Reliability Engineering and System Safety 202 (2020) 107041

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

A novel data mining approach for analysis of accident paths and performance assessment of risk control systems

Kritika Singh, J Maiti*

MS Student and Professor, Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India

ARTICLE INFO

Keywords:
Safety analytics & data mining
Frequent itemset generation
Temporal effect
Risk control system
Safety management

ABSTRACT

The data mining researches to facilitate the process of safety management is fairly new, compared to other industrial management domains. The implementation of appropriate, effective, and safe risk control systems (RCSs) is vital to ensure zero-accident and zero-harm vision of industrial work-systems. In this work, we propose a data mining based tool to analyze accident paths from incident data and assess the performance of RCSs. Our work upgrades the existing pattern analysis methods through three new types of analyses (i) temporal frequent itemset generation (T-FIG) for studying the time effect on patterns, (ii) elevated severity itemset generation (ESIG) for examining the risk reduction due to RCSs, and (iii) High impact itemset generation (High_impact_IG) to identify accident paths with high risk. T-FIG and ESIG assist in performance assessment of preventive and mitigating RCSs, respectively. The results from each of the analyses are compared and eight types of inferences regarding the performance of RCSs are drawn. The proposed methodology is applied to 612 incident records reported during steel making process in a steel manufacturing plant. It was found that there are four accident paths which have ineffective preventive and mitigating RCSs, have high risk and are probable to recur in future. Two among four of these paths include hot metal/steel/slag as the hazardous element and three of them are due to damaged/degraded/poorly maintained equipment. Moreover, the case study also demonstrates that proposed data mining approach is an effective and easy to use tool for performance assessment of RCSs and accident path analysis.

1. Introduction

1.1. Background

A risk control system (RCS) refers to one or more safety measures, which alone or in combination, prevents an accident to occur and/or mitigates the severity after the occurrence. The presence, effectiveness, and adequacy of RCS affect the overall safety performance of any system. Hence, the performance assessment of RCS holds high significance in safety management. The RCS are usually referred to as safety barriers in the existing literature. The performance assessment criteria and methods for RCSs are largely driven by the category to which the RCS belong. For instance, the assessment methods for administrative RCSs will be different from the methods effective for technical RCSs. Therefore, a concise, holistic and clear classification of plausible RCSs serves as a basic building block for designing performance assessment methodology. Several studies have provided different type of categorization [1–4]. NIOSH has provided hierarchy of control which has five hierarchal levels: elimination, substitution, engineering control, administrative control, and personal protective equipment. The effectiveness of these controls decrease from top to bottom of the hierarchy. On the other hand, HSE has classified the RCSs into nine categories for process industry: (i) inspection and maintenance, (ii) instrumentation and alarms, (iii) operating procedures, (iv) staff competence, (v) permit to work, (vi) plant design, (vii) plant change, (viii) communication, and (ix) emergency preparedness.

The RCS is closely related to accident path (AP). Therefore, assessment of RCSs can be done through analysis of accident paths. However, identification of complete accident path and appropriate RCSs for them may require huge human labor. Moreover, the list may not be exhaustive and monitoring the performance of each RCSs may also be difficult.

This issue can be addressed by analysis huge accident data collected by the organization using data mining approach. Frequent item-set generation, a data mining approach, can be used to identify which accident paths are frequently recurring.

In the work presented in this article, a new data mining based tool is presented to address these issues, analyze the accident paths and assess

* Corresponding author.
E-mail addresses: kritika.swati@gmail.com (K. Singh), jhareswar.maiti@gmail.com (J. Maiti).

https://doi.org/10.1016/j.ress.2020.107041
Received 15 June 2019; Received in revised form 18 February 2020; Accepted 22 May 2020
Available online 27 May 2020
0951-8320/

the performance of RCSs. The methodology performs four types of analysis on accident paths for the assessment. The methodology contributes to the studies of accident path analysis as well as performance assessment of RCSs. There have been scarce study of analysis of complete accident paths using data mining approach. There have been no study to assess the performance of RCSs using accident records and data mining methods.

The methodology is validated through accident records collected by a steel plant in India. Steel manufacturing industry has highly complex socio-technical work-system, hence more prone to hazardous situations and accidents. The plants comprise huge machineries and activities are also labor intensive, hence, the man-machine interaction in unavoidable and the severity of any accident is very high. Though steel industry is not a process industry, however, safety management systems focus on both workplace and process safety. Safety management system (SMS) primarily focusses on reducing the accident frequency, controlling the risk and continuously improving these conditions in the organization. World Steel Association has provided six safety and health principles for steel industry [5]. The safety management in steel manufacturing organizations are driven by safety objectives: (i) leadership and accountability, (ii) hazard identification, risk assessment and management, (iii) compliance assurance, (iv) design, construction and operational control, (v) people, competency and behaviors, (vi) communication, consultation and empowerment, (vii) incident reporting, investigation and learning, (viii) asset management, (ix) management of change, (x) emergency preparedness, response and crisis management, (xi) document control and record management, and (xii) measuring performance, audit and review. The presented work assists in achieving objective 7 and 12 through data-driven approach.

1.2. Concept of accident path

An accident path (AP) comprises five components, namely hazardous element (HE), accident causation mechanism (ACM), accident (Acc), target, and threat (see Figure 1). ACM comprises two components, initiating event (IE) and pivotal event (PV). An IE is an event which triggers the HE and may be followed by a PV [6]. An active state IE may cause accident while the passive IE is propagated to accident through PVs. The accident possess some threat which causes harm to the target. The escalation of ACMs is prevented by preventive risk control systems (P-RCSs). The P-RCSs are physical or non-physical barriers in the workplace to prevent the occurrence of an incident. The number of incidents is inversely proportional to the performance of P-RCS. Further, the severity associated with a threat depends on the performance of mitigating risk control systems (M-RCS). The M-RCSs are the physical or non-physical barriers at the workplace that reduce the severity of an incident. The severity of the incident is inversely proportional to the performance of M-RCS.

RCSs are inherent part of the conceptual chain of events model. If the experts analyze how the events propagated to cause an accident, they can identify how, when and which RCS failed to cause the accident or increased the severity. Therefore, analysis of complete accident path provides direct insights about how the RCSs have performed. Due to this reason accident path serves as base for all the analysis in this work.

1.3. Related work

1.3.1. Performance assessment of RCSs

The studies on the performance assessment of RCS are very few and can be broadly classified into two categories. Firstly, performance assessment of P-RCS for prevention of accident. Groot [7] has used BowTie methodology, and Tayab et al. [8] has analyzed accident paths to identify barriers and provide a framework for their management. Other studies, which have incorporated an assessment of barrier performance as a part of analyzing and preventing domino effects, are [9,10]. Landucci and coworkers have proposed a method to quantify barrier performance considering the effectiveness and availability of RCS. Their quantitative assessment is based on the expected performance of RCSs. Bucelli et al. [11] aimed at developing a structured methodology to the quantitative performance assessment of safety barriers in a harsh environment. Janssens et al. [12] proposed an approach and a computer program to determine the most optimal safety barrier investment decision for dealing with domino effects in existing industrial settings. Their approach helps in resource allocation for prevention of domino effects.

Second, the performance assessment of M-RCS to reduce the risk by mitigating domino events through implementation of effective RCSs [13–15]. The main focus of these studies is to propose a quantitative risk model by analyzing the effect of performance of M-RCS. Different techniques have been used to study this effect such as event tree analysis [14], barrier and operational risk analysis (BORA) [15], etc.

Two limitations can be observed in the above mentioned studies. First, they studied the effect of performance of RCS on the propagation of events. These performance assessments are computed before any actual incident has occurred, i.e., the real time performance may differ from the anticipated performance. Hence, it must be noted that the actual performance of the RCS can be assessed only after an incident has occurred. Analyzing accident patterns will help in identifying weak areas in RCSs' implementation. Second, studies are performed on specific workplace units and accident scenarios associated with it. Implementation of these researches in large plants may be cumbersome and require huge human labor. Moreover, it may also be prone to subjective biasness in assessment. The existing studies didn't use incident data and data mining techniques for performance assessment. The accident path patterns from incident reports can be analyzed to assess the performance of RCSs. Considering both the limitations, we propose a data mining approach for performance assessment of RCSs through accident path analysis.

1.3.2. Accident path analysis using data mining

The studies on accident path analysis using data mining have considered different individual components of accident path. A complete accident path has not been developed. The data mining applications in accident path analysis can be broadly studied in three domains: hazard

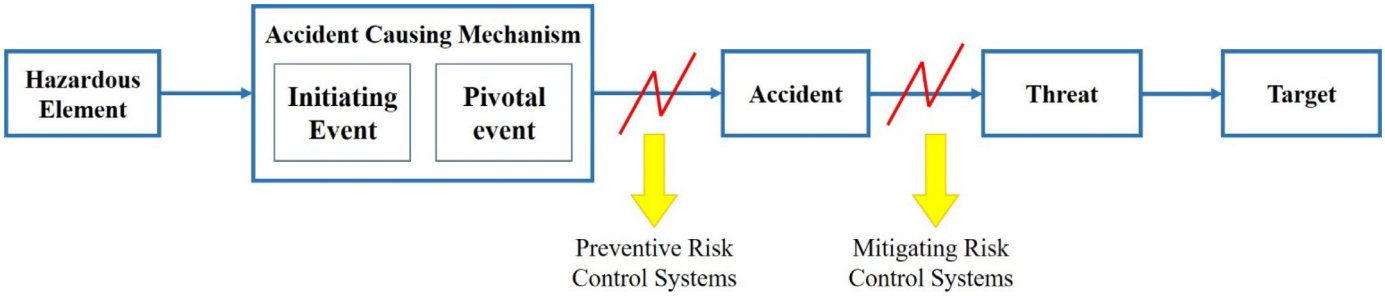

Figure 1. Conceptual model of chain of events in accident path.

