safety culture, abnormal temperature, and an unknown ignition source. These factors are mentioned in Table 6-4 and are given as hard evidence to fuzzy nodes and soft evidence to monitored nodes. The model gives the likelihood of 100% for both LOC and F&E when met with an unknown ignition source. The result shows that the model can predict the failure scenario correctly and coherently with the actual condition. Similarly, scenario 2 factors in

Table 6-4 are simulated in the BN model. Similar to scenario 1, LOC and F&E likelihood is 100%. The ignition source, in this case, is the heat gun. It ignites flammable resin when came into contact. The model results depict the same outcome when scenario 2 factors are given to the BN model. Based on these two cases, it is evident that the generalized model can be used to assess individual accident pathways. In scenario 3, sensor failure hinders operator action as operator action is based on alarm. Another factor is the lack of PSM; these are the reasons for LOC. The model result shows 100% LOC likelihood, with 11% related to F&E. No F&E was reported in this incident. However, there is a high chance that the hazardous substance (i.e., hydrogen sulfide) is highly flammable and may ignite when met with an ignition source. According to the model, there is 11% chance of F&E, given the release of hydrogen sulfide. Similarly, scenarios 4 & 5 are based on their identified factors listed in Table 6-4. The result shows that LOC is sure to occur with 100% likelihood and 11% F&E. In both cases no F&E occurred, which is also verified by the model results.

Table 6-4 Scenario-based hard and soft evidence for verification

Scenario	**Hard and soft evidence**	**Incident name**	**Model result**	
			Loss of containment	**F&E**
1	Lack of sensor (gas detection), safety culture, temperature, unknown ignition source	AB Specialty Silicons	100%	100%
2	Hot resin, heat gun as ignition source	Evergreen Packaging Paper Mill -	100%	100%

		Fire During Hot Work		
3	Sensor failure, operator, lack of PSM	Aghorn Operating Waterflood Station Hydrogen Sulfide Release	100%	11%
4	Weak safety culture, PSM, detection device, high pressure, valve opening	DuPont La Porte Facility Toxic Chemical Release	100%	11%
5	Valve opening/closing malfunctioning, hose failure	Emergency Shutdown Systems for Chlorine Transfer	100%	11%
6	Inappropriate design, inadequate safeguards such as gas detectors, Inadequate management of change (MoC) , inadequate operator training, inadeqaute emergency	Multiple storage tank ruptures, San Juan	100%	100%

	response, ground flare (ignition source)	Ixhuatepec, Mexico		
7	Sensor failure, inadequate operating procedure, inadequate monitoring, inadequate MoC, inadequate maintenance, human factor, ignition source	Gasoline storage tank overfilled, Buncefied, UK	100%	100%
8	Absence of alarms, valve inconsistency, human factor, inadequate level sensor, inadequate procedure, ignition source	Gasoline storage tank overfilled, Bayamon, Puerto Rico	100%	100%

Scenarios 6-8 are LOC accidents resulting from oil product storage shown on the lessons learned database comprising 52 major process accidents (*IChemE Safety and Loss Prevention*, 2022). Scenario 6 is a multiple LPG storage tank rupture in 1984 Mexico. The LOC occurred due to a rupture in the liquified petroleum gas (LPG) transfer line. This rupture resulted in leakage of LPG, when met with a ground flare propagated into a series of boiling liquid expanding vapor expansion (BLEVE). This catastrophe led to the evacuation of 200,000 people. The identified factors leading to this incident are listed in Table 6-4. Based on the evidence, BN model gives the likelihood of 100% for both LOC and F&E. Scenario 7 depicts the Buncefield, U.K. incident of gasoline overfill in which multiple F&E occurred. The unconfined vapor cloud ignition took place from overfilled gasoline, followed by fire that lasted for five days. This incident comprises basic factors such as failure of the automatic tank gauging system and high-level switch for the automatic high-level shutdown system. The root

factors are shown in Table 6-4, comprised of factors related to an unsafe act, unsafe conditions, management, and organizational failures. These causation factors led to the Buncefield LOC incidents. Based on the provided evidence, the model gives a 100% likelihood of LOC and F&E. The last scenario is again based on a gasoline storage tank overfilled that occurred in Bayamon, Puerto Rico, in 2009. In this incident, a gasoline storage tank was overfilled during an unloading operation. The incident resulted in a vapor cloud explosion followed by fire that lasted 66 hours. The causes of this incident are listed in Table 6-4. The model results show 100% likelihood for LOC and F&E. This scenario-based verification aims to show the model's efficacy in predicting LOC accidents. The model is verified on LOC accidents from the CSB database from scenarios 1-5, which were also part of model development. Scenarios 6-8 are unseen to the model and do not participate in model development. All scenarios 1-8 are tested and verified on the hybrid BN model. The result shows that based on each incident scenario, the model result is coherent with the actual incident. In scenarios 1, 2, 6 & 8, F&E occurred due to the release of contained material met with the ignition source. The model also suggests 100% F&E likelihood in those scenarios. Scenarios 3-5 show 11% likelihood due to the possibility of ignition in released flammable substance. The model provides promising results in each scenario. Therefore, the model is useful in predicting LOC incidents.

