Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
DOI : 10.47974/JIOS-2189

Integrating text mining and Bayesian networks for root-cause identification of near-miss incidents to enhance safety measures

Abhishek Verma *
Department of Management
Birla Institute of Technology and Science
Pilani
Jhunjhunu District 333031
Rajasthan
India

Deeshant Rajput †
Glance Digital Experience Private Limited
Bangalore
Karnataka
India

Saptarshi Mandal §
Fractal.ai
Bangalore
Karnataka
India

Jhareswar Maiti ‡
Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Kharagpur 721302
West Bengal
India

Abstract

Near-miss incidents (as leading safety indicators) provide implicit signals regarding the various limitations of the prevalent safety management system at a workplace. Implementing

* *E-mail:* **abhishekverma.cs@gmail.com (Corresponding Author)**
† *E-mail:* **deeshantrajput@gmail.com**
§ *E-mail:* **pikumandal123@gmail.com**
‡ *E-mail:* **jhareswar.maiti@gmail.com**

© Taru Publications

proactive safety measures requires the identification of root causes and, therefore, warrants in-depth analysis of these near-miss incidents. Effective root-cause management is crucial to avoid these incidents and improve overall safety. However, identifying root causes from the raw text-based near-miss incident reports is challenging because they are unstructured. To this purpose, we provide a methodology that combines text mining and Bayesian Networks to identify root causes and develop a casual knowledge model. The developed model has been used to analyse seven near-miss incident scenarios for validation. The findings indicate that incidents are more likely to occur due to factors such as defective switches or cables, poor stair or floor conditions, damaged wires, ropes, or slings, unsafe crane operations, malfunctioning valves or hoses, and the involvement of heavy vehicles like trucks and dumpers during loading or transportation activities. This work contributes to dealing with narrative text data and raising the organisation's awareness of various circumstances in which these incidents occur.

Subject Classification: *Primary 68T50, Secondary 62F15.*

Keywords: *Near-miss, Bayesian network, Word cloud, Visualization, Text mining, Safety, Steel plant.*

1. Introduction

Industrial accidents can cause loss of human capital, including competencies, knowledge, creativity, etc., and economic, production, and technology damage. While earlier studies in the field of safety management primarily concentrated on examining and interpreting past accident data [1][2][3][4], recently research trends shifted toward preventive approaches to enhance employee protection and strengthening organizational safety culture. From managerial standpoint, there is growing recognition of the need for adopting proactive/leading safety indicators (e.g. near-miss incidents) as more effective tool that than traditional lagging indicators (i.e. accident data). Analysis of near-miss incident data combined with expert opinion can guide decision-makers to take necessary steps to ensure enhanced safety management in organizations.

Schaaf in 1992 [5] defined a near-miss incident as any potential incident detected, diagnosed, and corrected accurately in time without causing any actual consequence (accident). Occupational Safety and Health Administration (OSHA) describes near-miss as "incidents where no property was damaged and no personal injury sustained, but where, given a slight shift in time or position, damage and/or injury easily could have occurred". Additionally, unsafe behaviour or conditions have been categorised as near-miss events [6][7]. Furthermore, near-miss incidents have been identified as practical as they provide an opportunity to

undertake proactive measures to improve environmental, health, and safety practices, which can lead to severe consequences if neglected [6].

To understand a near-miss incident, consider a scenario where a worker crosses a conveyor belt, carries crushed coke in idle condition, and suddenly starts moving (dangerous occurrence). There can be three possible scenarios: first, an accident scenario where the worker slips and falls; second, a near-miss scenario (involving chance factor), with the worker managing to cross the belt before it starts moving; and third, another near-miss scenario (due to management guidelines), where the worker recalls that crossing the conveyor belt was prohibited and restrains.

The traditional safety management approach entails detailed analysis of accident and injury data to unearth their root causes and guide operational changes in the workplace. However, these incidents are rare; as a result, extracting useful information that can be transformed into actionable inputs becomes challenging. Unlike this, the frequency of near-miss incidents is substantially high [6][8][9], thus making it effective for quantitative prediction models. Also, note that near-miss incident reports contain information similar to the accident reporting process. Moreover, using near-miss incident data (a pseudo-leading indicator of an organisation's safety performance) enables the decision-maker to develop proactive measures. For example, Phimister et al. [6] and Schaaf [5] showed that industries that collect more near-miss incidents are safer than industries that don't.

Hence, it is essential to encourage employees to report near-miss indents to their supervisors using various reporting modes such as written, verbal, or electronic forms. After disclosing the incident, it should be resolved effectively and timely so that the reoccurrence of the same could be avoided. Other than finding the root cause, taking corrective action, and following up, there is a need to find the mapping between a near-miss incident and actual injury cases, predict future incidents, and then apply an appropriate mitigation approach.

A quantitative evaluation of near-miss occurrences is crucial because these event may escalate into serious accidents if proper control measures or preventive constraints are not applied in time. A successful assessment can provide the consequences of such failures and recommend prioritization strategies for capital and human resource spending. Prioritized near-miss incidents are identified as having more potential for developing into accidents [10]. The prioritized incidents require further investigation, possibly by adding new safeguards or making existing ones more reliable or robust. This study aims to investigate the potential of

near-miss incidents reported in the safety management system of the steel plant due to the complex, diverse and hazardous nature of tasks and working environment. The workforce of the steel plant is also diverse in terms of training and safety culture because it is a mixture of permanent and contractual workers. To achieve this objective, both structured and narrative text data were examined using Bayesian network (BN) analysis to gain deeper insights into the factors contributing to incidents. Text mining is used to find the factors behind the primary causes logged onto the system by examining the narrative text because it provides qualitative insight into the incident process. It will give the prior probability of the causal factors feeding in BN to find probabilistic relationships among different aspects. The developed framework has been implemented to prioritize the underlying causes associated with near-miss cases. Techniques are used for overall prioritization of near-miss reporting status, considering the data of April 2010 - Dec 2013 to support decision-making for safety and budget allocation to mitigate the causes for safety performance improvement.

The structure of the remaining paper is as follows: Section 2 presents the literature review. Methodology, data collected from the industry and selection of factor is discussed in Section 3. Employed techniques, word cloud and Bayesian network are also discussed in Section 3. Section 4 outlines the outcomes obtained from the model along with their practical implications, while Section 5 summarizes the conclusions of the study and highlights its limitations.

2. Literature review

Research on near-miss management remains at nascent stage, with numerous challenges, and opportunities for development and structuring. The first and most comprehensive study on near-miss management was conducted by Schaaf [5]. The thesis presents a comprehensive framework for (1) classifying near misses, (2) analyzing near-miss reports, and (3) implementing recommendations derived from near-miss incident analysis. Phimister et al. [6] proposed a reference framework for designing a near-miss management system. Oktem, Wong, and Oktem [11] also presented a model for creating a near-miss management system. This model outlines various stages, starting with identifying the near-miss incident and ending with implementing a solution to prevent near misses. Gnoni et al. [12] introduced a similar modelling approach that leverages lean thinking. All of these researchers focused on two main areas: firstly, ensuring accurate

documentation of near-miss accidents, and secondly, analyzing near-miss events to gain a deeper understanding of the reported data. A well-structured database was required to capture the underlying reasons behind various types of occurrences.

Basso et al. [13] developed a database system for incident investigation designed to record the causes of specific incidents and the subsequent corrective actions. This framework enables the monitoring of safety management system performance. Nivolianitou et al. [14] also demonstrated the creation of a database design to record the reasons and factors contributing to incidents, accidents, and near-miss events in the petrochemical industry. This database aims to assist users in conducting statistical analyses. The incident databases assist in classifying incidents such as occupational incidents, process incidents, operational incidents, and near misses, as well as creating safety indicators. The shortcomings of incident reporting databases in accurately capturing both the quality and the quantity of information have been addressed by Schaaf and Kanse [15]. Efforts have been made in the construction sector to establish a system for reporting and analyzing near misses, identifying patterns, and tracking incidents in real time [16][17][18][19]. The influence of near-miss events on future decision-making of organization managers to improve safety performance was thoroughly studied by Dillon & Tinsley [20]. Therefore, it is crucial to learn from the analysis of near-miss data and prioritize the main causal factors behind near-miss incidents.