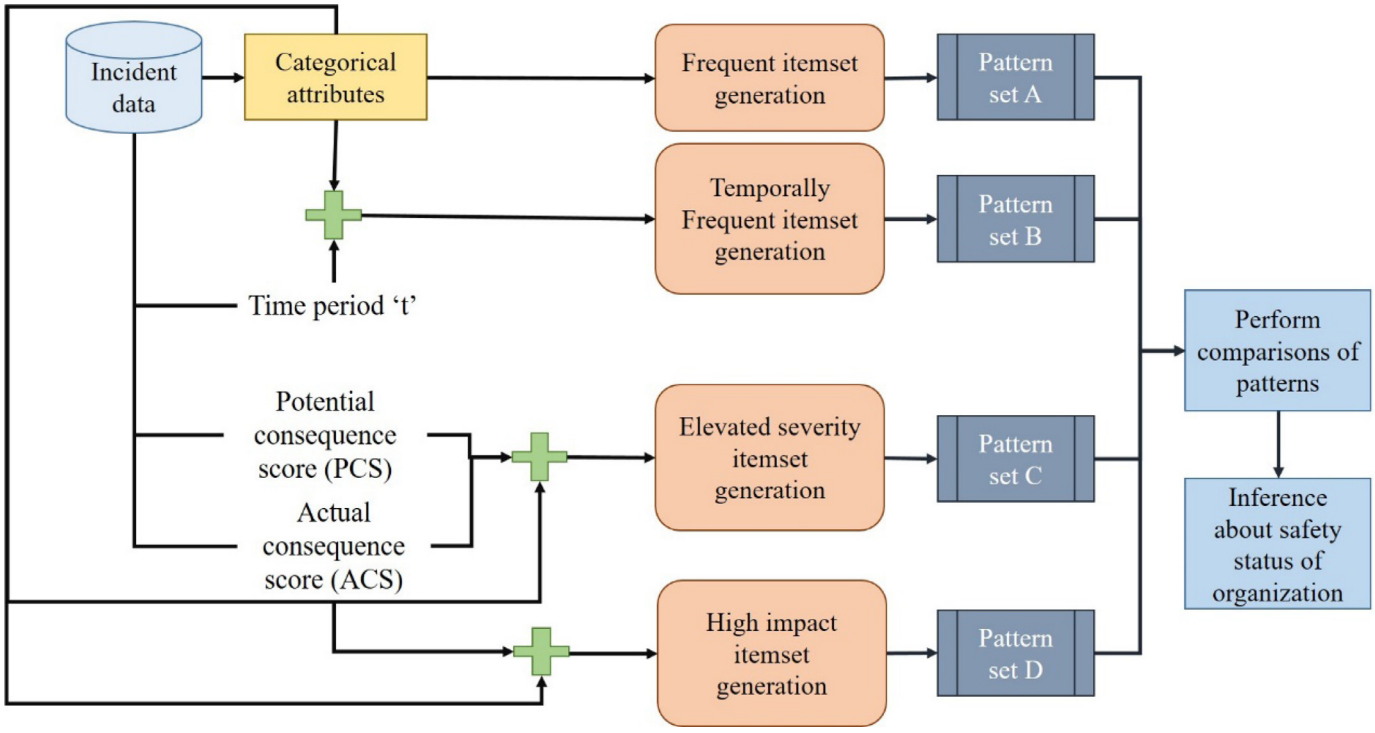

Figure 2. Framework for assessment using incident records.

identification [16], causal pattern analysis [17,18] and severity analysis [19]. The hazard identification literature is fairly less, however, the cause and effect relationship has been analyzed by many researchers using association rule mining (ARM) or frequent itemset generation (FIG) [20,21] and decision tree [22–24]. There have been few studies on application of data mining in safety domain for steel industry. Verma et al. [25] used association rule mining for identifying safety patterns from incident reports. Dhalmahapatra et al. [26] proposed decision support framework for generating safety rules from EOT crane incidents in steel manufacturing plant. Singh et al. [6] and Verma et al. [27] used text clustering for causal factors and recurring accident paths identification from incident narratives.

The studies in accident pattern analysis using FIG or ARM have equally weighed the data over the time of occurrence. The behavior of a pattern may change after preventive actions are taken by management. If the preventive actions are effective, the frequency of pattern will decrease over time, and if it is ineffective, the frequency increases or remains unchanged in the future. Also, some managerial actions may result in a new type of accident scenarios, which are probable to recur in the future. The traditional FIG or ARM method is unable to capture these changes. To the best of the knowledge of authors, none of the studies have considered the temporal effect in the pattern analysis. Moreover, studies have considered frequency as the major parameter in identifying patterns. So, the traditional approaches can be augmented with consequence scores to identify high impact patterns. Additionally, the relationship between potential and actual consequence scores provide inferences regarding performance of M-RCS. For instance, it is expected that the more the difference between the two scores, the better the performance of the M-RCS. Many such inferences are exploited to assess performance of M-RCS.

1.4. Contribution

Several contributions are made from the work. First, the work considers temporal effect in recurrence of accident patterns and provides an approach for mining temporally frequent accident paths (APs). In this work, an algorithm is presented that exponentially weighs the patterns, i.e., the older the pattern, the smaller the weight (similar to exponential smoothing in forecasting [28]). This method captures the temporal effect of data.

Second, the study provides a simple to implement and easy to interpret data mining methodology for assessment of P-RCS and M-RCS. The results from temporal FIG (T-FIG) is compared with that of traditional FIG and inference are drawn about the effectiveness of P-RCS. Moreover, the impact is associated with every accident path, which is reduced by employing M-RCS. Hence, when an accident is realized, the actual consequence score is expected to be less than the potential consequence score. The effectiveness of M-RCS is dependent on the difference between actual and potential consequence scores, i.e., the more the difference the better the effectiveness of M-RCS. In this work, we have exploited this property to examine the effectiveness of M-RCS.

Our work also augments the traditional frequent itemset generation approach with consequence score associated with each incident. This will help in identifying high impact accident patterns. Overall, the proposed approaches will assist safety management in decision making by answering the following questions:

(1) What are the frequent accident paths?
(2) For which accident paths, the risk control systems have performed successfully?
(3) For which accident paths, either the preventive risk control systems are absent or ineffective?
(4) Which accident paths have a high probability of occurring in the future?
(5) For which accident paths, the mitigating risk control systems fail to mitigate the impact?
(6) What are the high impact accident paths?

In this paper, we propose a tool for assessment of the performance of RCSs by modifying traditional FIG in three different forms to analyze accident patterns in an incident dataset.

Table 1
Attributes required for each analysis.

Attributes	Analysis techniques
Hazardous element, Accident causing mechanism, Accident, Incident category	FIG, T-FIG, ESIG, High_Impact_IG
Date of accident	T-FIG
Actual consequence score	ESIG, High_Impact_IG
Potential consequence score	ESIG

2. Methodology

The primary objective of this study is to upgrade the existing pattern analysis methods so that we can draw inferences regarding P-RCS, M-RCS, and high impact accident paths. We propose a framework (see Figure 2) which comprises four different types of analyses (i) frequent itemset generation (FIG), (ii) temporal frequent itemset generation (T-FIG), (iii) elevated severity itemset generation (ESIG), and (iv) High impact itemset generation (High_impact_IG). The results obtained from each analysis are compared with each other to draw conclusions about the performance of risk control systems. Each analysis is performed with different set of input data. The categorical attributes, i.e., HE, ACM, Acc, considered for itemset generation remain same in all the four analysis. The latter three analyses are the modified version of FIG; therefore each requires some additional attributes for itemset generation. For example, T-FIG required time of occurrence of incident to weigh it with temporal effect (explained in C part of this Section). Figure 2 demonstrates how different attributes serve as input for different analyses and the attributes required for each analysis is tabulated in Table 1. The overall schema of the dataset considered is

$$D: \langle HE, ACM, Acc, ACS, PCS, Date, IC|, ACS \leq PCS > ,$$

where HE is hazardous element/material, ACM is accident causing mechanism, Acc is accident, ACS is actual consequence score, PCS is potential consequence score, Date is the date of the incident and IC is the incident category, i.e., whether a record is about an incident or a near-miss.

2.1. Terminologies

Definition 1. A **near miss** is an incident which did not result in property damage or personal injury but it could easily cause injury or damage if there was slight shift in position and time.

Definition 2. The **chain of event** analysis studies accident path as latitudinal propagation from hazardous element to the accident by accident causing mechanism.

Definition 3. [29]**.** An **accident** an unplanned and uncontrolled event in which the action or reaction of an object, substance, person or radiation results in personal injury or the probability thereof.

Definition 4. [29]**. An incident** is work-related event(s) in which an injury or ill health (regardless of severity) or fatality occurred, or could have occurred.

Definition 5. A **hazardous element/ material (HE)** is a source or situation having the potential to cause harm or loss.

Definition 6. An **accident causing mechanism (ACM)** is an unsafe act or condition which actuates the HE to realize into an accident.

2.2. Problem formulation

An accident path is defined with five tuples as AP: <HE, ACM, Acc, Tar, Th>, where HE is a hazardous element, ACM is accident causing mechanism, Acc is an accident, Tar is target and Th is a threat. The performance of P-RCS and M-RCS is effected by the patterns of the set

Table 2
One-to-one comparison matrix for assessing the performance of RCSs.

FIG P	FIG A	T-FIG P	T-FIG A	ESIG P	ESIG A	High-impact-IG P	High-impact-IG A
Ineffective P-RCSs	Effective P-RCSs	Have ineffective or no M-RCSs and are frequently recurring	Are frequent but no incident reported with ineffective or absent M-RCSs	Are frequently recurring and caused high severity in past*	Are frequent but no incident reported with high severity*
Have been recurrent in recent time, currently absent RCSs, and probable to recur in future		Have ineffective or no M-RCSs		Are not frequent but hold high severity*	
		Have increasing trend and ineffective or absent M-RCSs	Have increasing trend but no incident reported with ineffective or absent M-RCSs	Have increasing trend and also hold high severity*	Have increasing trend, however, no high impact incident reported in past*
		Have ineffective or no M-RCSs		Less probable to recur in future but hold high severity*	
				Have ineffective or absent M-RCSs and hold high severity	Have ineffective or absent M-RCSs but hold low severity
				Hold high severity*	
Not applicable			P: AP is present in the result of the corresponding analysis		
Same as the upper half of the matrix			A: AP is absent in the result of the corresponding analysis		

*Require further comparison with some other analysis technique to draw inference about RCSs.