6.4.2 Sensitivity Analysis

Sensitivity analysis is performed to determine each factor's sensitivity towards LOC. The analysis can be performed directly from GeNie software (GeNie Software, 2023) by setting the target node of LOC and performing the sensitivity of each root node. The present study uses MS Excel to perform sensitivity analysis. Each root node, whether fuzzy or monitored, is given a percentage. This percentage can vary from $\pm 10, 20, 30,$ and so on. In the present case, ± 50 of each node is considered to capture its effect on the LOC pivotal node. This process is

repeated until all fuzzy and monitored percentage change is done. Based on the outcome, a tornado chart is developed to visualize the effect of each fuzzy and monitored node on LOC, as shown in Figure 6-8. Two factors have the highest sensitivity towards LOC. The first factor is an inadequate written and operating procedure that comes under management and organizational failure. The second is a sensor failure which is considered an unsafe condition. The results suggest that management, organizational oversight, and unsafe conditions are primary causation factors for LOC and, subsequently, F&E. Recently IChemE (*IChemE Safety and Loss Prevention*, 2022) analyzed major process safety incidents and stated their root causes. There were 52 incidents; out of them, 40 had the common root cause of inadequate procedure. Therefore, the sensitivity analysis result can be validated through the IChemE root cause map. The sensitivity analysis highlighted an important factor that needs proper attention in process industries. Moreover, this result also shows the importance of methodology in developing the hybrid BN model.