The significance of analyzing data on near-miss incidents for enhancing safety inside the organization was studied [21][22][23], and an inverse relationship between near-misses and accident occurrences was found. However, near-miss incidences were classified as a successful outcome in their research. This viewpoint was challenged by Dillon and Tinsley [20], who claimed that near-miss events should be seen as failures to promote risk reduction. Gnoni and Lettera [24] proposed a methodology for assessing the potential risk of near-miss incidents using matrices and indices. The advantages and drawbacks of both approaches were also highlighted. Dee, Cox, and Ogle [25] proposed a game theory approach to assess risk perception using data on near-miss events. They also assigned a value to these events to incentivize top management to take steps to reduce hazards.

Bayesian auto-coding approaches have been utilized in the fire service industry to reduce the manual coding needed for large descriptive datasets of near-miss data [26]. BN can integrate historical data from the incident database with the expertise of professionals. The algorithm can provide

precise and computationally efficient estimations for massive datasets [27] [28]. Neil, Fenton, and Tailor [29] found that it has the potential to enhance the precision of the model. In their study, Bier and Mosleh [30] utilized a probabilistic model to assess near-miss incidents in nuclear plants. They concluded that when making decisions about the frequency of accidents in an organization, near-miss events should be given more significant consideration than experts' claims. Data on accident precursors have been utilized to predict aberrant occurrences (such as near-misses, incidents, and accidents) at chemical plants [31].

The formulation of project risk management in the Korean ship industry was facilitated by using Bayesian Networks (BN). The primary hazards associated with the projects include modifications to the design, availability of design personnel, procurement of raw materials, and fluctuations in exchange rates. Within medium-sized shipbuilding enterprises, the primary risks are related to adhering to the designated timeline and budget. Conversely, in larger companies, the challenge lies in minimizing the risk of dissatisfaction with the specifications [32]. A previous study evaluated and ranked the risks associated with water main failures using a Bayesian belief network (BBN) model [33]. The BBN framework has further been employed to prioritize various situations using weighted scores derived from expert opinions. Recently, researchers have integrated the Bayesian network with analytic hierarchy process (AHP) to identify the importance of factors contributing to near-miss incidents [34][35]. In these studies, the prior probabilities of the factors were determined based on the expert evaluation. Experts are often scarce in most sectors. Additionally, it is advisable to incorporate the near-miss incidences reported by organization personnel.

There is a dearth of studies on the near-miss management of an integrated steel plant. The steel industry is more complex and diverse because it combines processes, workplaces, and construction-related factors [36]. Our study incorporates inputs from all the workers who logged the incident rather than relying solely on the expert judgements to define prior probabilities in Bayesian network. Insights from the industry visits and discussions with senior management revealed several common challenges in incident prioritization: (a) time and labour cost, (b) the absence of effective tools to distinguished between high and low priority incidents, (c) sometimes report does not reach to relevant experts due to incomplete reporting, and (d) inadequate understanding of the specific incidents.

3. Methodology

Building a Bayesian Network (BN) is a highly intricate approach to problem-solving, with its network topology tailored to the individual problem at hand. This study creates BN hierarchies using text-mining concepts and converting the identified components into a BN framework. The conceptual framework of this study illustrated in Figure 1, integrates qualitative and quantitative aspects derived from earlier studies. The network structure of a Bayesian Network (BN) is considered the qualitative aspect since it focuses on detecting event factors. On the other hand, assessing the probability of these factors and making inferences from the network is considered the quantitative aspect. The work incorporated expert experiences and data to establish meaningful extra relationships between BN nodes in the model. The factors, or root causes, are determined by applying text mining techniques to the narrative text and categorizing them as coded primary causes. Multiple components can sometimes be shared among coded primary causes, which can only be derived from unstructured text. The components were derived from the word cloud by utilizing expert opinion, taking into account the root cause's role in the

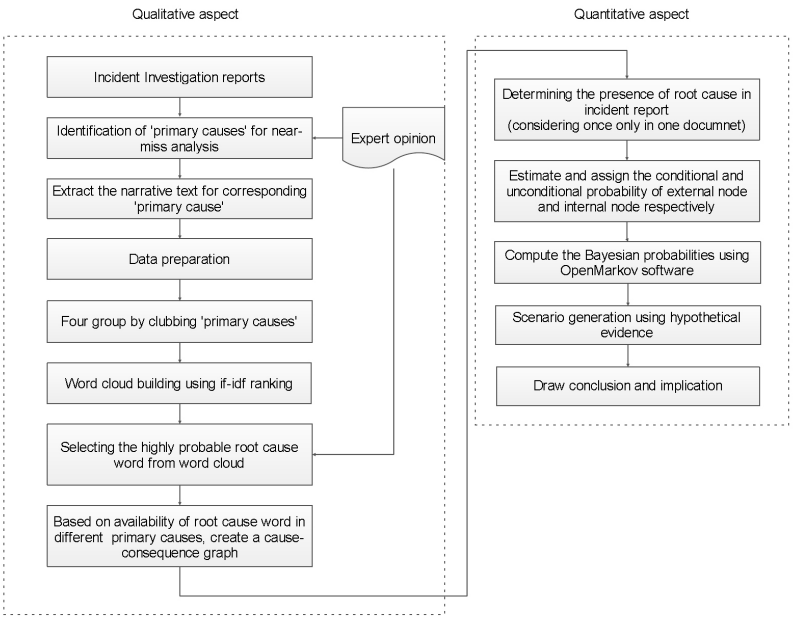


Figure 1

Conceptual framework of the study

occurrence of the incident. The prior probability for each factor is estimated by calculating its occurrence frequency within event reports linked to the corresponding primary cause. Each factor is considered only once per incident report. To establish the cause-effect relationship among the dependent nodes in a Bayesian Network (BN), conditional probabilities are then computed based on the prior probabilities of the root causes. Subsequently, hypothetical evidence is examined to determine the impact on the Bayesian Network when no identifiable root cause exists. Scenarios were created to determine the primary underlying factors most impacted.

3.1 *Case-study*

Efficient safety management is a necessity. The organization under consideration in this study comprises seven primary divisions: Coke, Sinter and Iron Division, Steel Making Division, Wire Making Division, Raw Material Handling Division, Maintenance Division, Operations Division, and Administration Division. The steel plant has implemented an intranet-based safety management system (SMS) to enhance safety performance and ensure accessibility throughout the organization's local network. The plant gathers information about safety performance through incident investigation reports using this online system. Incident investigation data includes information about near-miss incidents, injuries, and property damage. There is currently no database specifically designed to record near-miss incidents. Verma et al. [1] briefly outlined the comprehensive data-gathering process. Near-miss data are collected using two different methods. In the first approach, regular employees document incidents directly in the SMS. In the second approach, contractor supervisors record near-miss incident manually on standardized paper template. The digital version of each report is then emailed to the safety manager of the relevant section or subsection, who subsequently uploads the information into the online SMS platform. For this study, the authors extracted the compiled incident investigation data in MS Excel for further analysis. Each incident is assigned a particular 'Primary cause' code, which can be selected from the dropdown menu in the SMS system. Twenty-three key reasons were identified in the SMS database for near-miss, accident, and property damage incidents. After consulting with a safety expert, 16 of the 23 primary reasons were selected for causation investigation. The criteria and field to be included or excluded were determined through a collaborative brainstorming session with industry professionals whose data was evaluated for the study.

3.2 *Data preparation*

Data preparation addresses several issues, including missing values, duplicate entries, typographical errors, non-vocabulary terms, and incomplete information. Duplicate records were removed using the built-in "Remove Duplicate" feature available in the MS Excel. The MS Excel feature and manual examination eliminated misspelt and irrelevant text. The raw file is taken from the database had a total of 9086 records. Following the pre-processing stage, 8877 records were selected for subsequent analysis. Out of the 8,877 data mentioned in SMS, 2,984 occurrences were determined to be near-miss incidents after evaluating 16 primary causes.