<HE, ACM, Acc>. Hence, we have considered these attributes (say p) for pattern analysis. Let the set of categorical values for each attribute is Cp, then the set of all possible patterns, say *accident_pat*, from these attributes can be given as:

$\text{accident}_{\text{path}} = \text{accident}_{\text{path}} > \text{Cp} \ \forall \ \text{p}$, (1) where accident_path is initialized with ∅, \ represents the Cartesian product of two sets.

Definition 7. In a dataset 'D' of N records, any accident path k' is considered as frequent if its support is greater than threshold support, i.e., $support(k') \geq \text{min\_}support$, where

$$\text{support}(\text{k}') = \frac{\text{N}(\text{k}')}{\text{N}} \tag{2}$$

$N(k')$ is number of records of the pattern k'.

Lemma 1. The total number of accident paths present in dataset must not exceed $\prod_{\text{p}=1}^{\text{p}} |\text{C}_\text{p}|$.

Proof. For any p^{th} attribute, the total number of categorical value is given as $|\text{C}_\text{p}|$. The Cartesian product of two sets results in the pairing of each categorical value in one set with all categorical values in the other set. Therefore for any two sets, $\text{C}_{\text{p}+1}$ and C_p, we have total $|\text{C}_{\text{p}+1}| \times |\text{C}_\text{p}|$ number of pairs. Hence, with p number of total attributes, the total number of possible combinations is $\prod_{\text{p}=1}^{\text{p}} |\text{C}_\text{p}|$. There will be some combinations in the set which are practically not possible, hence will not appear in the dataset. If $D(k')$ represents the patterns present in dataset, then $\text{D}(\text{k}') \subseteq \text{accident\_path}\ (\text{k}')$.

For the analyses explained later, the meaning of different notations is as follows. For any instance 'i' in the dataset $D(i)$, the accident path associated with it is given by k′, the period of in which instance is recorded is 't' and the actual and potential consequence score associated with it is $\text{ACS}_{\text{k'i}}$ and $\text{PCS}_{\text{k'i}}$, respectively.

2.3. Frequent itemset generation

Association rule mining (ARM) is employed on large database for generating recurring patterns. Different algorithms, such as Apriori algorithm [30], FP-Growth [31], etc., have evolved over the period to perform ARM. One such method is the frequent itemset generation (FIG). The combination of different itemset in a database is generated, and frequent itemset is selected based on parameters support, lift, and confidence of each pattern. In this work, each accident path component is considered as items, and an itemset represents one accident path.

The process of FIG involves generating all possible itemsets by Cartesian product of the categorical values of all the attributes (as explained in Equation 1). Further the support of each itemset is computed using Equation 2 and an itemset is considered frequent according to the criteria mentioned in Definition 3. The pseudo code of the FIG can be obtained by replacing the equation of support step 3 in Algorithm 1 with equation 2.

2.4. Temporally frequent accident path generation

In cases, when the data is spread over a long period, traditional FIG may result in biased result since patterns are influenced by temporal factor. There can be cases where an AP is very frequent for short time instance, and due to some new policy, its frequency may decrease. The case can be reverse as well, i.e., an AP may have an increasing trend in response to some new policy or installation. However, the overall frequency will not be very high. In traditional FIG, the former pattern will appear in the result, and the later may not appear, thus, giving some biases in the result. Usually, the management is interested in knowing the accident paths with increasing trend even if its overall frequency is not alarming.

The proposed T-FIG considers this temporal factor and weighs the patterns according to their corresponding period. The older the pattern, the smaller is its weight. This is achieved by exponentially weighing the patterns.

Definition 8. The exponential weight of an accident path k', at any instance of time period 't' is

$$\text{Fr}(\text{k}')_\text{t} = \alpha(1-\alpha)^{\text{T}-\text{t}}\text{fr}(\text{k}')_\text{t} \tag{3}$$

where, T is the total time period, $fr(k')_t$ is the frequency of accident path k' in time period 't', α is smoothing constant such that $0 < \alpha < 1$.

Proof. For a time series data of the total period 'T', the forecasted value for period 't+1' using exponential smoothing is

$$\text{F}_{\text{t}+1} = \alpha\text{x}_\text{t} + (1-\alpha)\text{F}_\text{t};\ \text{t} > 0,\ \ \text{F}_1 = \text{x}_1 \tag{4}$$

where x_t is the actual value for the period 't'. Equation 4 can be re-written as

$$\text{F}_{\text{t}+1} = \alpha\text{x}_\text{t} + (1-\alpha)(\alpha\text{x}_{\text{t}-1} + (1-\alpha)\text{F}_{\text{t}-1})$$

$$\text{F}_{\text{t}+1} = \alpha\text{x}_\text{t} + (1-\alpha)\alpha\text{x}_{\text{t}-1} + (1-\alpha)^2\text{F}_{\text{t}-1}$$

$$\text{F}_{\text{t}+1} = \alpha\text{x}_\text{t} + (1-\alpha)\alpha\text{x}_{\text{t}-1}\ldots + (1-\alpha)^{\text{T}-1}\alpha\text{x}_1 + (1-\alpha)^{\text{T}}\text{F}_1$$

In the above equation it can be noticed that at any time instance 't', the coefficient of x_t is $\alpha(1-\alpha)^{T-t}$. This means that at any time instance 't', the weight of x_t is $\alpha(1-\alpha)^{T-t}$, i.e., it is being exponentially weighed with respect to time 't'. In T-FIG, the variable x_t represents the frequency of an accident path k'in the time period 't', i.e., $Fr(k')_t$.

The value of α is computed by performing exponential smoothing based forecasting of each accident path. This means that we have a different value of α for each accident path. The α is selected such that the mean square error (MSE) is minimized. The MSE is

$$\text{MSE} = \frac{\sum_{\text{t}=1}^{\text{T}} (\text{x}_\text{t} - \text{F}_\text{t})^2}{\text{T}} \tag{5}$$

Therefore, the optimal value of α is calculated as:

$$\min \sum_{t=1}^{T} (x_t - F_t)^2$$

$$s.\ t.\ \ \{0 < \alpha < 1 \tag{6}$$

Definition 9. In a dataset of N records, an accident path k' is considered as temporally frequent in T-FIG if its temporal support is greater than threshold support, i.e., $\text{support}_{\text{temporal}}(\text{k}') \geq \text{min\_support}_{\text{temporal}}$ where

$\text{support}_{\text{temporal}}(\text{k}') = \frac{\sum_{\text{t}=1}^{\text{T}} \text{Fr}(\text{k}')_\text{t}}{\sum_{\text{t}=1}^{\text{T}} \alpha(1-\alpha)^{\text{T}-\text{t}_{\text{N}_\text{t}}}}$ (7) where, N_t is the number of records in time period 't' and $\sum_{t=1}^{T} Fr(k')_t < N(k')$. The pseudo code for T-FIG is given as Algorithm 1.

Proposition 1. The time complexity of T-FIG is $O(p + \prod_{p=1}^{p} |C_p|) \cong O(\prod_{p=1}^{p} |C_p|)$. If average number of categorical values for an attribute 'p' is C_{pavg}, then time complexity is $O(C_{pavg}^{p})$.

Proof. In Algorithm 1, the first loop will iterate for the number of attributes considered for analysis, i.e., p. The expression $\text{accident}_{\text{path}} = \text{C}_\text{p} > \text{C}_{\text{p}-1}$ will take the time of $O(1)$ in each iteration. Therefore, the time complexity of the first loop is $O(p)$. The variable accident_path comprises $\prod_{p=1}^{p} |C_p|$ number of accident path, hence, the second loop will iterate for $\prod_{p=1}^{p} |C_p|$ number of times and the steps within the loop are simple mathematical, conditional and assignment operations, therefore they will take $O(1)$ running time in each iteration. Hence, the complexity of the second iteration is $O(\prod_{p=1}^{p} |C_p|)$. The total complexity will be $O(p + \prod_{p=1}^{p} |C_p|)$. Since $p \ll \prod_{p=1}^{p} |C_p|$, therefore time complexity is approximately equal to $O(\prod_{p=1}^{p} |C_p|)$. Now, we know an average frequency is given as

$\frac{|\text{C}_1| + |\text{C}_2| \ldots + |\text{C}_\text{p}|}{\text{p}} = \text{C}_{\text{pavg}}$,

Therefore,

$|\text{C}_1|*|\text{C}_2|.....*|\text{C}_\text{p}| \leq \text{C}_{\text{pavg}}*\text{C}_{\text{pavg}}.....\text{p times}$

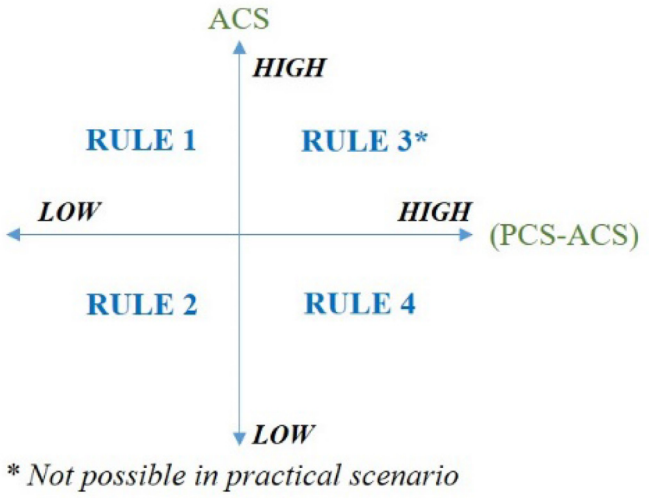

Figure 3. The four quadrant representation of relationship between ACS and (PCS-ACS).

$$\prod_{p=1}^{p} |C_p| \leq C_{pavg}^{p}$$

The running time complexity can be rewritten as $O(C_{pavg}^{p})$.