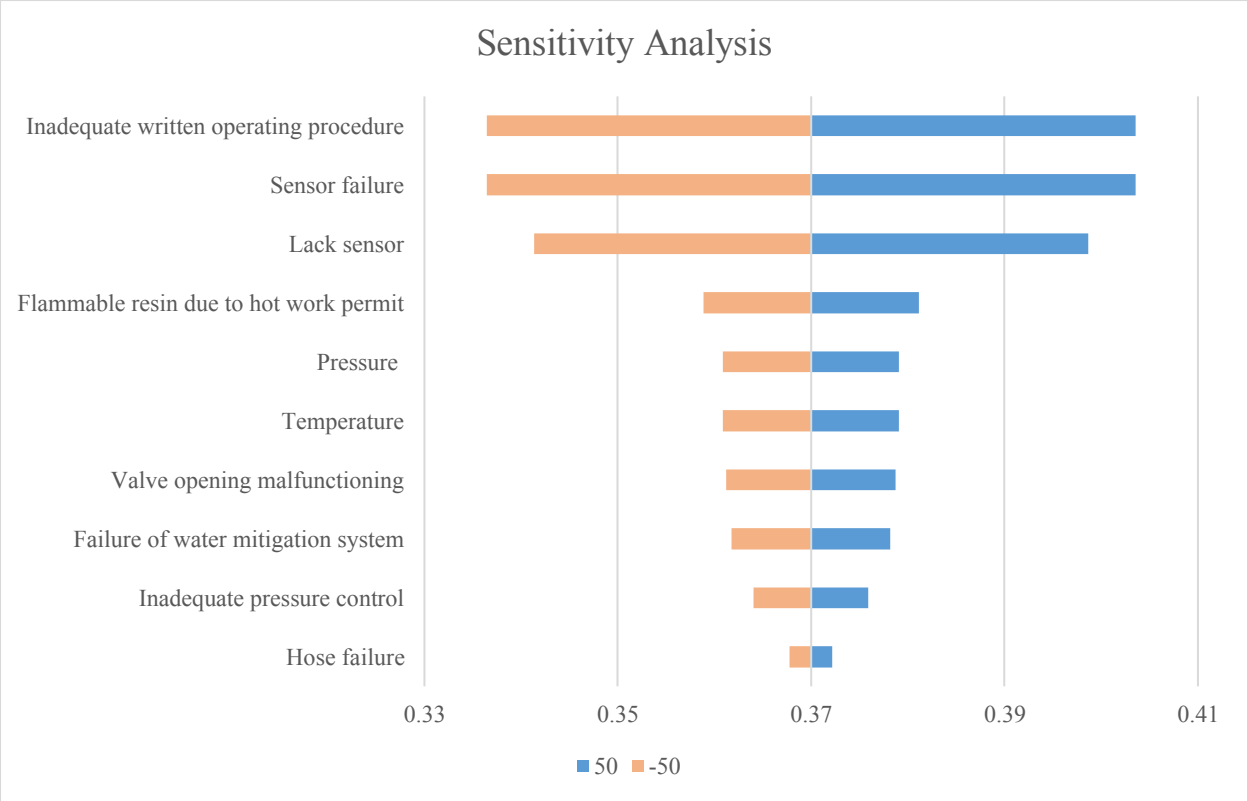


Figure 6-8 Tornado chart to analyze the sensitivity of fuzzy and monitored nodes

6.5 Conclusions

Safety 4.0 demands integrating NLP with a data-driven approach. This study introduces a hybrid BN modeling approach that analyzes, interprets, and organizes multi-source data into meaningful information. The research demonstrates a robust approach to predicting LOC incidents based on past experiences from the CSB database and contemporary data from real-time monitored parameters. Reliance on historical and contemporary data provides a comprehensive picture of accident causation. The knowledge gap of leveraging multi-source heterogeneous data integration for accident causation analysis has been addressed in this study. The analysis unfolds the potential LOC incidents pathways that lead to catastrophe. These pathways can be a precursor to avoiding potential adverse events. The ISM digraph highlights the complex interrelationships of factors associated with LOC incidents. The unique features of this study are as follows:

- Providing a provision to integrate textual data and numerical data into accident likelihood.
- Developing a generalized hybrid BN model for LOC incidents from the past and real-time data.
- Identifying inadequately written procedures and sensor failure as having the highest sensitivity toward LOC incidents.
- Gaining insights from multiple data sources into what went wrong.
- Developing strategies to minimize LOC incidents based on the hierarchy of factors in the digraph.
- Automating causation extraction from the accident database using the co-occurrence network.

The key goal is to develop a safety 4.0 tool that can automate insights from the accident database and leverage real-time data to enhance informed safety-related decisions. Textual and

numerical are two data sources comprised in the hybrid BN model to drive meaningful information. The advantage of this model is the use of NLP to develop causation from unstructured data. NLP helps assess hazards and potential causation pathways. Employing NLP with real-time data introduces a novel way of real-time risk monitoring of LOC incidents. In this way, features from unstructured data (textual data) and combination with structured data (sensor data) capture more system information that is otherwise not possible from either of the data sources. The model consists of factors responsible for LOC involved in real-world industrial accidents and control room sensor data to develop a way to identify precursors for accident causation. The model is verified on 8 major LOC incidents to determine its efficacy in predicting adverse events. Comprehensive validation is challenging due to the unavailability of real-time sensor data.

The uncertainties in the model arise due to the actual sensor data paucity and expert opinion usage in the ISM and BN methods. Moreover, the subjectivity introduced in the filtering step of the co-occurrence network can be reduced by introducing domain expertise in determining words that do not add value to causation. Addressing these concerns can be a direction for future studies.

6.6 Acknowledgments

The first author would like to thank Dr. Md. Tanjin Amin for assisting with the sensor data simulation. The authors acknowledge the financial support provided by Genome Canada and their supporting partners through the Large Scale Applied Research Project; the Canada Research Chair (CRC) Tier I Program in Offshore Safety and Risk Engineering; and the Mary Kay O'Connor Process Safety Center at Texas A&M University, TX, USA.

6.7 References

1. 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*. https://doi.org/10.1016/j.psep.2021.09.022
2. Amin, M. T., Imtiaz, S., & Khan, F. (2018). Process system fault detection and diagnosis using a hybrid technique. *Chemical Engineering Science*, *189*, 191–211. https://doi.org/10.1016/j.ces.2018.05.045
3. Amin, Md. T., Khan, F., Ahmed, S., & Imtiaz, S. (2021). A data-driven Bayesian network learning method for process fault diagnosis. *Process Safety and Environmental Protection*, *150*, 110–122. https://doi.org/https://doi.org/10.1016/j.psep.2021.04.004
4. Amin, Md. T., Khan, F., & Imtiaz, S. A. (2019). Fault detection and pathway analysis using a dynamic Bayesian network. *Chemical Engineering Science*.
5. Amyotte, P., Irvine, Y., & Khan, F. (2018). Chemical safety board investigation reports and the hierarchy of controls: Round 2. *Process Safety Progress*, *37*(4), 459–466. https://doi.org/10.1002/prs.12009
6. Amyotte, P. R., Macdonald, D. K., & Khan, F. I. (2011). An analysis of CSB investigation reports concerning the hierarchy of controls. *Process Safety Progress*, *30*(3), 261–265. https://doi.org/10.1002/prs.10461
7. Attri, R., Dev, N., & Sharma, V. (2013). Interpretive Structural Modelling (ISM) approach: An Overview. In *Research Journal of Management Sciences* (Vol. 2, Issue 2). www.isca.in
8. Bao, H., Khan, F., Iqbal, T., & Chang, Y. (2011). Risk-based fault diagnosis and safety management for process systems. *Process Safety Progress*, *30*(1), 6–17. https://doi.org/https://doi.org/10.1002/prs.10421