3.3 *Word cloud*

Word clouds have become increasingly popular for visually representing textual data in graphs. It can extract information from the content and concisely summarise the key elements. Typically, a word cloud algorithm arranges words one at a time in a greedy manner. It optimally allocates each word inside the given space. The words were arranged according to a ranking function that evaluates their significance using metrics like term frequency and inverse document frequency. The words can be arranged in many orientations, such as horizontally, vertically, or a mixed fashion. In our study, to find the valuable term out of different primary causes, the 'comparison' and 'commonality' word cloud is implemented using the R language code. A brief overview of the process for generating a word cloud with R programming is presented below:

1. The short narratives of each incident are extracted to determine their respective primary causes. Individual MS Excel files are then created for each primary causes to avoid overlap or mixing of text data from different categories.
2. The free text provided under the 'short of description incident' column for each primary cause is processed by converting the text to lowercase, deleting digits, punctuation, and additional white space. Additionally, stop-words such as "a", "an", "the", "and", and "or" were eliminated.
3. The next phase involves constructing a corpus and generating a structured data matrix from a text file. A corpus is created by

utilizing the 'Corpus()' function, whereas a structured dataset is generated by employing the 'TermDocumentMatrix()' function.

4. The ranking in the model is determined by the tf-idf statistic, where *term frequency (tf)* represents how often a word appear within a specific document, and *inverse document frequency (idf)* indicate the extent of information the word provide across the entire dataset.

$$tf(t,d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

$$idf(t,D) = \ln \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$tfidf(t,d,D) = tf(t,d).idf(t,D)$$

$f_d(t)$ = frequency of term t in document d

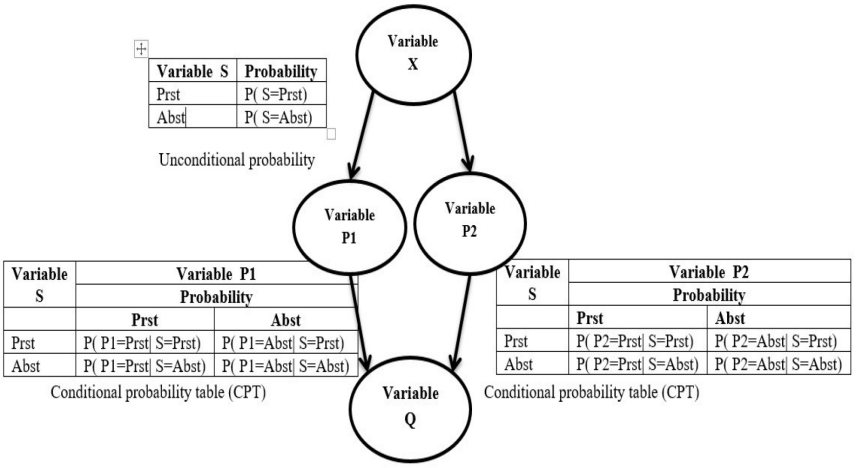
D = corpus of documents

5. After obtaining the structured format of the text file, we proceed to create a matrix that consists of words and their respective estimated weightage based on tf-idf. This matrix will ultimately be passed as an argument to the 'wordcloud()' method to generate a word cloud. The vertical dimension of each word in this image represents the word's significance in the overall text.

By following the steps mentioned above, the desired output will be obtained.

3.4 *Bayesian Network*

A Bayesian networks (BNs) represents probabilistic relationships among variables through a directed acyclic graph (DAG) [37]. Every node corresponds to a variable, and the directed arcs express conditional dependencies. As shown in Figure 2, a BN comprises a group of variables (such as *S, P1, P2*, and *Q*) connected by directed edges that represent their probabilistic dependencies. Each variable can exist in mutually exclusive states; for instance, variables *S, P1, P2,* and *Q* may take the states {Prst=present, Abst=absent}. Nodes without incoming links are referred as *root node* (such as *S* in Figure 2), while nodes receiving directed links are *child nodes*, those sending out links are *parent nodes*. In this configuration, *P1* and *P2* act as children of S and simultaneously as parents of *Q*. The



Variable S	Probability
Prst	P(S=Prst)
Abst	P(S=Abst)

Unconditional probability

Variable S	Variable P1	
	Probability	
	Prst	Abst
Prst	P(P1=Prst| S=Prst)	P(P1=Abst| S=Prst)
Abst	P(P1=Prst| S=Abst)	P(P1=Abst| S=Abst)

Conditional probability table (CPT)

Variable S	Variable P2	
	Probability	
	Prst	Abst
Prst	P(P2=Prst| S=Prst)	P(P2=Abst| S=Prst)
Abst	P(P2=Prst| S=Abst)	P(P2=Abst| S=Abst)

Conditional probability table (CPT)

Variable P1	Variable P2	Variable Q	
		Probability	
		Prst	Abst
Prst	Prst	P(Q=Prst| P1=Prst, P2=Prst)	P(Q=Abst| P1=Prst, P2=Prst)
Prst	Abst	P(Q=Prst| P1=Prst, P2=Abst)	P(Q=Abst| P1=Prst, P2=Abst)
Abst	Prst	P(Q=Prst| P1=Abst, P2=Prst)	P(Q=Abst| P1=Abst, P2=Prst)
Abst	Abst	P(Q=Prst| P1=Abst, P2=Abst)	P(Q=Abst| P1=Abst, P2=Abst)

Figure 2

Sample representation of Bayesian Network model

directed edges indicate dependency relationships, where value of each node is determined by the values of its parent nodes in the DAG. Consequently, the values of *P1* and *P2* depend on *S*, and the occurrence of *Q* is dependent on both *P1* and *P2*. These interdependencies are quantified using conditional probability tables (CPTs) associated with each node.

In the Bayesian framework, the dependency between multiple parent nodes *Pi (i=1,2,…,n)* and the corresponding evidence or child node *Q* can be calculated using the following expression:

$$p(P_i \mid Q) = \frac{p(Q \mid P_i) \times p(P_i)}{\sum_{j=1}^{n} p(Q \mid P_j) \times p(P_j)}$$

In this expression, $p(P \mid Q)$ denotes the posterior probability of P given that Q has occurred, $p(P)$ represents the prior probability of P, and $p(Q \mid P)$ indicate the conditional probability or likelihood of observing Q when P occurs. The conditional probabilities can be derived from either expert knowledge or from the available data.

4. Analysis of results and discussion of findings

4.1 *Extraction of root causes using word cloud visualization*

Following data pre-processing, both comparison and commonality clouds are created. The comparison cloud facilitates the identification of differences and similarity among two or more primary causes by contrasting their respective word clouds. In contrast, the commonality cloud highlights the terms that appear across multiple documents. The resulting comparison clouds are shown in Figures 3-6. The relative importance of each word in the visualization corresponds to its tf-idf weight, with larger font sizes representing greater significance. The root cause terms are chosen using word clouds that display comparisons and commonalities. Keywords are determined by assessing their importance and relying on the expertise of professionals. Figure 3 shows a cloud diagram comparing incidents connected to processes. These incidents are categorized based on their principal causes, which include hot metal, fire, poisonous chemicals, gas leakage, and process incidents. Figure 4 displays a cloud diagram that compares events linked to collisions and falls. The incidents are classified according to their primary causes, namely dashing or collision, skidding, slip/trip/fall, and working at height. Figure 5 illustrates a cloud diagram that compares occurrences associated with "material & equipment" based on their primary causes such as material

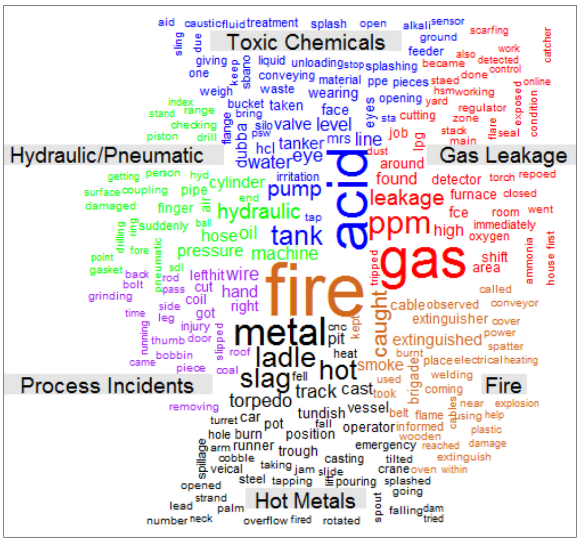


Figure 3

Comparison word cloud for process-related incidents

handling, equipment or machinery damage, structural integrity, and issues with lifting tools & tackles. Figure 6 displays a comparative cloud of energy-related incidents, categorized by their primary causes: electrical flash and energy isolation. Appendix A provides a list of most significant terms extracted from the word clouds corresponding to each primary cause.