2.5. Elevated severity itemset generation

The M-RCSs are placed between an accident and target in the chain of events approach to reduce the threat to target. If the actual consequence score associated with an accident path is high, it is interpreted that the path has caused severe harms. The high potential consequence score indicates that the path has the potential to cause severe harm. In order to assess the performance of M-RCSs, we exploit the relation between actual and potential consequence scores and calculated M-RCS failure index. There are four possible combinations of relation between value of ACS and difference between PCS and ACS, which is depicted in Figure 3. The following rules can be generated:

Rule 1: If the ACS is high and the difference between PCS and ACS is low, then it can be inferred that M-RCS is either absent or ineffective.

Rule 2: If the ACS is low and the difference between PCS and ACS is low, then it does not provide any information about the ineffectiveness of M-RCS.

Rule 3: If the ACS is high and the difference between PCS and ACS is high, such a case is not practically possible.

Rule 4: If the ACS is low and the difference between PCS and ACS is high, then M-RCS is effective for the corresponding accident path.

Calculation of M-RCS failure index

$$\text{ACS above limit}_i(\text{AAL}_i) = \begin{cases} \frac{\text{ACS}_i - \text{th}_{rs}}{\text{th}_{rs}}, & \text{if } \text{ACS}_i \geq \text{th}_{cs} \\ 0, & \text{if } \text{ACS}_i < \text{th}_{cs} \end{cases} \quad (8)$$

$$\text{Difference score}_i(\text{DS}_i) = \begin{cases} \frac{\text{PCS}_i - \text{ACS}_i}{\text{th}_d}, & \text{if } (\text{PCS}_i - \text{ACS}_i) \geq \text{th}_d \\ 0, & \text{if } \text{ACS}_i < \text{th}_d \end{cases} \quad (9)$$

$$\text{Failure Score}_i(\text{FS}_i) = \begin{cases} (\text{AAL})_i + (\text{DS})_i, \\ \text{if } (\text{ACS}_i \geq \text{th}_{cs} \text{ AND } (\text{PCS}_i - \text{ACS}_i) \geq \text{th}_d) \\ 0, \quad \text{otherwise} \end{cases} \quad (10)$$

$$(\text{M}-\text{RCS failure index})_i = \begin{cases} \frac{\text{FS}_i}{\min(\text{FS})}, & \text{if } (\text{ACS}_i \geq \text{th}_{cs} \text{ AND } (\text{PCS}_i - \text{ACS}_i) \geq \text{th}_d) \\ 1, & \text{otherwise} \end{cases} \quad (11)$$

where i refers to a particular instance in dataset, and th_{cs} and th_d are the minimum allowable consequence score and minimum allowable difference between PCS and ACS. The support formula in (2) is modified to incorporate M-RCS failure index and the new support formula is given as follows:

$$\text{support}_{\text{ESIG}}(\text{k}') = \frac{\sum_{n=1}^{N_{k'}} (\text{M} - \text{RCS failure index})_n * \text{N}(\text{k}')_n}{\sum_{i=1}^{N} (\text{M} - \text{RCS failure index})_i * \text{N}_i} \quad (12)$$

where $N_{k'}$ is the number of records in D having accident path k'. The pseudo code for implementing this work is provided as Algorithm 2.

Proposition 2. The time complexity of ESIG is $O(p + \prod_{p=1}^{p} |C_p| + N + N + N)$. Since, $O(p) \ll O(\prod_{p=1}^{p} |C_p|) \ll O(N)$, therefore, time complexity of the algorithm can be given as $O(3N) \cong O(N)$.

Proof. The first loop iterates 'p' number of times and the step in this loop takes $O(1)$ time, therefore, time complexity of this loop is $O(p)$. The second and third loop iterates N number of times, therefore, the running time complexity for each iteration is $O(N)$. The Min(FS) function, used for implementing Equation 11, returns the minimum value from one-dimensional array. The 1D array is of the size N as each entry is the computed FS value for each instance. The time complexity of this function in case of non-arranged array is O(N). Running time complexity of fourth loop is $O(\prod_{p=1}^{p} |C_p|)$. The reason is explained earlier in proposition 1. Therefore, the time complexity is $O(p + \prod_{p=1}^{p} |C_p| + N + N + N)$. For a large dataset following inequality holds

$O(p) \ll O(\prod_{p=1}^{p} |C_p|) \ll O(N) and O(3N) \cong O(N)$

Therefore, for large value of N, the running time complexity is $O(N)$.

2.6. High impact itemset generation

Along with the frequency of accident path, we have also considered the consequence score associated with it. In this analysis, we have considered both near miss and incident data. This is so because the paths followed by near miss are also of significant importance if the consequence associated with it is severe. The actual consequence score is considered for this analysis. The ACS for any record '*i*' having path k' is given as $ACS_{k'i}$. The support for each AP is calculates as

$$\text{support}_{\text{highimpact}}(\text{k}') = \frac{\sum_{i=1}^{N} (\text{ACS}_{\text{k}'i})}{\sum_{i=1}^{N} \text{ACS}_i} \quad (13)$$

where $\sum_{i=1}^{N} (ACS_{k'i})$ is the sum of ACS for all instances where accident path is k'. The pseudo code of the proposed High-impact-IG is obtained by replacing the equation of support in step 3 of Algorithm 1 by Equation 13.

2.7. Performance assessment of RCSs

Several interpretations can be obtained by comparing the findings of the analyses (e.g., FIG, T- FIG, ESIG, and High-impact-IG) performed in the methodology. Suppose the findings of two analysis techniques is set A and B, then four conditions are possible for an accident path: (i) the accident path is present in both A and B, (ii) present in A and absent in B, (iii) absent in A and present in B, (iv) absent in both. Such a one-to-one comparison among all the four analyses is provided in Table 2. The P and A denote the presence and absence of an accident path in the result of the corresponding analysis method, respectively. The texts in each cell of Table 2 show the interpretation of the performance of RCSs. For instance, an accident path which is present in the results of both T-FIG and ESIG have increasing trend and ineffective or absent M-RCSs. This is so because if the accident path is present in T-FIG, this means it has high temporal weights in recent time period and therefore probable to recur in future, while its presence in ESIG denotes that its M-RCS failed to reduce the severity, resulting in small difference between

actual and potential consequence score.

Apart from one-to-one comparison, comparison among three or four sets can also be carried out. This can be done by interpreting the distribution of accident paths across the analysis methods through a four-set Venn diagram. The inferences that can be obtained from comparison of results from more than two analyses methods are discussed in the Section 3.

3. Case study

3.1. Studied plant

In this work, we have studied the occupational accidents in a steel manufacturing plant in India. The plant broadly comprises seven safety domains: mining, electrical, rail/road, process, contractor, construction and electric overhead transfer (EOT) crane safety. The focus of this case study is analysis of risk control systems involved in process safety. The scope of process safety includes all the safety requirements during steel making process and therefore, this domain ensures prevention of extremely high severity accidents. Therefore, analysis of process safety data is highly important. There are two major units involved in steel making process: (i) iron making unit, and (ii) steel manufacturing unit. The major areas of concern in iron making unit are blast furnace, agglomeration unit (sinter and pellet plant) and coke oven plant. The output from the iron making unit is hot molten metal, which is transferred through torpedo on rail to hot metal treatment (HMT) plant. The unit which deals with hot metal transfer is known as hot metal logistics. In HMT, the impurities are removed from hot metal and transferred to caster unit using EOT crane, where the casting is done to produce the desired steel products of different quality. A part of process safety also overlaps with the scope of EOT crane safety.

3.2. Dataset

The data is collected through an IT-based system present as the part of safety management system in the studied plant. An intra-net web interface is used for the data collection. In case an accident occurs, the worker present at the scenario logs the details of the accident in the web application. To ease the process of logging process, the system is designed to collect important information regarding the accident in form of categorical dropdowns. This prevents troublesome narration of accident in the text format, saves time of the logger, and also reduces subjectivity. The text box is also provided for additional information. After the details are recorded, based on the consequence score, further investigation of the accident is carried out by the safety experts. In order to perform the data logging, a standard procedure is documented and followed by the organization provided by their safety management. In order to achieve the consistency in data collection, periodic workshops are conducted to train the employees on the logging system.

The dataset was collected for the period from FY'16 to FY'18. The collected data comprises 612 records, where each record details the facts associated with a particular accident. The dataset consists of 22 attributes of mixed data type, i.e., numerical, categorical, and text. Among 612, there are 242 incident records and 370 near miss records. For ESIG, only incident reports are considered as mitigating RCSs are required only after an incident has occurred. As demonstrated in Figure 2 and Section 2, six attributes are used for this analysis. The detailed definition of each attribute is provided in Table 3.

3.3. Results and analysis

In this study, the temporal effect is studied on quarterly basis. This is done because the concerned steel plant reviews its safety status quarterly. Hence most of the measures are taken and implemented quarterly. Therefore, for the considered dataset of three years, the total time period, T, is 12. The three attributes HE, ACM, and Acc have 8, 10, and 13 categorical values, respectively. There are 190 unique accident paths in the dataset. The number of paths that can be obtained from the Cartesian product of categorical values is 1040. This provides evidence that Lemma 1 is valid. The support for all the four analyses is considered as 0.01. The threshold th_d and th_{cs} values are taken as 30 and 70, respectively. The threshold is provided by safety experts of the concerned organization. The α value is calculated by performing exponential smoothing based forecasting and minimizing mean square error.

21, 14, 24, and 31 accident paths are obtained from FIG, T-FIG, ESIG, and High-impact-IG which are provided in Tables A1, A2, A3, and A4, respectively in the appendix. Let each set of results as shown in Tables A1-A4 be referred to as A, B, C, and D, respectively. Some of the paths are overlapping in the four sets. A Venn diagram is given in Figure 4 to demonstrate the distribution of paths among four sets. The left diagram represents the Venn diagram for the general case of four sets. The right side is the data distribution of four sets that are obtained from our analyses. For example, the set "ABCD" in the left diagram refers to the elements which are present in all the four sets. In our results, it is observed that there are four accident path patterns which are obtained from all four analyses. Therefore, on the right side, the set "ABCD" is written as four. Different inferences can be drawn from the set of paths obtained, and they are discussed below.