9. Baybutt, P. (2016). Insights into process safety incidents from an analysis of CSB investigations. *Journal of Loss Prevention in the Process Industries*, *43*, 537–548. https://doi.org/10.1016/j.jlp.2016.07.002
10. Bhusari, A., Goh, A., Ai, H., Sathanapally, S., Jalal, M., & Mentzer, R. A. (2021). Process safety incidents across 14 industries. *Process Safety Progress*, *40*(1). https://doi.org/10.1002/prs.12158
11. Bloomberg. (2022). *BP's Ohio Refinery May Stay Shut Into 2023 After Deadly Fire*. https://www.bloomberg.com/news/articles/2022-09-27/bp-toledo-refinery-fire-repairs-may-extend-into-early-2023?leadSource=uverify%20wall
12. Chen Shu-Jen and Hwang, C.-L. (1992). Fuzzy Multiple Attribute Decision Making Methods. In *Fuzzy Multiple Attribute Decision Making: Methods and Applications* (pp. 289–486). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-46768-4_5
13. Clemen, R. T., & Winkler, R. L. (1999). Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis*, *19*(2), 187–203. https://doi.org/10.1023/A:1006917509560
14. CSB News Release. (2022). *U.S. Chemical Safety and Hazard Investigation Board*. https://www.csb.gov/csb-releases-new-chemical-incident-data-and-calls-for-increased-attention-to-process-safety-management-during-winter-period-/
15. Ferret, O. (2004). Discovering word senses from a network of lexical cooccurrences. *Proceedings of the 20th International Conference on Computational Linguistics*, 1326–1332.
16. Fu, G., Xie, X., Jia, Q., Li, Z., Chen, P., & Ge, Y. (2020). The development history of accident causation models in the past 100 years: 24Model, a more modern accident causation model. *Process Safety and Environmental Protection*, *134*, 47–82. https://doi.org/10.1016/j.psep.2019.11.027

17. GeNie *software*. (2023). https://www.bayesfusion.com
18. Goodman, I. R. , Mahler, R. P. , & Nguyen, H. T. (2013). *Mathematics of data fusion* (3rd ed., Vol. 37). Springer Science & Business Media.
19. Higuchi, K. (2016). *KH Coder 3 Reference Manual*. https://khcoder.net/en/manual_en_v3.pdf
20. U.K. HSE discovering safety. (2021). *Loss of containment insights project*. https://www.discoveringsafety.com/works/loss-containment-insights-project
21. Huang, W., Zhang, Y., Kou, X., Yin, D., Mi, R., & Li, L. (2020). Railway dangerous goods transportation system risk analysis: An Interpretive Structural Modeling and Bayesian Network combining approach. *Reliability Engineering and System Safety*, *204*. https://doi.org/10.1016/j.ress.2020.107220
22. *IChemE Safety and Loss Prevention*. (2022). https://www.icheme.org/membership/communities/special-interest-groups/safety-and-loss-prevention/resources/lessons-learned-database/
23. 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. https://doi.org/https://doi.org/10.1016/j.psep.2022.11.074
24. Kamil, M. Z., Taleb-Berrouane, M., Khan, F., & Amyotte, P. (2021). Data-driven operational failure likelihood model for microbiologically influenced corrosion. *Process Safety and Environmental Protection*. https://doi.org/10.1016/j.psep.2021.07.040
25. 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*, *00*, 1–20. https://doi.org/https://doi.org/10.1111/risa.14100

26. Kaszniak, M. (2010). Oversights and omissions in process hazard analyses: Lessons learned from CSB Investigations. *Process Safety Progress*, *29*(3), 264–269. https://doi.org/10.1002/prs.10373

27. Li, F., Wang, W., Dubljevic, S., Khan, F., Xu, J., & Yi, J. (2019). Analysis on accident-causing factors of urban buried gas pipeline network by combining DEMATEL, ISM and BN methods. *Journal of Loss Prevention in the Process Industries*, *61*, 49–57. https://doi.org/10.1016/j.jlp.2019.06.001

28. 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. https://doi.org/https://doi.org/10.1016/j.psep.2021.05.036

29. Liu, K., & El-Gohary, N. (2020). Fusing Data Extracted from Bridge Inspection Reports for Enhanced Data-Driven Bridge Deterioration Prediction: A Hybrid Data Fusion Method. *Journal of Computing in Civil Engineering*, *34*(6). https://doi.org/10.1061/(asce)cp.1943-5487.0000921

30. Mannan, M. S., & Waldram, S. P. (2014). Learning lessons from incidents: A paradigm shift is overdue. *Process Safety and Environmental Protection*, *92*(6), 760–765. https://doi.org/10.1016/j.psep.2014.02.001

31. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.