Figure 4

Comparison word cloud for collision and fall-related incidents

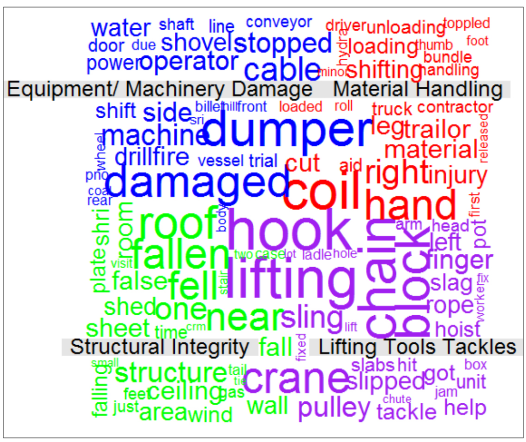


Figure 5

Comparison word cloud for material handling & equipment damage related incidents

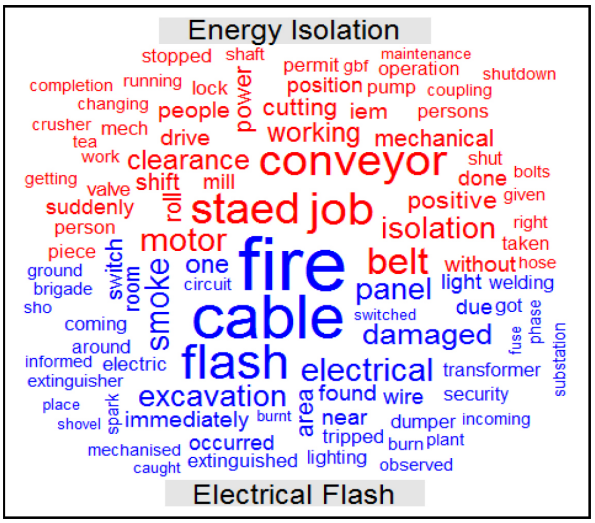


Figure 6

Comparison word cloud for energy-related incidents

4.2 *Causal modeling using Bayesian Network*

As illustrated in Figure 7, the proposed model is formulated based on the root causes extracted through word cloud analysis. It comprises four hierarchical levels, i.e., root causes, primary causes, cause group and near-miss index, for a profound understanding of the near-miss risk factors. The commonality between the root causes is found in the groups using the word cloud. The major root causes found from the word cloud are then picked out for every listed primary cause, which is given in Appendix A. Appendix A further assist in identifying the qualitative relationships among various factors and defining the structural dependencies between the variables.

Within the Bayesian network (BN) framework, each conditional probability table (CPT) is defined by node's immediate parent node, following the principle of conditional independence. In this study, every node is assigned two possible states: *present* and *absent*. The probabilities of independent nodes are derived from their frequencies in incident reports corresponding to specific primary cause. Model construction and CPT estimation are performed using OpenMarkov software [38]. The probability values are generated from empirical data, while the grouping of primary causes is determined through consultation of domain experts. Table 1 summarizes these four groups after the discussion with safety

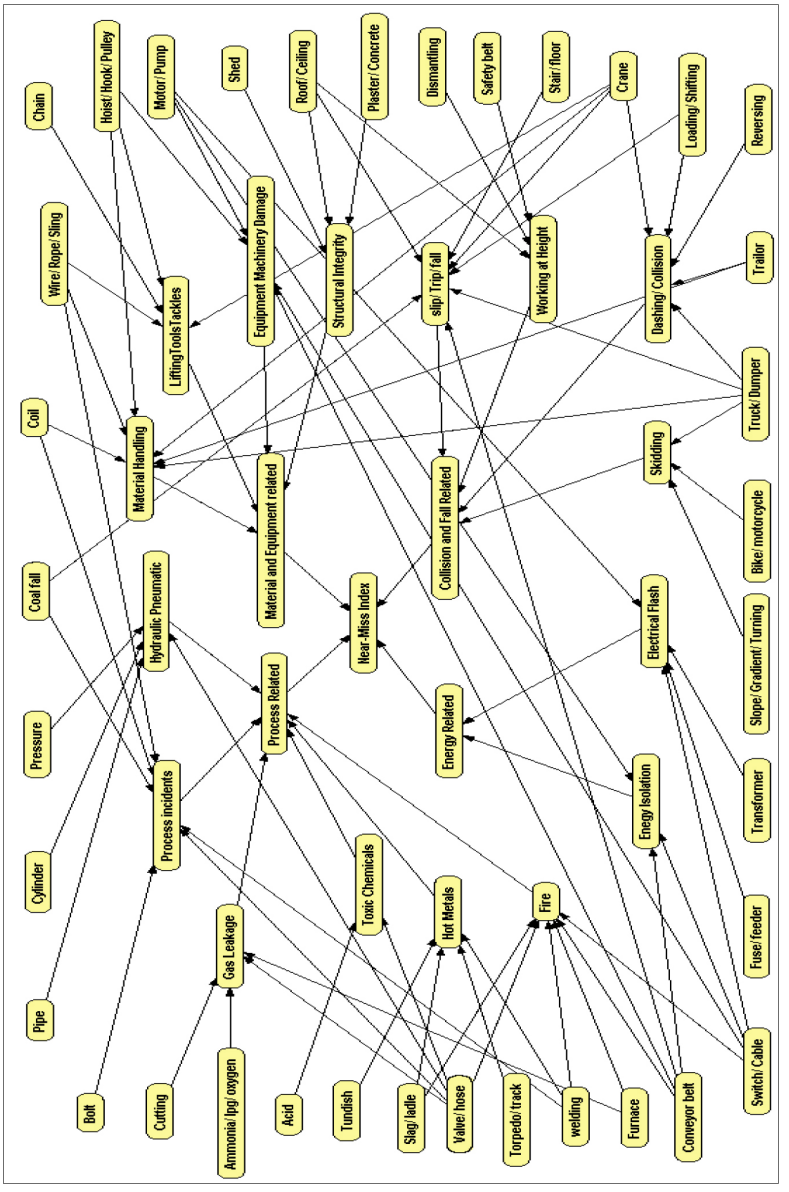


Figure 7

Bayesian network for evaluating near-miss incidents

experts of the steel plant by combining primary causes. The BN are created across three levels of causal factors: group, primary, root levels, with each node represented by distinct, quantifiable state. Finally, the conditional probabilities of each child nodes are computed by evaluating all possible combinations of their parent node states (present or absent) based on the given probabilities. A Bayesian network is constructed with 57 nodes, consisting of 36 independent and 21 dependent nodes. The network also had 88 linkages and required the estimation of 1200 conditional probabilities. Figure 8 displays a screenshot of the OpenMarkov software application, which depicts the near-miss index by considering all relevant criteria. The respective nodes display the marginal probability. The primary causes were grouped based on commonality and to reduce the need for complex mathematical calculations, as shown in Table 1. Given the high level of connectedness and the low likelihood of a root cause node, paying close attention to even the most minor changes in the near-miss index produced by a specific node is crucial.

Table 1

Groups and clubbed primary causes

Sr. No.	Group name	Primary Causes
1	Collision and Fall related	Slip/Trip/Fall Dashing collision Skidding Lifting tools tackles Working at height
2	Process related	Fire Hot metals Gas leakage Hydraulic/Pneumatic Toxic chemicals Process incidents
3	Energy related	Electrical flash Energy isolation
4	Material and Equipment related	Material handling Equipment/ machinery damage Structural handling