Inference 1: The set $P = A \cap B$ are the APs for which the P-RCS are either ineffective or absent.

There are nine APs which are present in both the set, A and B, and are provided in Table 4. These patterns are frequent, and their frequency has not decreased over the period or have a recent increasing trend. It can be inferred that either the management did not take any measure to prevent them or the measure taken is not working effectively. Figure 5 demonstrates the time series plot of the temporal weight of these nine patterns.

Inference 2: The set $Q(x) = \{A | \, x \notin ((\mathrm{A} \cap \mathrm{B})\mathrm{OR}(\mathrm{A} \cap \mathrm{C})\mathrm{OR}(\mathrm{A} \cap \mathrm{D})\}$ have APs with decreasing trend and their corresponding P-RCS and M-RCS are effective.

Q is a set of APs which are present only in results of FIG and not in any other set. There are two APs in set Q and are given in Table 5. The absence of these APs in T-FIG denotes that they have a decreasing trend. Hence, the P-RCS to prevent these patterns are effective. Further, their absence in the result of ESIG denotes that M-RCS are also effective for these accident patterns. Figure 6 provides the time series representation with temporal weights of the APs with decreasing trend and effective M-RCS and P-RCS.

Inference 3: The set $R(x) = \{x \in (B - A)\}$ has APs of increasing trend, hence are most probable to occur in future.

The set R comprises five APs which are provided in Table 6. These are the APs which are absent in set A. These APs had either no occurrence or low occurrence in the initial period; however, their frequency has increased in recent periods. This is the reason its overall weight is more than the threshold since new records are weighed more than the older records in T-FIG. Figure 7 represents the temporal weight based time series plot of the five APs having an increasing trend.

Inference 4: The set $S(x) = \{x \in ((B \cap D) - A - C)\}$ comprises APs which have increasing trend and result in high impact.

The set S comprises 3 APs which are given in Table 7. These patterns have a high probability of occurring in the future and have resulted in high severity incidents in past incidents. Therefore, immediate appropriate mitigating risk control systems are required for severity control of the incidents. The temporal weight based trend is provided in Figure 7.

Inference 5: The set $T(x) = \{x \in (B \cap C \cap D) - A\}$ comprises APs which have increasing trend, hold high risk and have either ineffective M-RCS or it is absent.

These are the patterns which hold a high priority in management's attention. This is so because firstly they have the probability of occurring in the future. Secondly, they have resulted in high severity in the

Table 3
Details of the attributes considered.

Attribute	Data Type	Description	Categorical value
Hazardous element	Categorical	This attribute comprises 8 categorical values describing the hazardous material/ elements present at workplace.	Hot metal/steel/slag Gas Chemical Dust and steam Activity related hazard Uncontained water/oil Coke High pressure material
Accident causing mechanism	Categorical	This attribute comprises 10 categorical values describing the unsafe act or condition.	Damaged/degraded/poorly maintained equipment Improper material/material quality/ equipment design Improper work by operator/worker Inadequate/ unavailable SOP Improper working environment Failure/malfunctioning of equipment Presence of unwanted/flammable material Abnormal process parameter Improper/exceeded process/operating parameters Lack of communication/ supervision
Accident	Categorical	This attribute comprises 13 categorical values describing the outcome of ACM.	Spillage of hazardous material Leakage of hazardous material Fire/ flame generation Disturbance in process parameters/ equipment shutdown Degradation/breaking of equipment Overflow of hazardous material Explosion Exposure to toxic gas/ high temperature Splash Dashing/collision Derailment Fall of material Water logging/flooding situation
Actual consequence score	Numerical	This is the consequence score associated with accident path. It is less than or equal to PCS since M-RCS are present to reduce the impact.	ACS ∈ [15, 130]
Potential consequence score	Numerical	This is the score of accident path in case if all the M_RCS fails.	PCS ∈ [15, 130]
Date of incident	Date	The date on which the incident occurred.	
Incident Category	Categorical	This provides information whether the reported record is an incident or a near miss. It comprises of only two categories.	Near Miss Incident

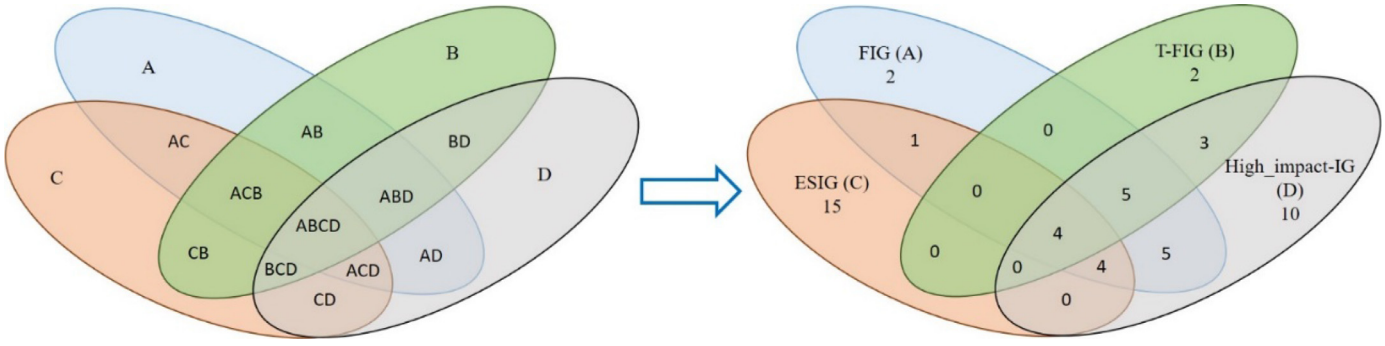

Figure 4. Venn diagram for pattern distribution of four analysis.

past and lastly, the M_RCS for these accident paths were either ineffective in controlling the severity, or they were not implemented by management. In the dataset considered, no such pattern is identified.

Inference 6: The set $U(x) = \{x \in ((B \cap C) - A - D)\}$ comprises APs which have increasing trend and either the M-RCS ineffective or it is absent.

The set U comprises APs which are absent in set A, hence have an increasing trend. This can be interpreted that they have a high probability of occurrence in the future. These patterns are important since they are probable to occur in the future. Since, these APs are also present in set C therefore, M-RCS for these are either ineffective or it is absent. Hence, safety management must take immediate action to reduce the frequency and severity of these patterns. It should also be noted that, since they are not present in set D, therefore, these patterns must have not resulted in high impact accidents in past. In the considered dataset, no such path is observed in set U.

Inference 7: The set $V(x) = \{x \in (C \cap D) - A - B\}$ comprises APs which have severe consequences and either M_RCS is ineffective or it is absent.

The AP in this set has resulted in high severity incidents in the past, and the risk control for controlling the severity was either ineffective or absent.

Table 4
Accident patterns whose frequency do not decrease over time (P).

Path Code	HE	ACM	Accident
Path 1.1	Hot metal/ steel/ slag	Improper work by operator/worker	Dashing/collision
Path 1.2	Hot metal/ steel/ slag	Improper work by operator/worker	Degradation/breaking of equipment
Path 1.3	Hot metal/ steel/ slag	Damaged/degraded/poorly maintained equipment	Spillage of hazardous material
Path 1.4	Hot metal/ steel/ slag	Improper work by operator/worker	Spillage of hazardous material
Path 1.5	Gas	Failure/malfunctioning of equipment	Exposure to toxic gas/ high temperature
Path 1.6	Gas	Damaged/degraded/poorly maintained equipment	Fire/ flame generation
Path 1.7	Gas	Damaged/degraded/poorly maintained equipment	Leakage of hazardous material
Path 1.8	Gas	Failure/malfunctioning of equipment	Leakage of hazardous material
Path 1.9	Chemical	Damaged/degraded/poorly maintained equipment	Leakage of hazardous material

Inference 8: The set $W(x) = \{x \in (A \cap B \cap C \cap D)\}$ comprises APs which are frequent, have severe consequence and have either ineffective RCSs or they are absent.

The APs in set W holds the highest priority for the attention of management. These are the paths for which both the RCSs are ineffective or absent and if not prevented will result in a highly severe incident. Moreover, they have an increasing trend. The four accident paths from our analyses belong to set W and are provided in Table 8. The temporal weight based time-series plot of these four paths is provided in Figure 5.

4. Discussion

4.1. Data mining as performance assessment tool

The presented methodology is first data mining based approach in the domain of performance assessment of RCSs. This method helps in analyzing large amount of data with minimum human involvement. Moreover, the algorithms reflect the exact information that is present in the data, therefore, subjective biasness in the result is absent. However, if there are factors other than performance of RCSs which may affect the trend, such factors cannot be captured or comprehended through data mining methods. Moreover, in case of situations where the trend experiences sudden peak, in such scenarios as well, the proposed methodology may not be able to provide considerable reason for the peak. In such cases, the human judgments are more reliable than data mining inferences.

4.2. Managerial implications

The results are discussed with respect to the six questions asked in Section 1.4 and necessary recommendations are provided. It must be noted that the results about the accident patterns and performance of RCSs are quite different and more insightful from the previous works on the applications of data mining methods in occupational safety.

What are the frequent accident paths?

The results obtained from FIG, as provided in Table A1 , are the frequent accident paths. Among them, “spillage of hot metal/steel/slag due to improper work by operator/worker” is the most frequent accident path. There could be several reasons for improper work by operator/worker such as, negligence, incompetence, lack of safety awareness, etc. Since this path is highly frequent, the safety experts must analyze each record separately and identify root causes. If the workers are found to be incompetent, then performance of the worker during training should be checked. If the worker is found to be skillful during training, then training methods should be revisited, and further the skill and task mapping for each worker shall be done properly. In case of lack of safety awareness or negligence, safety sessions should be conducted by safety experts periodically to motivate workers towards safety. Moreover, the upper management shall examine the pressure of high production on workers, since this may also effect safety willingness of workers. The next most frequent path is “leakage of gas due to damaged/degraded/poorly maintained equipment”. There could be possible reasons for damaged/degraded/poorly maintained equipment, such as, poor quality of product, incorrect maintenance period, equipment used incorrectly, etc. In this scenario, team of engineers must design checklist for checking the quality of products after they are delivered from vendors and another checklist should be designed for maintenance activity. Moreover, the safety and engineers team must also timeline for maintenance activity for each equipment and an online logging system should be designed for keeping record of inspection and maintenance schedules. Similarly, interventions for other paths should also be designed.