32. NBC Boston. (2022). *Day After Deadly Gas Leak, Ammonia Levels Drop at Norwood Food Facility*. https://www.nbcboston.com/news/local/probe-underway-after-ammonia-leak-leaves-one-dead-in-norwood/2924959/

33. NRDC. (2023, February 21). *Ohio Train Disaster Reveals Gaping Holes in Hazardous Chemical Controls*. https://www.nrdc.org/bio/jennifer-sass/ohio-train-disaster-reveals-gaping-holes-hazardous-chemical-controls
34. Nurmi, H. (1981). Approaches to collective decision making with fuzzy preference relations. *Fuzzy Sets and Systems*, *6*(3), 249–259. https://doi.org/https://doi.org/10.1016/0165-0114(81)90003-8
35. Onisawa, T. (1988). An approach to human reliability in man-machine systems using error possibility. *Fuzzy Sets and Systems*, *27*(2), 87–103. https://doi.org/https://doi.org/10.1016/0165-0114(88)90140-6
36. Ramzali, N., Lavasani, M. R. M., & Ghodousi, J. (2015). Safety barriers analysis of offshore drilling system by employing Fuzzy Event Tree Analysis. *Safety Science*, *78*, 49–59. https://doi.org/https://doi.org/10.1016/j.ssci.2015.04.004
37. Reuters. (2022). At Least 10 People Killed in India Factory Explosion. https://www.reuters.com/world/india/least-six-killed-india-chemical-factory-explosion-2022-06-04/
38. Romesburg, C. (2004). *Cluster analysis for researchers*.
39. Saeed, M. S., Halim, S. Z., Fahd, F., Khan, F., Sadiq, R., & Chen, B. (2022). An ecotoxicological risk model for the microplastics in arctic waters. *Environmental Pollution*, *315*, 120417. https://doi.org/https://doi.org/10.1016/j.envpol.2022.120417
40. Sajid, Z., Khan, F., & Zhang, Y. (2017). Integration of interpretive structural modelling with Bayesian network for biodiesel performance analysis. *Renewable Energy*, *107*, 194–203. https://doi.org/10.1016/j.renene.2017.01.058
41. Sugeno, M., & Kang, G. T. (1986). Fuzzy modelling and control of multilayer incinerator. *Fuzzy Sets and Systems*, *18*(3), 329–345. https://doi.org/https://doi.org/10.1016/0165-0114(86)90010-2

42. The Times of India. (2022). *Ammonia leak at meat plant in Aligarh leaves 59 unconscious*. https://timesofindia.indiatimes.com/city/agra/ammonia-leak-at-meat-plant-in-aligarh-leaves-59-unconscious/articleshow/94545945.cms
43. U.S. Chemical Safety and Hazard Investigation Board. (2022). Incident Reporting Rule Submission Information and Data. https://www.csb.gov/news/incident-report-rule-form-/
44. Wang, F., Gu, W., Bai, Y., & Bian, J. (2023). A method for assisting the accident consequence prediction and cause investigation in petrochemical industries based on natural language processing technology. *Journal of Loss Prevention in the Process Industries*, *83*, 105028. https://doi.org/https://doi.org/10.1016/j.jlp.2023.105028
45. Warfield, J. N. (1974). Developing Interconnection Matrices in Structural Modeling. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-4*(1), 81–87. https://doi.org/10.1109/TSMC.1974.5408524
46. Wu, W. S., Yang, C. F., Chang, J. C., Château, P. A., & Chang, Y. C. (2015). Risk assessment by integrating interpretive structural modeling and Bayesian network, case of offshore pipeline project. *Reliability Engineering and System Safety*, *142*, 515–524. https://doi.org/10.1016/j.ress.2015.06.013
47. Yu, J., & Rashid, M. M. (2013). A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, *59*(7), 2348–2365. https://doi.org/10.1002/aic.14013
48. Yuan, C., Cui, H., Ma, S., Zhang, Y., Hu, Y., & Zuo, T. (2019). Analysis method for causal factors in emergency processes of fire accidents for oil-gas storage and transportation based on ISM and MBN. *Journal of Loss Prevention in the Process Industries*, *62*. https://doi.org/10.1016/j.jlp.2019.103964
49. Zadeh, L. A. (1965). Fuzzy Sets. *Informational and Control*, *8*, 338–353.

50. Zarei, E., Khakzad, N., Cozzani, V., & Reniers, G. (2019). Safety analysis of process systems using Fuzzy Bayesian Network (FBN). *Journal of Loss Prevention in the Process Industries*, *57*, 7–16. https://doi.org/https://doi.org/10.1016/j.jlp.2018.10.011
51. Zhang, Z., Zweigenbaum, P., & Yin, R. (2018). Efficient Generation and Processing of Word Co-occurrence Networks Using corpus2graph. *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, 7–11. https://doi.org/10.18653/v1/W18-1702

Appendix

Table 6-5 Structural self-interaction matrix (SSIM) Created using pair-wise comparison of each factor

Factors	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	**XIV**	**XV**	**XVI**	**XVII**	**XVIII**	**XIX**	**XX**
I	X	A	O	O	O	V	O	O	O	O	O	O	0	O	0	O	A	O	O	O
II		X	O	A	O	O	A	A	O	O	O	V	O	O	V	O	O	O	O	O
III			X	O	O	V	O	O	O	O	O	O	O	O	O	O	O	O	O	O
IV				X	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
V					X	O	A	O	A	A	A	V	A	A	O	O	O	O	O	O
VI						X	O	O	O	O	O	O	O	O	O	O	O	O	O	O
VII							X	O	O	O	O	A	O	O	O	A	O	O	O	O
VIII								X	O	O	O	O	O	O	O	O	O	O	A	A
IX									X	O	O	O	V	O	O	O	O	O	A	A
X										X	O	O	O	O	O	V	O	O	O	O
XI											X	O	O	O	O	O	O	O	O	O
XII												X	O	O	O	O	V	O	O	O