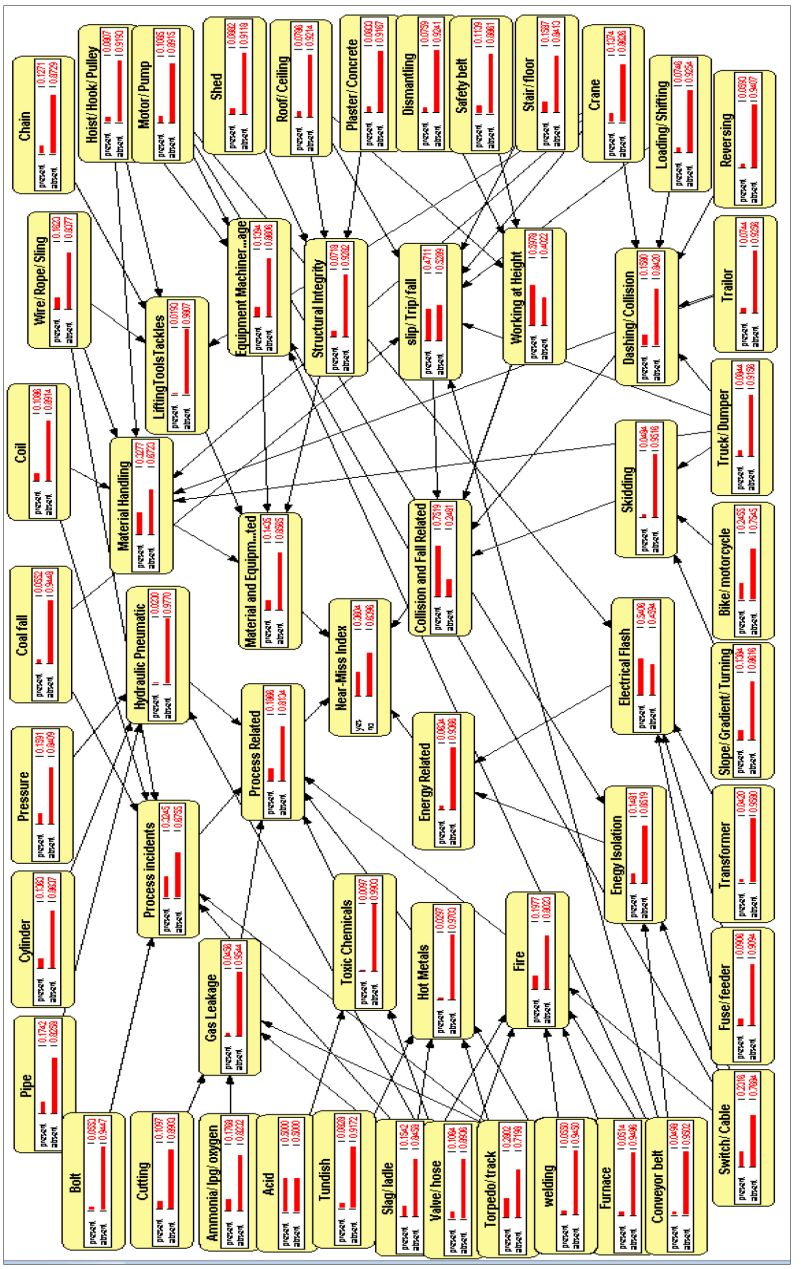


Figure 8

Probabilistic model of Bayesian network

4.2.1 Scenario analysis

The relative influence of causal factors across the three tiers of the network was evaluated by analyzing how the near-miss index changes when the probability of each factor's occurrence is assumed to be zero. This analysis provides insight into the degree to which each node influences the overall near-miss index. Table 2 presents the fundamental underlying reasons derived from the incident's description (narrative text) by applying text mining. The prioritized order of a particular root cause level and their unique effects on the near-miss index are provided in Tables 2.

Similarly, the status of BN is analysed after introducing the hypothetical cases reflecting the absence of certain nodes. These simulated scenarios illustrate how the network responds when particular evidence nodes are observed or removed. Since one root cause is associated with more than one primary cause, it is interesting to see the effect of that evidence node propagated to every node of the causal path to understand the importance of mitigation or control of the particular hazard.

Scenario 1, shown in Figure 9, for the absence of switch/cable-related incidents (23.16%), the BN suggests that there is a reduction possible in primary causes like fire (18.66%), energy isolation (13.45%), electric flash (49.34%) and equipment/ machinery damage (13.48) primary cause incidents. At the group level, a reduction is possible for energy (5.79%), process (18.38%), and material/ equipment damage (14.30%) related incidents. The near-miss index comes down to a probability of 36.02%.

Table 2

Prioritized ranking of root causes identified from word cloud representation

Order	Root cause extracted from word cloud	Estimated Probability	Probability of near-miss index in the absence of a root cause
1	Switch/ Cable	23.16	36.02
2	Stairs/ Floor condition	15.87	36.02
3	Wire/ rope/ sling	16.23	36.03
4	Crane	13.74	36.03
5	Valve/ hose	10.64	36.03
6	Truck/ Dumper	8.44	36.03
7	Loading/ Shifting	7.46	36.03

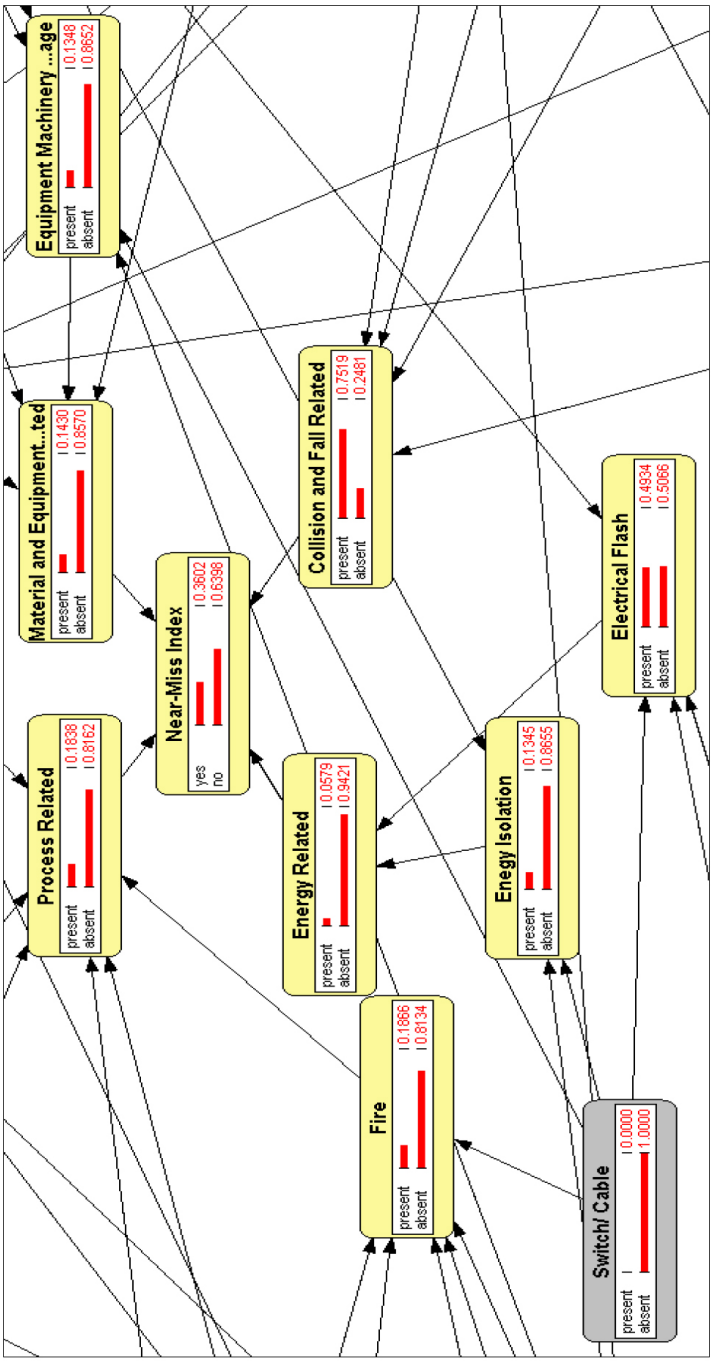


Figure 9

Status of affected node after propagation of additional evidence of absence of switch/ cable related incidence.

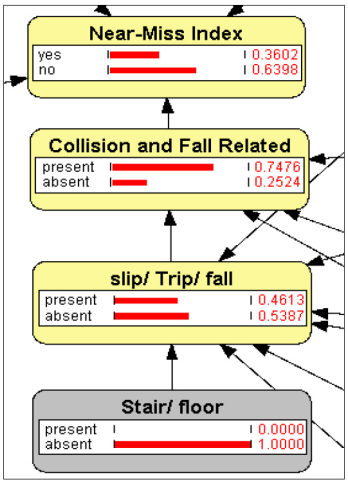


Figure 10

Status of affected node after propagation of additional evidence of absence of stair or floor related incidence.

Scenario 2, shown in Figure 10, depicts a situation where all incidents related to stair and floor conditions (15.87%) are absent. The BN suggests that there is a reduction possible in the primary cause level of slip/ trip/ fall (46.13%) related incidents. At the group level, a reduction is possible for collision and fall-related incidents (74.70%). The near-miss index comes down to a probability of 36.02%.