For which accident paths, the risk control systems have performed successfully?

The paths provided in Table 5 are the paths for which the RCSs were

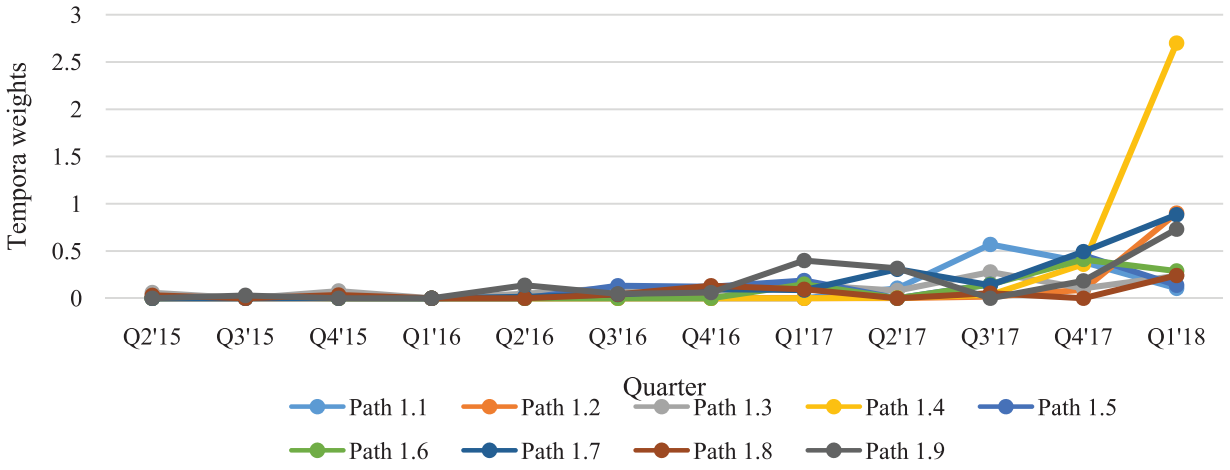

Figure 5. Temporal weight time series representation of APs for which the P-RCS are either ineffective or absent.

Table 5
Accident patterns with decrease in frequency over time (set Q).

Path Code	HE	ACM	Accident
Path 2.1	Hot metal/ steel/ slag	Damaged/degraded/poorly maintained equipment	Leakage of hazardous material
Path 2.2	Gas	Failure/malfunctioning of equipment	Disturbance in process parameters/equipment shutdown

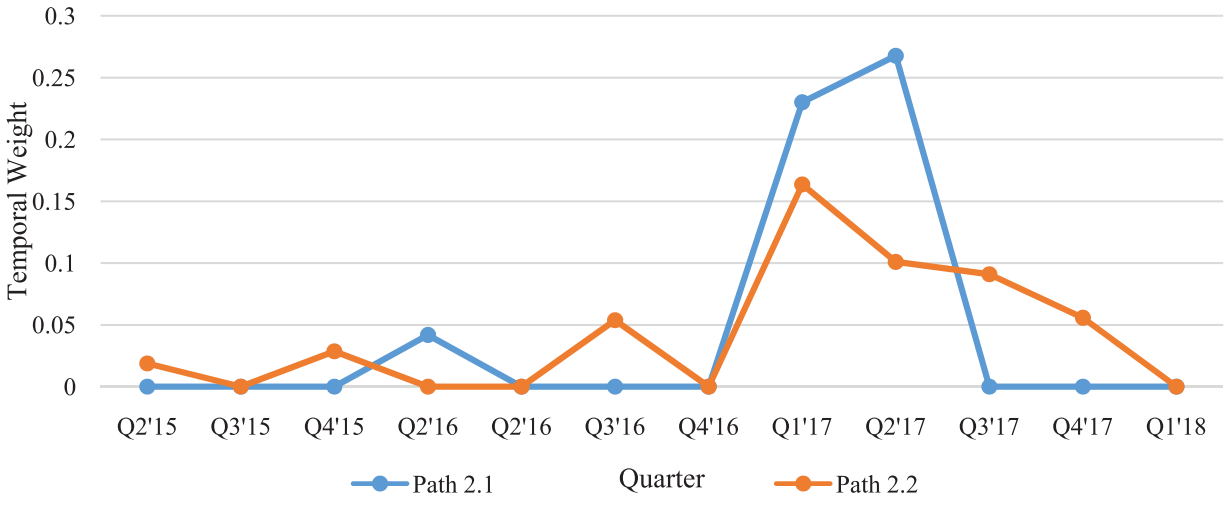

Figure 6. Temporal weight based time series representation of APs with decreasing trend.

Table 6
Accident patterns present with increasing trend (set R).

Path Code	HE	ACM	Accident
Path 3.1	Hot metal/ steel/ slag	Damaged/degraded/poorly maintained equipment	Degradation/breaking of equipment
Path 3.2	Hot metal/ steel/ slag	Presence of unwanted/flammable material	Explosion
Path 3.3	Hot metal/ steel/ slag	Improper work by operator/worker	Overflow of hazardous material
Path 3.4	Chemical	Failure/malfunctioning of equipment	Fire/ flame generation
Path 3.5	Chemical	Inadequate/ unavailable SOP	Spillage of hazardous material
	SOP: Standard Operating Procedure		

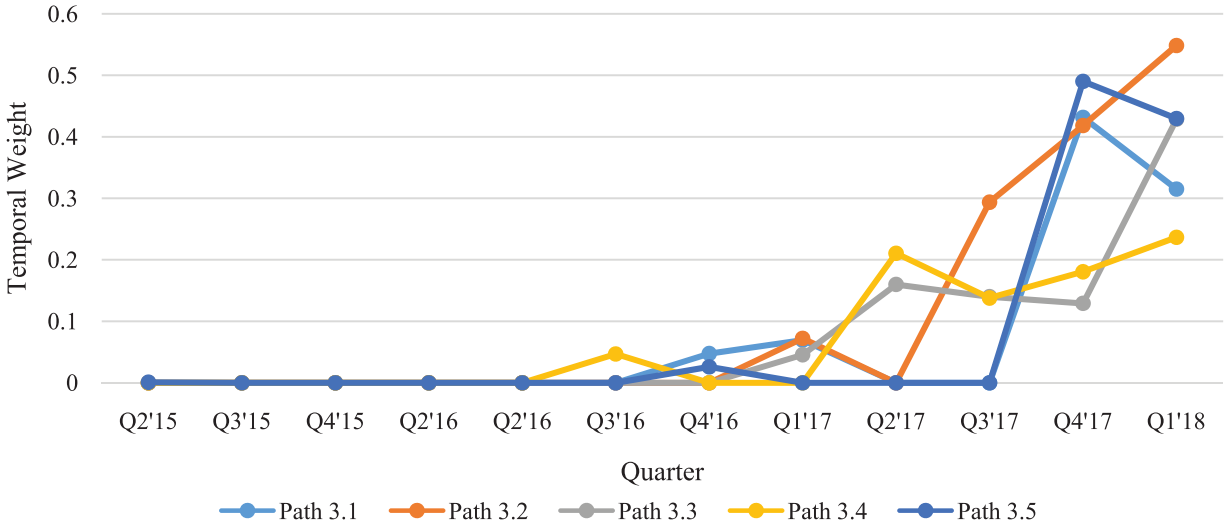

Figure 7. Temporal weight time series representation of *APs of increasing trend.*

effective. These are the paths which are present only in set A. The safety management shall revisit the interventions decisions taken for these accident paths. If similar cases are observed in other divisions or departments, similar interventions may be effective in those cases as well.

For which accident paths, either the preventive risk control systems are absent or ineffective?

The accident paths presented in Table 4 are the paths for which the P-RCSs are either ineffective or absent. Among the nine accident paths identified, “spillage of hot metal/steel/slag due to improper work by operator/worker” is the most temporally recurring accident path. The second and third most temporally recurring accident paths are “leakage of gas and chemical due to damaged/degraded/poorly maintained

Table 7
High impact accident patterns with increasing trend (set S).

Path Code	HE	ACM	Accident
Path 4.1	Chemical	Inadequate/ unavailable SOP	Spillage of hazardous material
Path 4.2	Hot metal/ steel/ slag	Damaged/degraded/poorly maintained equipment	Degradation/breaking of equipment
Path 4.3	Hot metal/ steel/ slag	Presence of unwanted/ flammable material	Explosion

Table 8
High frequency-high severity patterns (set W) due to ineffective RCS.

Path Code	HE	ACM	Accident
Path 8.1	Gas	Damaged/degraded/poorly maintained equipment	Leakage of hazardous material
Path 8.2	Chemical	Damaged/degraded/poorly maintained equipment	Leakage of hazardous material
Path 8.3	Hot metal/ steel/ slag	Damaged/degraded/poorly maintained equipment	Spillage of hazardous material
Path 8.4	Hot metal/ steel/ slag	Improper work by operator/worker	Spillage of hazardous material

Algorithm 1
TemporalFrequentAccidentPathGeneration(D).

Input: Cp <set of categorical values of pth attribute>, p <number of attributes to be considered>, T<Time period considered>
Output: Temporally frequent item-set
Initialize: Temporal_Frequent_accident_path =φ, accident_path = φ, C0 = φ
Begin
For $p' \in p$
$\text{accident}_{\text{path}} = C_p > C_{p-1}$
End for
For $k' \in$ accident_path
$support_{temporal}(k') = \frac{\sum_{t=1}^{T} Fr(k')_t}{\sum_{t=1}^{T} \alpha(1-\alpha)^{T-t} N_t}$
If $support_{temporal}(k') \geq \min\_support_{temporal}$
Temporal_Frequent_accident_path $\leftarrow k'$
End if
End For
Return Temporal_Frequent_accident_path
End Begin

Algorithm 2
MitigationViolationFrequentItemsetGeneration(D).