	XIII													X	O	O	O	O	O	O	O
	XIV														X	O	O	O	O	O	O
	XV															X	O	V	O	O	A
	XVI																X	V	O	O	O
	XVII																	X	A	O	O
	XVIII																		X	O	O
	XIX																			X	O
	XX																				X

Table 6-6 Final reachability matrix (FRM)

Factors	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	Driving Power
I	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
II	1	1	0	0	0	1*	0	0	0	0	0	1	0	0	1	0	1*	0	0	0	6
III	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
IV	1*	1	0	1	0	1*	0	0	0	0	0	1*	0	0	1*	0	1*	0	0	0	7

V	1*	0	0	0	1	1*	0	0	0	0	0	1	0	0	0	0	1*	0	0	0	5
VI	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
VII	1*	1	0	0	1	1*	1	0	0	0	0	1*	0	0	1*	0	1*	0	0	0	8
VIII	1*	1	0	0	0	1*	0	1	0	0	0	1*	0	0	1*	0	1*	0	0	0	7
IX	1*	0	0	0	1	1*	0	0	1	0	0	1*	0	0	0	0	1*	0	0	0	6
X	1*	0	0	0	1	1*	0	0	0	1	0	1*	0	0	0	1	1*	0	0	0	7
XI	1*	0	0	0	1	1*	0	0	0	0	1	1*	0	0	0	0	1*	0	0	0	6
XII	1*	0	0	0	0	1*	0	0	0	0	0	1	0	0	0	0	1	0	0	0	4
XIII	1*	0	0	0	1	1*	0	0	0	0	0	1*	1	0	0	0	1*	0	0	0	6
XIV	1*	0	0	0	1	1*	0	0	0	0	0	1*	0	1	0	0	1*	0	0	0	6
XV	1*	0	0	0	0	1*	0	0	0	0	0	0	0	0	1	0	1	0	0	0	4
XVI	1*	0	0	0	0	1*	0	0	0	0	0	0	0	0	0	1	1	0	0	0	4
XVII	1	0	0	0	0	1*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3
XVIII	1*	0	0	0	0	1*	0	0	0	0	0	0	0	0	0	0	1	1	0	0	4
XIX	1*	1*	0	0	1*	1*	0	1	1	0	0	1*	0	0	1*	0	1*	0	1	0	10
XX	1*	1*	0	0	1*	1*	0	1	1	0	0	1*	0	0	1*	0	1*	0	0	1	10

Dependence Power	18	6	1	1	9	20	1	3	3	1	1	13	1	1	7	2	17	1	1	1	

Table 6-7 Level Partitioning

Elements (Xi)	Reachability Set R(Xi)	Antecedent Set A(Xi)	Intersection Set R(Xi)∩A(Xi)	Level
I	1,	1.2,4,5,7,8,9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,20	1,	2
II	2,	2,4, 7, 8, 19, 20,	2,	5
III	3,	3,	3,	2
IV	4,	4,	4,	6
V	5,	5,7,9, 10, 11, 13, 14, 19, 20,	5,	5
VI	6,	1,2.3,4,5,6,7,8,9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,20,	6,	1
VII	7,	7,	7,	6
VIII	8,	8, 19, 20,	8,	6
IX	9,	9, 19,20,	9,	6
X	10,	10,	10,	6

XI	11,	11,	11,	6
XII	12,	2,4,5,7,8,9, 10, 11, 12, 13, 14, 19, 20,	12,	4
XIII	13,	13,	13,	6
XIV	14,	14,	14,	6
XV	15,	2,4,7,8, 15, 19, 20,	15,	4
XVI	16,	10,16,	16,	4
XVII	17,	2,4,5,7,8,9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,	17,	3
XVIII	18,	18,	18,	4
XIX	19,	19,	19,	7
XX	20,	20,	20,	7

Table 6-8 Conical matrix

Factors	**VI**	**I**	**III**	**XVII**	**XII**	**XV**	**XVI**	**XVII**	**II**	**V**	**IV**	**VII**	**VII**	**IX**	**X**	**XI**	**XII**	**XIV**	**XIX**	**XX**	**Driving Power**	**Level**

VI	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
I	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
III	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
XVII	1*	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
XII	1*	1*	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
XV	1*	1*	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
XVI	1*	1*	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4
XVIII	1*	1*	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4	4
II	1*	1	0	1*	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	6	5
V	1*	1*	0	1*	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	5	5