Scenario 3, shown in Figure 11, represents a situation when all incidents related to wire, rope and sling conditions (16.23%) are absent. The BN suggests that there is a reduction possible in primary causes like material handling (31.06%), process incidents (31.90%) and lifting tools (1.44%) related incidents. At the group level, reduction is possible for process-related (18.47%) and material & equipment-related (13.68%) incidents. The near-miss index comes down to a probability of 36.03%.

Scenario 4, shown in Figure 12, represents a situation where all incidents related to crane operation (13.74%) are absent. The BN suggests that there is a reduction possible in primary causes like material handling (30.89%), slip/ trip/ fall (46.77%), dashing/ collision (15.78%) and lifting tools tackles related (1.37%) incidents. At the group level, the reduction is possible for collision & fall (74.93%) and material & equipment-related (13.61%) incidents. From that, the near-miss index comes down to a probability of 36.03%.

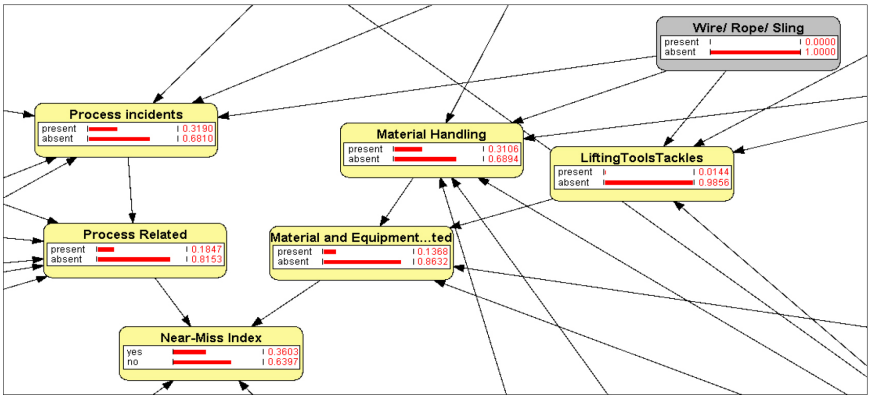


Figure 11

Status of affected node after propagation of additional evidence of absence of wire/ rope/ sling related incidence.

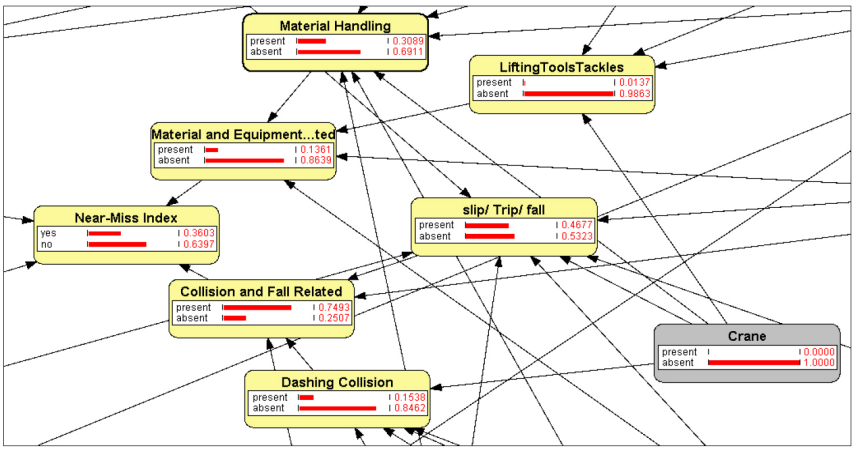


Figure 12

Status of affected node after propagation of additional evidence of absence of crane related incidence.

Scenario 5, shown in Figure 13, represents a situation where all incidents related to valve or hose dysfunction (10.64%) are absent. The BN suggests that there is a reduction possible in primary causes like gas leakage (4.38%), hydraulic/ pneumatic (2.03%), process incident (32.28%), fire incidents (19.49%) and toxic chemical (0.92%) related incidents. At the group level, valve and hose incidents only impact process-related incidents (18.49%) to reduce it. The near-miss index comes down to a probability of 36.03%.

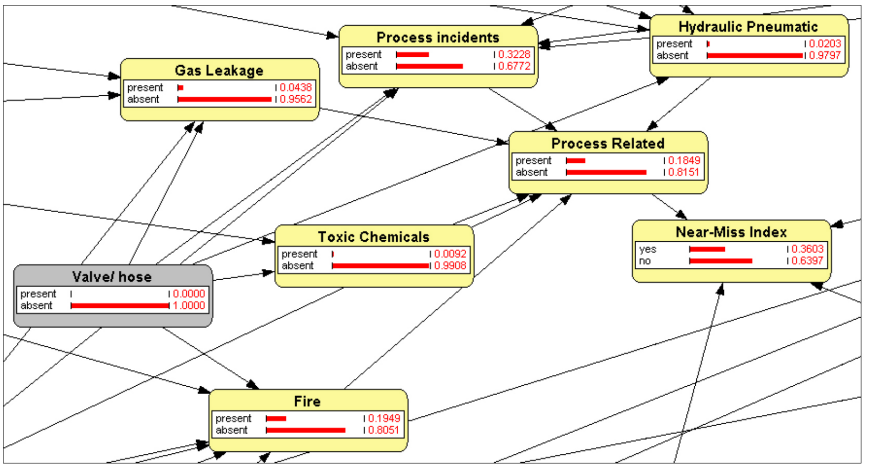


Figure 13

Status of affected node after propagation of additional evidence of absence of valve or hose related incidence

Scenario 6, shown in Figure 14, represents a situation when all incidents related to heavy vehicles, like a truck or dumper-related incidents (8.44%), are absent. The BN suggests that there is a reduction possible in primary causes like skidding (4.79%), slip/ trip/ fall (47.00%), dashing/

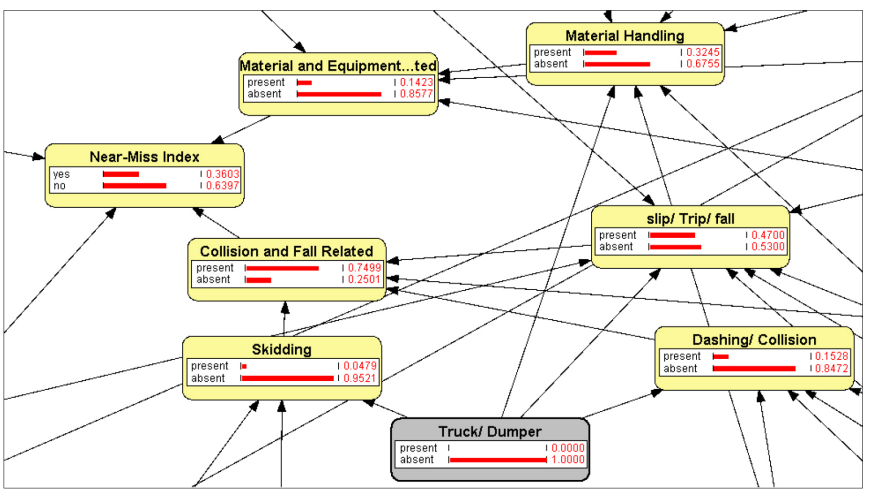


Figure 14

Status of affected node after propagation of additional evidence of absence of truck or dumper related incidence.