Input: C_p <set of categorical values of p^{th} attribute>, p <number of attributes to be considered>, ARS, PRS, D
Output: Elevated severity accident path
Initialize: Elevated_severity_accident_path =φ, accident_path = φ, C_0 = φ
Begin
For $p' \in p$
$\text{accident}_{\text{path}} = C_p > C_{p-1}$
End for
For $n \in N$
If $(ACS_n \geq th_{cs}$ AND $(PCS_n - ACS_n) \geq th_d)$
Implement equation 8
Implement equation 9
Implement equation 10
End If
End For
For $n \in N$
If $(ACS_n \geq th_{cs}$ AND $(PCS_n - ACS_n) \geq th_d)$
Implement equation 11
Else
$(M-RCS\ failure\ index)_n = 1$
End If
End For
For $k' \in Itemset$
$support_{violation}(k') = \frac{\sum_{n=1}^{N_{k'}} (M - RCS\ failure\ index)_n * N(k')_n}{\sum_{i=1}^{N} (M - RCS\ failure\ index)_i * N_i}$
If $support_{violation}(k') \geq \min\_support_{violation}$
Elevated_severity_accident_path $\leftarrow k'$
End if
End For
Return Elevated_severity_accident_path
End Begin

equipment". From these finding it is clear that leakage of hazardous material is one of the most common temporally recurring event and major cause is damaged/degraded/poorly maintained equipment. The interventions for all the three paths have been discussed in question 1.

Which accident paths have a high probability of occurring in the future?

The five paths presented in Table 6 are most probable to recur in future. Among these paths, "explosion due to presence of unwanted/ flammable material like hot metal/steel/slag" has highest temporal weight in last quarter, therefore, this path holds highest probability to recur in future. Second most probable path to recur in future is "spillage of chemical due to inadequate/unavailable SOP". For the former path, periodic workplace cleaning and housekeeping should be done. Moreover, operators should be check the ladle and slag-pot for presence of any flammable material before pouring hot metal into the ladle and disposing slag into slag-pot. For the latter accident path, it is to be first identified if the SOP was present to perform the activity or not. If it was present, then SOP must not be adequate to perform the activity safely. In case, the SOP is not present, proper SOP should be prepared for the activity comprising possible hazards and risk of the activity.

For which accident paths, the mitigating risk control systems fail to mitigate the impact?

The results to be obtained from inference 5 and 6 will be the accident paths for which M-RCSs are either ineffective or absent. In the considered case study, no such accident paths were obtained.

What are the high impact accident paths?

The results obtained from High_impact_IG are the accident paths which hold high severity. The accident path holding highest risk is "leakage of gas due to damaged/degraded/poorly maintained equipment". "Leakage of chemical due to damaged/degraded/poorly maintained equipment" and "spillage of hot metal/steel/slag due to failure/malfunctioning of equipment" are next two most risky accident paths having same severity weight as 0.028.

5. Conclusions

5.1. Contribution to methodology

The methodology presented in the work contributes to performance assessment methodologies existing in occupational safety domain by using data mining approach. A novel methodology is provided by looking at the basic safety concepts from data point of view. The T-FIG is built on the concept that an effective P-RCS will result in decrease in corresponding accident frequency. Similarly, ESIG is built on the concept that an effective M-RCS will decrease the actual consequence score. Further, replacing frequency in FIG with ACS will result in

Table A1
Results of FIG.

Hazardous Element	Accident Causing Mechanism	Accident/incident	Support
Hot metal/ steel/ slag	improper work by operator/worker	dashing/collision	0.013
Hot metal/ steel/ slag	improper work by operator/worker	degradation/breaking of equipment	0.011
Hot metal/ steel/ slag	Inadequate/ unavailable SOP	Explosion	0.013
Hot metal/ steel/ slag	Inadequate/ unavailable SOP	Fire/ flame generation	0.011
Hot metal/ steel/ slag	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.011
Hot metal/ steel/ slag	damaged/degraded/poorly maintained equipment	spillage of hazardous material	0.028
Hot metal/ steel/ slag	failure/malfunctioning of equipment	spillage of hazardous material	0.026
Hot metal/ steel/ slag	improper work by operator/worker	spillage of hazardous material	0.062
Hot metal/ steel/ slag	Inadequate/ unavailable SOP	spillage of hazardous material	0.011
Gas	failure/malfunctioning of equipment	Disturbance in process parameters/equipment shutdown	0.011
Gas	improper work by operator/worker	Explosion	0.013
Gas	failure/malfunctioning of equipment	exposure to toxic gas/ high temperature	0.023
Gas	improper work by operator/worker	exposure to toxic gas/ high temperature	0.013
Gas	damaged/degraded/poorly maintained equipment	Fire/ flame generation	0.011
Gas	failure/malfunctioning of equipment	Fire/ flame generation	0.011
Gas	improper work by operator/worker	Fire/ flame generation	0.011
Gas	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.036
Gas	failure/malfunctioning of equipment	Leakage of hazardous material	0.021
Gas	improper work by operator/worker	Leakage of hazardous material	0.013
Chemical	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.033
Chemical	improper work by operator/worker	Overflow of hazardous material	0.011

Table A2
Results of T-FIG.

Hazardous Element	Accident Causing Mechanism	Accident/incident	Support
Hot metal/ steel/slag	improper work by operator/worker	dashing/collision	0.018
Hot metal/ steel/slag	damaged/degraded/poorly maintained equipment	degradation/breaking of equipment	0.014
Hot metal/ steel/slag	improper work by operator/worker	degradation/breaking of equipment	0.017
Hot metal/ steel/slag	presence of unwanted/flammable material	Explosion	0.013
Hot metal/ steel/slag	improper work by operator/worker	Overflow of hazardous material	0.013
Hot metal/ steel/slag	damaged/degraded/poorly maintained equipment	spillage of hazardous material	0.02
Hot metal/ steel/slag	improper work by operator/worker	spillage of hazardous material	0.051
gas	failure/malfunctioning of equipment	exposure to toxic gas/ high temperature	0.017
gas	damaged/degraded/poorly maintained equipment	Fire/ flame generation	0.016
gas	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.034
gas	failure/malfunctioning of equipment	Leakage of hazardous material	0.01
chemical	failure/malfunctioning of equipment	Fire/ flame generation	0.01
chemical	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.031
chemical	Inadequate/ unavailable SOP	spillage of hazardous material	0.016

accident paths with high impact. The existing FIG approach is modified to include these simple concepts and a comparative methodology is presented which can provide information about overall safety state of the organization and performance of RCSs.

5.2. Contribution to theory

In the existing studies of performance assessment of RCSs, the analysis was based primarily on human judgment and was subjective to researcher's opinion and expertise. The inferences presented in this work are completely supported by the data and reflect the actual safety state of the workplace. Moreover, the work contributes to studies in accident path analysis as well. As discussed earlier, there is lack of studies in the analysis of complete accident path, however, components of accident path have been analyzed well in literature. This study analyses complete accident path and from the results it can be observed that analysis of complete accident path provides more insights about the safety state of the organization which is otherwise not possible from only component's analysis.

5.3. Contribution to industry

The presented methodology not only provides information about the safety state of the plant but also assists safety experts in decision making. It will help organizations to monitor the performance RCSs from the actual workplace situation and also measure the effectiveness of the previous preventive decisions made by the organizations. Since process safety incidents are highly severe but less probable, therefore, the data set can be of relatively small size. In the case study, it can be observed that the proposed methodology is able to provide significant insights with small dataset as well. Therefore, the methodology is highly useful for the industries like steel manufacturing. Moreover, the work presented is generic in nature and can be applied to other industries as well.

5.4. Assessment of proposed method

The proposed method is simple to implement, easy to understand, and assists the management in decision making. Moreover, it requires no human involvement in analyzing the data. Huge amount of accident data can be analyzed using Python or any other programming language code. In the methodology development, instead of focusing only on causal patterns, the work has focused on complete accident path. Also, the methodology captures the change in patterns in response to managerial actions by including temporal effect. It helps in generating different safety driven inferences. The significant APs are considered based on both frequency and consequence score. One major highlight of the work is it provides a tool to assess the performance of both preventive and mitigating risk control systems. The methodology is validated through a real world case study and the results obtained from the

Table A3
Results of ESIG.

Hazardous Element	Accident Causing Mechanism	Accident/incident	Support
activity related hazard	damaged/degraded/poorly maintained equipment	Explosion	0.013
chemical	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.027
chemical	improper work by operator/worker	Overflow of hazardous material	0.013
chemical	improper work by operator/worker	splash	0.015
coke	improper work by operator/worker	spillage of hazardous material	0.021
dust and steam	Inadequate/ unavailable SOP	Fire/ flame generation	0.012
dust and steam	Inadequate/ unavailable SOP	splash	0.024
gas	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.013
gas	Inadequate/ unavailable SOP	degradation/breaking of equipment	0.016
gas	Inadequate/ unavailable SOP	Explosion	0.024
gas	Inadequate/ unavailable SOP	exposure to toxic gas/ high temperature	0.015
Hot metal/ steel/slag	damaged/degraded/poorly maintained equipment	spillage of hazardous material	0.027
Hot metal/ steel/slag	failure/malfunctioning of equipment	Fire/ flame generation	0.042
Hot metal/ steel/slag	failure/malfunctioning of equipment	Overflow of hazardous material	0.011
Hot metal/ steel/slag	failure/malfunctioning of equipment	spillage of hazardous material	0.011
Hot metal/ steel/slag	failure/malfunctioning of equipment	splash	0.034
Hot metal/ steel/slag	improper material/material quality/equipment design	degradation/breaking of equipment	0.012
Hot metal/ steel/slag	improper material/material quality/equipment design	Fire/ flame generation	0.016
Hot metal/ steel/slag	improper material/material quality/equipment design	spillage of hazardous material	0.039
Hot metal/ steel/slag	improper work by operator/worker	Explosion	0.019
Hot metal/ steel/slag	improper work by operator/worker	Fire/ flame generation	0.032
Hot metal/ steel/slag	improper work by operator/worker	spillage of hazardous material	0.039
Hot metal/ steel/slag	improper work by operator/worker	splash	0.056
Hot metal/ steel/slag	improper working environment	spillage of hazardous material	0.038
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Explosion	0.023
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Fire/ flame generation	0.051
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Leakage of hazardous material	0.015
Hot metal/ steel/slag	Inadequate/ unavailable SOP	spillage of hazardous material	0.016
Hot metal/ steel/slag	lack of communication/supervision	fall of material	0.013
Hot metal/ steel/slag	lack of communication/supervision	Fire/ flame generation	0.013
Hot metal/ steel/slag	lack of communication/supervision	spillage of hazardous material	0.028

Table A4
Results of High_impact_IG.