IV	1*	1*	0	1*	1*	1*	0	0	1	0	1	0	0	0	0	0	0	0	0	0	7	6
VII	1*	1*	0	1*	1*	1*	0	0	1	1	0	1	0	0	0	0	0	0	0	0	8	6
VIII	1*	1*	0	1*	1*	1*	0	0	1	0	0	0	1	0	0	0	0	0	0	0	7	6
IX	1*	1*	0	1*	1*	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	6	6
X	1*	1*	0	1*	1*	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	7	6
XI	1*	1*	0	1*	1*	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	6	6
XIII	1*	1*	0	1*	1*	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	6	6
XIV	1*	1*	0	1*	1*	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	6	6

XIX	1*	1*	0	1*	1*	1*	0	0	1*	1*	0	0	1	1	0	0	0	0	1	0	10	7
XX	1*	1*	0	1*	1*	1*	0	0	1*	1*	0	0	1	1	0	0	0	0	0	1	10	7
Dependence Power	20	18	1	17	13	7	2	1	6	9	1	1	3	3	1	1	1	1	1	1		
Level	1	2	2	3	4	2	4	4	5	5	6	6	6	6	6	6	6	6	7	7		

7 Summary, Conclusions and Recommendations

7.1 Summary

The advancement in process operations requires advanced data-driven approaches for safety. This study presents advanced probabilistic methods for developing an accident causation model that supports safety 4.0 in process operations. The main contribution of this work is to extend the modeling power of Bayesian networks (BN) by introducing novel algorithms that use structured and unstructured data.

The thesis comprises various complex failure scenarios from an application perspective. Failures from offshore platforms due to MIC in process industries oil and refining, LOC, and fire & explosion are covered in the thesis. The modeling perspective introduces several innovations in developing the learning-based model, which offers a new dimension to risk analysis. These innovations incorporate NLP to develop objective risk models and generalized hierarchical causation models.

The thesis offers a direct solution to learning from structured data instead of relying on process knowledge. The integration of operational and microbiological data are shown to predict the occurrence of MIC. The salient features are the ability to learn from data and the handling of missing values. The introduction of LBN is beneficial to determine vulnerable process equipment.

Unstructured data are never used to analyze what went wrong. With the advancement of NLP, it is possible to text-mine data and automate the feature extraction process. The NER model is trained and tested to gain insights into MIC accidents. A novel algorithm for mapping NER to BN is introduced to develop the risk model from an accident database.

Employing NLP creates opportunities to develop new models, such as the generalized causation model, by introducing a systematic approach of combining three techniques NLP,

ISM and BN. A novel algorithm for ISM mapping into BN is also introduced. The advantage is to analyze various accident pathways and understand commonalities among accidents.

The generalized causation model approach is further extended to incorporate structured data to gain more insight and monitor risk. The co-occurrence network method is used to automate insights from the accident database, ISM, and BN to transform into an accident likelihood model. The thesis provides a paradigm shift into developing a causation model from multi-source data. The structured and unstructured datasets used in the thesis are obtained from industry and accident investigation reports of PHMSA and CSB.

7.2 Conclusions

The specific conclusions are listed below.

7.2.1 Development of a learning-based likelihood model

Advancing towards process digitalization demands an approach that relies on data and is adaptable to ensure the safety of a process operation. An integrated model is introduced that can use field and laboratory data to assess MIC threats. The proposed model, called the LBN model, possesses several advantages over existing MIC models. It exhibits the ability to learn BN structure and parameters from data even though the dataset comprises missing values. The LBN model has undergone testing and validating using the training and testing data set, demonstrating acceptable performance and establishing the model's efficacy in predicting the MIC likelihood. By employing a data-driven approach, valuable insights can be gained to identify the vulnerable process equipment susceptible to MIC. Consequently, the model can be utilized to assess MIC risk in CPI and improve overall operational safety.

7.2.2 Risk estimation and evaluation from textual data

A new dimension to risk assessment is introduced by gaining insights from textual data. Automated feature extraction from a database enables the analysis of the causes and consequences of incidents. A unique approach is introduced that combines NER with BN, using a defined mapping algorithm. The approach has unique aspects, including its ease of implementation, self-explanatory feature labeling, incremental annotations, and applicability to various domains. The methodology demonstrated a new way of analyzing accident scenarios. Five PHMSA database incidents are taken to evaluate objective risk from the textual data and verify the introduced methodology's applicability. The risk levels are deemed unacceptable and consistent with actual conditions. A total of 8 causation models are developed from the trained NER model to establish NER efficacy. Unlike previous studies limited to feature extraction, this methodology provides a pathway for predicting risk from textual data. The development of BN from textual data demonstrated that unstructured textual data are a valuable source and can be used to assess objective risk.