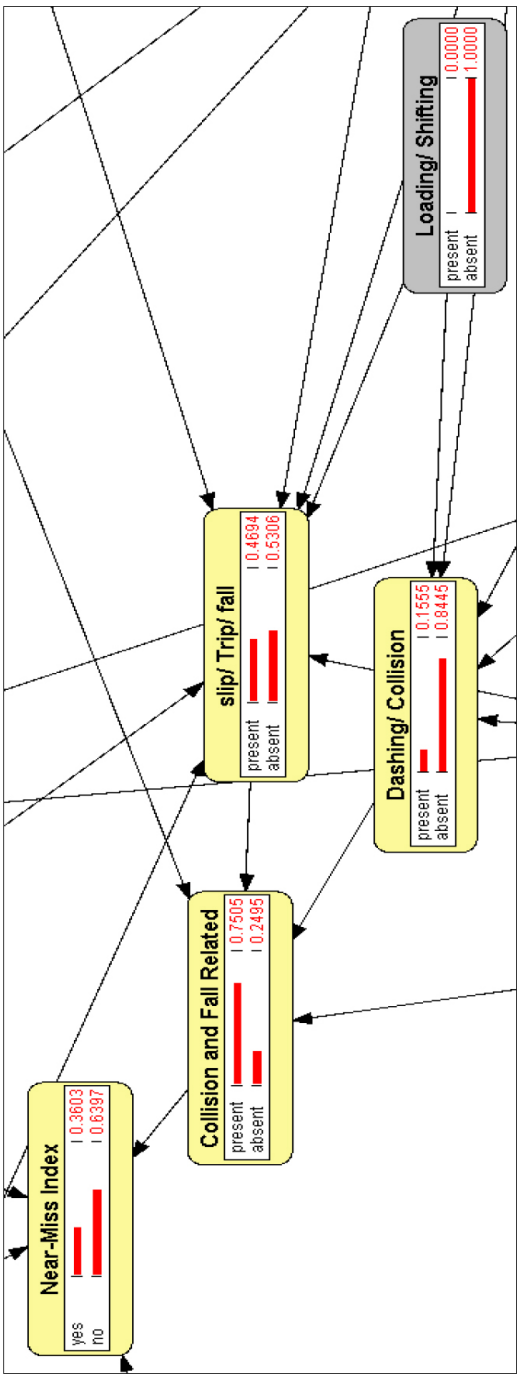


Figure 15

Status of affected node after propagation of additional evidence of absence of loading or shifting operation related incidence.

collision (15.28%) and material handling (32.45%) incidents. At the group level, reduction is possible for collision & fall (74.99%) and material & equipment (14.23%) related incidents. The near-miss index comes down to a probability of 36.03%.

Scenario 7, shown in Figure 15, represents a situation where all incidents related to loading and shifting operations (7.46%) are absent. The BN suggests a reduction possible in primary causes like slip/ trip/ fall (46.94%) and dashing/ collision (15.55%) related incidents. At the group level, reduction is possible for collision & fall (75.05%) related incidents. The near-miss index comes down to a probability of 36.03%.

5. Conclusions

A comprehensive analysis is conducted to determine the underlying reasons for near-miss accidents. An ordered list of root causes can significantly assist in formulating a safety policy for the steel plant and allocating the budget for this purpose. The proposed technique utilizes a word cloud to extract information at the most basic level and then integrates it with a Bayesian network to assist in prioritizing near-miss incidents. The causal relationships among these factors were determined using data from a preliminary investigation report and expert knowledge. The model can provide qualitative and quantitative information on many parameters related to near-miss incidents. At the fundamental level, 36 underlying causes were recognised, and their impact on the overall network is evaluated by simulating hypothetical scenarios. The results provide valuable insights for management to concentrate on improving specific operational areas. The factors with high probability of contributing to incidents faulty switch or cable, unsafe stair or floor conditions, damaged wires, ropes, or slings, improper crane operation, valve or hose malfunctions, and large vehicles such as trucks or dumpers involved in material loading or transportation. Additional underlying factors can also be ranked according to their likelihood and influence on intermediary elements. Examining the data using a Bayesian network yields four significant rules for the steel plant under investigation.

Rule 1: Near-miss reduction requires tracing all three hierarchy levels, namely group, primary cause, and root cause.

Rule 2: As the level of group participation rises, the likelihood of near-miss occurrences also increases, which could escalate into accidents.

Rule 3: The higher the probability of the primary cause within a group, the greater the likelihood of near-miss incidents.

Rule 4: At the root causes level, both the individual probability and their involvement in the near-miss index are essential.

The proposed methodology is efficient because it utilizes data to build causal networks, reducing the challenge of finding domain experts and mitigating inconsistencies in their opinions. This approach enables the management to effectively handle unstructured data and discover the root causes that are most vulnerable, sensitive, and highly prone to causing incidents within the plant. Furthermore, based on the study, it is recommended that measures be implemented that will lead to continuous enhancement in the intranet-based reporting, manual reporting, and monthly survey system across various departments. This will enable a more accurate description of the events around each occurrence and facilitate the identification of hidden near-misses.

Limitations

The effectiveness of the proposed approach relies heavily on the availability of data-rich environments; however, such conditions are often difficult to achieve in many real-world applications. Moreover, the quality of recorded data is largely depend on the comprehension and interpretation of supervisors documenting the incidents in the database [39]. Cleaning data is itself a big issue. Under-reporting of near-miss incidents limits the exploration of more causing factors. Most of the near-miss cases are overlooked or go unreported, which checks the opportunity for workplace safety improvement. Williamsen [40] provided the following issues found at different levels of hierarchy in the organization regarding near-miss incident reporting:

1) Managers and supervisors consider reporting and investigating incidents and informing senior management about the actions taken to be burdensome tasks that do not provide value.
2) Safety professionals face challenges in motivating and instructing staff to provide accurate and dependable quantitative data.
3) Only significant occurrences are reported by daily employees and contract workers to their supervisors. They consider minor near-miss events to be a part of the job.

4) Data analyst find it difficult to analyze the scarce and unreliable data to generate meaningful outcomes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work is carried out under the UAY project, Government of India (GOI), jointly funded by Ministry of Education & Ministry of Steel (GOI), and Tata Steel Limited. The authors gratefully acknowledge the Centre of Excellence in Safety Engineering and Analytics (CoE-SEA) (www.coesea.iitkgp.ac.in) and Safety Analytics & Virtual Reality (SAVR) Laboratory (www.savr.iitkgp.ac.in) of Department of Industrial & Systems Engineering, IIT Kharagpur, for providing experimental, computational and research facilities used in this study. The authors also extend sincere thanks to plant management for their support, cooperation, and providing the necessary data for this study.

Appendix A

Sl. No.	Group name	Primary Causes	Root cause words
1.	Process related	Fire	1. Ladle/ slag
			2. Valve/ hose
			3. Welding
			4. Furnace
			5. Switch/ cable
			6. Conveyor belt
		Hot metals	1. Ladle/ slag
			2. Tundish
			3. Torpedo/ track
			4. Welding
		Gas leakage	1. Valve/ hose
			2. Furnace

Contd...

			3. Ammonia/ lpg/ oxygen
			4. Cutting
		Hydraulic/ Pneumatic	1. Pressure
			2. Pipe
			3. Cylinder
			4. Valve/ hose
			5. Welding
		Toxic chemical	1. Valve/ hose
			2. Acid
		Process incident	1. Valve/ hose
			2. Coal fall
			3. Bolt
			4. Coil
			5. Furnace
			6. Welding
2.	Collision and Fall related	Dashing/ collision	1. Truck/ dumper
			2. Loading/ shifting
			3. Crane
			4. Reversing
			5. Trailor
		Skidding	1. Truck
			2. Bike/ motor cycle
			3. Slope/ gradient/ turning
		Slip/ trip/ fall	1. Conveyor belt
			2. Coal
			3. Truck/ dumper
			4. Loading/ shifting
			5. Crane
			6. Stair/ floor
			7. Roof/ ceiling
		Working at height	1. Safety belt
			2. Roof/ ceiling
			3. Dismantling

Contd...

3.	Material and Equipment related	Material handling	1. Coil
			2. Truck/ dumper
			3. Crane
			4. Wire/ rope/ sling
			5. Hoist/ hook/ pulley
			6. Trailor
		Equipment damage	1. Switch/ cable
			2. Conveyor belt
			3. Hoist/ hook/ pulley
			4. Motor/ pump
		Lifting tools tackles	1. Crane
			2. Wire/ rope/ sling
			3. Hoist/ hook/ pulley
			4. Chain
		Structural Integrity	1. Roof/ ceiling
			2. Plaster/ concrete
			3. Shed
4.	Energy related	Energy isolation	1. Switch/ cable
			2. Conveyor belt
			3. Motor/ pump
		Electrical flash	1. Switch/ cable
			2. Motor/ pump
			3. Fuse/ feeder
			4. Transformer

Reference

[1] A. Verma, S. Das Khan, J. Maiti, and O. B. Krishna, "Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports," *Saf. Sci.*, vol. 70, pp. 89–98 (Dec. 2014).

[2] A. Verma and J. Maiti, "Text-document clustering-based cause and effect analysis methodology for steel plant incident data," *Int. J. Inj. Contr. Saf. Promot.*, vol. 25, no. 4, pp. 416–426 (2018), doi: 10.1080/17457300.2018.1456468.