Hazardous Element	**Accident Causing Mechanism**	**Accident/Incident**	**Support**
chemical	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.028
chemical	improper work by operator/worker	Overflow of hazardous material	0.012
chemical	damaged/degraded/poorly maintained equipment	spillage of hazardous material	0.011
chemical	Inadequate/ unavailable SOP	spillage of hazardous material	0.013
dust and steam	failure/malfunctioning of equipment	Leakage of hazardous material	0.017
gas	improper work by operator/worker	Leakage of hazardous material	0.012
gas	failure/malfunctioning of equipment	Leakage of hazardous material	0.023
gas	damaged/degraded/poorly maintained equipment	Leakage of hazardous material	0.047
gas	improper work by operator/worker	Fire/ flame generation	0.011
gas	failure/malfunctioning of equipment	Fire/ flame generation	0.011
gas	damaged/degraded/poorly maintained equipment	exposure to toxic gas/ high temperature	0.017
gas	failure/malfunctioning of equipment	exposure to toxic gas/ high temperature	0.025
gas	improper work by operator/worker	exposure to toxic gas/ high temperature	0.015
gas	improper work by operator/worker	Explosion	0.014
Hot metal/ steel/slag	improper work by operator/worker	dashing/collision	0.013
Hot metal/ steel/slag	damaged/degraded/poorly maintained equipment	degradation/breaking of equipment	0.013
Hot metal/ steel/slag	improper work by operator/worker	degradation/breaking of equipment	0.015
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Explosion	0.018
Hot metal/ steel/slag	presence of unwanted/flammable material	Explosion	0.018
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Fire/ flame generation	0.011
Hot metal/ steel/slag	Inadequate/ unavailable SOP	Overflow of hazardous material	0.011
Hot metal/ steel/slag	damaged/degraded/poorly maintained equipment	spillage of hazardous material	0.019
Hot metal/ steel/slag	failure/malfunctioning of equipment	spillage of hazardous material	0.028
Hot metal/ steel/slag	improper work by operator/worker	spillage of hazardous material	0.09

data of case study were discussed with the experts of the studied plant. The experts found adherence of the results obtained from our methodology with their workplace observations. Moreover, they also found the results to be easy to comprehend and useful for taking further interventions.

However, the methodology does not provide any information related to the degree of effectiveness of RCSs. The methodology is based on the assumption that changes in trend of the pattern is based on the performance of RCSs. Moreover, it fails to distinguish between absence and ineffectiveness of RCSs, i.e., from results it can be comprehended that the RCS was either absent or ineffective, but the exact causal inference is difficult. Further, there could be other factors which may affect the trend of accident path pattern. Exploratory studies in this line based on the proposed methodology will bring further insights.

5.5. Future work

This work can further be enhanced by incorporating the parameters

like degree of effectiveness and ineffectiveness in frequent itemset generation. Moreover, a decision support framework can be proposed by incorporating different safety datasets such as formal audit or close call reports. This will help in comparing the results from different datasets so that an overall assessment of complete safety management can be done. The trend of proactive indicators may affect the trend of accident patterns. This effect can be analyzed and the relationship between proactive indicators and accident pattern can be quantified**.** This relationship can be exploited to reduce the accidents by controlling the proactive indicators. Interestingly, our proposed methodology, if applied, will also trigger plant personnel to detect plant specific factors that influence the RCS performance, which can be explored in future studies. Additionally, incorporation of control chart principles in the temporal analysis to identify out-of-control patterns and change points in the trend.

Acknowledgement

The work is funded by UAY project (Project Code: IITKGP_022). We acknowledge **Safety Analytics & Virtual Reality (SAVR) Laboratory** (www.savr.iitkgp.ac.in), Department of Industrial & Systems Engineering, IIT Kharagpur for experimental, computational and research facilities for this work. We would like to thank the management of the plant for providing relevant data and their support and cooperation during the study.

Declaration of Competing interests

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ress.2020.107041.

References

[1] Svenson O. The Accident Evolution and Barrier Function (AEB) Model Applied to Incident Analysis in the Processing Industries. Risk Anal 1991;11(3).
[2] Sklet S. Safety barriers : Definition, classification, and performance. J. Loss Prev. Process Ind. 2006;19:494–506.
[3] C. of chemical process safety CCPS, “Layer of Protection Analysis: Simplified Process Risk Assessment,” 2001.
[4] NIOSH, “Hierarchy of Controls,” 2016.
[5] “Safety and Health Recognition Programme, ” *World Steel Association*, 2019.
[6] Singh K, Maiti J, Dhalmahapatra K. Chain of events model for safety management : Data analytics approach. Saf. Sci. 2019;118:568–82. February.
[7] Groot A. Advanced Process Safety Barrier Management by Applying Proactive Incident Investigation to Failed or Impaired Barriers. SPE Int. Conf. Exhib. Heal. Safety, Secur. Environ. Soc. Responsib. Soc. Pet. Eng. 2016.
[8] Tayab MR, Villapil S, Kashwani G. Barrier Analysis & Strengthening to Eliminate Road Traffic Accidents in Oil Fields. SPE Int. Conf. Exhib. Heal. Safety, Secur. Environ. Soc. Responsib. Soc. Pet. Eng. 2018.
[9] Landucci G, Argenti F, Tugnoli A, Cozzani V. Quantitative assessment of safety barrier performance in the prevention of domino scenarios triggered by fire. Reliab. Eng. Syst. Saf. 2015;143:30–43.
[10] Landucci G, Argenti F, Spadoni G, Cozzani V. Domino effect frequency assessment: The role of safety barriers. J. Loss Prev. Process Ind. 2016;44:706–17.
[11] Bucelli M, Landucci G, Haugen S, Paltrinieri N, Cozzani V. Assessment of safety barriers for the prevention of cascading events in oil and gas offshore installations operating in harsh environment. Ocean Eng 2018;158:171–85. April.
[12] Janssens J, Talarico L, Reniers G, Sörensen K. A decision model to allocate protective safety barriers and mitigate domino effects. Reliab. Eng. Syst. Saf. 2015;143:44–52.
[13] Khakzad N, Landucci G, Reniers G. Application of dynamic Bayesian network to performance assessment of fire protection systems during domino effects. Reliab. Eng. Syst. Saf. 2017;167:232–47.
[14] Landucci G, Necci A, Antonioni G, Argenti F, Cozzani V. Risk assessment of mitigated domino scenarios in process facilities. Reliab. Eng. Syst. Saf. August 2016;160:37–53. 2017.
[15] Zhen X, Vinnem JE, Peng C, Huang Y. Quantitative risk modelling of maintenance work on major offshore process equipment. J. Loss Prev. Process Ind. 2018;56:430–43.
[16] Wang X, Huang X, Luo Y, Pei J, Xu M. Improving Workplace Hazard Identification Performance Using Data Mining. J. Constr. Eng. Manag. 2018;144(8).
[17] Ikeagwuani UM, John GA. Safety in maritime oil sector: Content analysis of machinery space fire hazards. Saf. Sci. 2013;51(1):347–53.
[18] Nenonen N. Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. Appl. Ergon. 2013;44(2):215–24.
[19] Silva JF, Jacinto C. Finding occupational accident patterns in the extractive industry using a systematic data mining approach. Reliab. Eng. Syst. Saf. 2012;108:108–22.
[20] Mirabadi A, Sharifian S. Application of association rules in Iranian Railways (RAI) accident data analysis. Saf. Sci. 2010;48(10):1427–35.
[21] Sjöblom O. Data Mining in Promoting Aviation Safety Management. In International Conference on Well-Being in the Information Society. 2014. p. 186–93.
[22] Mistikoglu G, Gerek IH, Erdis E, Mumtaz Usmen PE, Cakan H, Kazan EE. Decision tree analysis of construction fall accidents involving roofers. Expert Syst. Appl. 2015;42(4):2256–63.
[23] Amiri M, Ardeshir A, Zarandi MHF, Soltanaghaei E. Pattern extraction for high-risk accidents in the construction industry : a data-mining approach. Int. J. Inj. Contr. Saf. Promot. 2016;23(3):264–76.
[24] Dhalmahapatra K, Singh K, Jain Y, Maiti J. Exploring Causes of Crane Accidents from Incident Reports Using Decision Tree. Information and Communication Technology for Intelligent Systems.mart Innovation, Systems and Technologies. 2019.
[25] Verma A, Das Khan S, Maiti J, Krishna OB. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Saf. Sci. 2014;70:89–98.
[26] Dhalmahapatra K, Shingade R, Mahajan H, Verma A. Decision support system for safety improvement : An approach using multiple correspondence analysis, t-SNE algorithm and K-means clustering. Comput. Ind. Eng. 2019;128:277–89.
[27] Verma A, Maiti J, Gaikwad VN. A preliminary analysis of incident investigation reports of an integrated steel plant: some reflection. Int. J. Inj. Contr. Saf. Promot. 2017;25(2):180–94.
[28] Wheelwright S, Makridakis S. Forecasting Methods for Managers 1974.
[29] https://oshwiki.eu/wiki/Accidents_and_incidents, “Accidents and incidents”.
[30] Agarwal BR, Srikant R, Ahmad MA. Fast Algorithms for Mining Association Rules. In Proc. of the 20th VLDB Conference. 1994. p. 487–99.
[31] Zhang W, Liao H, Zhao N. Research on the FP Growth Algorithm about Association Rule Mining. 2008 International Seminar on Business and Information Management. 2008. p. 315–8.