7.2.3 Generalized causation likelihood analysis

A novel approach is introduced to developing a generalized causation model for oil and refining incidents. A systematic approach is introduced to gain insights from the accident database and transform it into a generalized model. The approach comprises NER, ISM and BN. The output from NER serves as an input to ISM, and the output from ISM is mapped to BN using a novel mapping algorithm. The developed BN model unfolds critical factors and commonalities among accidents. The model is tested on 10 incidents and verified on 6 incidents resulting in 100% likelihood which is coherent with actual conditions. Sensitivity analysis shows that MoC and lack of procedure and training are the highest sensitive parameters towards F&E. Management and regulatory oversights drive F&E accidents. The introduction of generalized

causation modeling is an important step toward accident causation analysis. It helps to analyze each incident pathway, assess commonalities and develop strategies based on the hierarchy of factors.

7.2.4 Multi-source data integration for generalized causation analysis

A robust approach is introduced to integrating textual data and numerical data into a likelihood model. Historical and contemporary data combined provide a comprehensive causation model. A hybrid BN model is developed for LOC incidents. Historical data, i.e., unstructured in nature due to textual data, from accident investigation reports are used to develop insights and integrate with real-time monitoring data from sensors. This way, features from accident investigation reports and real-time monitoring capture more information about what went wrong, which is otherwise not possible using a single source. The co-occurrence network is used combined with ISM and BN techniques to develop an approach that can integrate multi-source data. The hybrid BN provides meaningful information from two data sources. The model provides a new way of monitoring LOC accident risk based on real-time data and past accident causation factors. The model is useful for predicting precursors to LOC adverse events. Inadequately written procedures and sensor failure are the highest sensitive parameters to LOC. The model is tested and verified on 8 LOC accidents to predict LOC likelihood and F&E likelihood. In all cases, the model predicted 100% LOC and 11% when no F&E was reported and 100% F&E when an ignition source was present. Hence, the model is capable of predicting LOC adverse events.

7.3 Recommendations

This doctorate thesis introduces new concepts and addresses the shortcomings of existing techniques and limitations in process safety and risk analysis of oil and gas facilities and CPI. Nevertheless, this study can be further extended to incorporate the following recommendations.

7.3.1 Data requirements

The approaches developed in the thesis demand high quality and quantity of data which is often difficult to obtain. One of the benefits of the LBN model is the ability to learn BN's topology and parameters even when missing values are in the input dataset. According to the result, 150 data points are required for the LBN model stability comprised of 10 nodes. In the LBN study, the industry partner provides data but not exhaustive data points. Therefore, data are simulated between the lower and upper bound of the provided values. The model's performance is evaluated to check how many missing values can be handled by the model. The study does not aim to address handling missing values in process data. Further research is required to investigate how to deal with missing values. Moreover, the LBN can be improved to incorporate incremental learning. The incremental learning process, again, is data intensive.

Likewise, pair-wise comparison in the ISM requires comprehensive data to establish interrelationships among factors. The influence can only be established if the individual factor's relationship is known. When the ISM digraph is mapped into BN, CPTs must be defined, which requires considerable data. Thus, this makes expert opinion inevitable. Therefore, a database of structured data needs to be developed to leverage for assessing the relationship between factors, process monitoring, and developing risk assessment models. The unstructured textual databases like U.S. CSB and PHMSA for accidents are a valuable resource extensively utilized in the thesis. Similarly, a database of numerical data can be of great importance to learning from past experiences and fulfilling the data requirements of the data-driven approaches mentioned earlier.

7.3.2 Automated causation extraction

NER provides cause-effect features from an accident database. Custom NER is beneficial for extracting custom entities from the database, and the labels attached to them require less manual interpretation to develop a causation model. However, the manual interpretation step can be automated using custom NER with relation extraction. This can be achieved by incorporating automated relation extraction of entities, illustrating cause-effect scenarios using named entities and their relationships, thus eliminating manual interpretations. The expected outcome will be automated extraction of causation from textual data.

The spaCy library used to train custom NER also allows for relation extraction. Custom NER and relation extraction can be jointly performed during the annotation process. Thus, an automated causation analysis can be beneficial for learning lessons from past events.

7.3.3 Uncertainty handling

The thesis aimed to introduce methodologies to make informed safety-related decision-making rather than numbers accuracy. Both aleatory and epistemic uncertainties play an important role due to the stochastic nature of adverse events and incomplete information in modeling incidents in the study. Therefore, both uncertainties need to be overcome in future work. The epistemic uncertainty can be addressed using Dempster Shafer's theory and developing an evidential network instead of BN. Like BN, an evidential network is a directed acyclic graph for propagating epistemic uncertainty within a system's elements based on conditional belief mass. The evidential network is important in system safety when dealing with scarcity of accurate data, thus introducing epistemic uncertainty on the child node's belief mass from the parent nodes' belief mass.

In the case of aleatory uncertainty, Monte Carlo simulation technique can propagate aleatory uncertainty through the Bayesian network. Generate random samples from the assigned probability distributions and simulate the network repeatedly, incorporating the uncertainty in each iteration. This approach allows for estimating probabilistic outcomes, considering the aleatory uncertainty in the model.