[3] K. Singh, J. Maiti, and K. Dhalmahapatra, "Chain of events model for safety management: Data analytics approach," *Saf. Sci.*, vol. 118, pp. 568–582 (2019).

[4] A. Bagga, S. Srivastava, and R. S. Shekhawat, "Prediction of road crash attributes using machine learning," *J. Inf. Optim. Sci.*, vol. 45, no. 4, pp. 1217–1227 (2024).

[5] T. W. Van Der Schaaf, "Near miss reporting in the chemical process industry," Eindhoven University of Technology (1992).

[6] J. R. Phimister, U. Oktem, P. R. Kleindorfer, and H. Kunreuther, "Near-miss incident management in the chemical process industry.," *Risk Anal.*, vol. 23, no. 3, pp. 445–59 (Jun. 2003).

[7] U. G. Oktem, "Near-Miss : A Tool for Integrated Safety , Health , Environmental and Security Management," (2003).

[8] F. E. Bird and G. L. Germain, *Loss control management: Practical Loss Control Leadership*, Revised Ed. International Loss Control Institute Loganville (1986).

[9] H. . Heinrich, *Industrial Accident Prevention*, Second. New York: McGraw-Hill (1959).

[10] A. Verma, D. Rajput, and J. Maiti, "Prioritization of Near-Miss Incidents Using Text Mining and Bayesian Network," *Int. Conf. Adv. Comput. Data Sci.*, vol. 721, pp. 183–191 (2017).

[11] U. G. Oktem, R. Wong, and C. Oktem, "Near-Miss Management : Managing the Bottom of the Risk Pyramid," *Risk & Regulation-Magazine of the ESRC Centre for Analysis of Risk and Regulation*, no. July, pp. 12–13 (2010).

[12] M. G. Gnoni, S. Andriulo, G. Maggio, and P. Nardone, "' Lean occupational' safety: An application for a Near-miss Management System design," *Saf. Sci.*, vol. 53, pp. 96–104 (2013).

[13] B. Basso, C. Carpegna, C. Dibitonto, G. Gaido, A. Robotto, and C. Zonato, "Reviewing the safety management system by incident investigation and performance indicators," *J. Loss Prev. Process Ind.*, vol. 17, no. 3, pp. 225–231 (May 2004).

[14] Z. Nivolianitou, M. Konstandinidou, C. Kiranoudis, and N. Markatos, "Development of a database for accidents and incidents in the Greek petrochemical industry," *J. Loss Prev. Process Ind.*, vol. 19, no. 6, pp. 630–638 (Nov. 2006).

[15] T. van der Schaaf and L. Kanse, "Biases in incident reporting databases: an empirical study in the chemical process industry," *Saf. Sci.*, vol. 42, no. 1, pp. 57–67 (Jan. 2004).

[16] W. Wu, H. Yang, D. a. S. Chew, S. Yang, A. G. F. Gibb, and Q. Li, "Towards an autonomous real-time tracking system of near-miss accidents on construction sites," *Autom. Constr.*, vol. 19, no. 2, pp. 134–141 (Mar. 2010).

[17] W. Wu, A. G. F. Gibb, and Q. Li, "Accident precursors and near misses on construction sites: An investigative tool to derive information from accident databases," *Saf. Sci.*, vol. 48, no. 7, pp. 845–858 (Aug. 2010).

[18] F. B. Cambraia, T. A. Saurin, and C. T. Formoso, "Identification, analysis and dissemination of information on near misses: A case study in the construction industry," *Saf. Sci.*, vol. 48, no. 1, pp. 91–99 (Jan. 2010).

[19] M. S. Brian Thoroman and P. Salmon, "An integrated approach to near miss analysis combining AcciMap and Network Analysis," *Saf. Sci.*, vol. 130, no. June, p. 104859 (2020).

[20] R. L. Dillon and C. H. Tinsley, "How Near-Misses Influence Decision Making Under Risk: A Missed Opportunity for Learning," *Manage. Sci.*, vol. 54, no. 8, pp. 1425–1440 (Aug. 2008).

[21] S. Jones, C. Kirchsteiger, and W. Bjerke, "The importance of near miss reporting to further improve safety performance," *J. Loss Prev. Process Ind.*, vol. 12, no. 1, pp. 59–67 (Jan. 1999), doi: 10.1016/S0950-4230(98)00038-2.

[22] P. Marsh and D. Kendrick, "Near miss and minor injury information--can it be used to plan and evaluate injury prevention programmes?," *Accid. Anal. Prev.*, vol. 32, no. 3, pp. 345–54 (May 2000).

[23] A. Verma, J. Maiti, and V. N. Gaikwad, "A preliminary analysis of incident investigation reports of an integrated steel plant: some reflection," *Int. J. Inj. Contr. Saf. Promot.*, vol. 25, no. 2, pp. 180–194 (2018).

[24] M. G. Gnoni and G. Lettera, "Near-miss management systems: A methodological comparison," *J. Loss Prev. Process Ind.*, vol. 25, no. 3, pp. 609–616 (May 2012).

[25] S. J. Dee, B. L. Cox, and R. A. Ogle, "Using Near Misses to Improve Risk Management Decisions," *Process Saf. Prog.*, vol. 32, no. 4 (2013).

[26] J. a. Taylor, A. V. Lacovara, G. S. Smith, R. Pandian, and M. Lehto, "Near-miss narratives from the fire service: A Bayesian analysis," *Accid. Anal. Prev.*, vol. 62, pp. 119–129 (2014), doi: 10.1016/j.aap.2013.09.012.

[27] E. Fersini, E. Messina, and F. a. Pozzi, "Sentiment analysis: Bayesian Ensemble Learning," *Decis. Support Syst.*, vol. 68, pp. 26–38 (2014).

[28] R. P. Chaturvedi and U. Ghose, "Small object detection using retinanet with hybrid anchor box hyper tuning using interface of Bayesian mathematics," *J. Inf. Optim. Sci.*, vol. 43, no. 8, pp. 2099–2110 (2022).

[29] M. Neil, N. Fenton, and M. Tailor, "Using Bayesian networks to model expected and unexpected operational losses," *Risk Anal.*, vol. 25, no. 4, pp. 963–972 (2005).

[30] V. M. Bier and A. Mosleh, "The analysis of accident precursors and near misses: Implications for risk assessment and risk management," *Reliab. Eng. Syst. Saf.*, vol. 27, no. 1, pp. 91–101 (1990).

[31] A. Meel and W. D. Seider, "Plant-specific dynamic failure assessment using Bayesian theory," *Chem. Eng. Sci.*, vol. 61, no. 21, pp. 7036–7056 (Nov. 2006), doi: 10.1016/j.ces.2006.07.007.

[32] E. Lee, Y. Park, and J. G. Shin, "Large engineering project risk management using a Bayesian belief network," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5880–5887 (Apr. 2009).

[33] G. Kabir, S. Tesfamariam, A. Francisque, and R. Sadiq, "Evaluating risk of water mains failure using a Bayesian belief network model," *Eur. J. Oper. Res.*, vol. 240, no. 1, pp. 220–234 (2015).

[34] S. Park, H. Yang, G. Heo, M. Zubair, and R. K. Ur, "Study on Nuclear Accident Precursors Using AHP and BBN," *Sci. Technol. Nucl. Install.*, vol. 2014, p. 12 (2014).

[35] M. Zubair, S. Park, G. Heo, M. U. Hassan, and M. Aamir, "Study on nuclear accident precursors using AHP and BBN, a case study of Fukushima accident," *Int. J. energy Res.*, vol. 39, no. 1, pp. 98–110 (2015).

[36] A. Verma, K. Dhalmahapatra, and J. Maiti, "Forecasting occupational safety performance and mining text-based association rules for incident occurrences," *Saf. Sci.*, vol. 159, no. August 2022, p. 106014 (2023).

[37] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (1988).

[38] M. Arias, F. Díez, and M. Palacios, "OpenMarkovXML. A format for encoding probabilistic graphical models," (2010).

[39] M. Tiwari, P. Aital, and P. Joshi, "A comprehensive survey and review of machine learning techniques in document processing : Industry applications and future directions," *J. Inf. Optim. Sci.*, vol. 45, no. 4, pp. 1177–1188 (2024).

[40] M. Williamsen, "Near-Miss Reporting: A missing link in safety culture," *Prof. Saf.*, vol. 58, no. May, pp. 46–50 (2013).

Received December, 2024