Modelling the Relationships Between the Barriers to Implementing Machine Learning for Accident Analysis: the Indian Oil Industry

Abstract

Employees in the Indian oil industry operate in highly complex and hazardous environments. It is an industry that has been overshadowed by a history of catastrophic accidents that has affected both society and nature. Machine learning (ML) techniques are considered important in enabling an efficient and effective response to analyzing accident data and providing opportunities for lessons to be learned. Numourous studies have reported various barriers to the implementation of ML techniques in the Indian oil industry. Yet, there is limited knowledge about the weighting of importance of these barriers and the relationship between them. The purpose of this two-part study is to identify and rank the 10 most reported barriers to the implementation of ML in accident data analysis in the context of the Indian oil industry, and to analyse the relationships between these barriers. This is the first study to rank and analyse the relationship of these barriers using a number of advanced analytical tools, as well as engaging with experts in the oil industry. The findings provide sobering implications for accident investigators, and practitioners in the global oil industry. The findings also open new opportunities for research in the areas of ML and accident data analysis.

Keywords: Machine learning; Delphi; DEMATEL; TOPSIS; COPRAS; Oil industry

1. Introduction

Effective accident prevention requires organizations to have the ability to analyse large, complex datasets to identify the cause of the accident (Akbari & Do, 2021; Qureshi *et al*., 2020). For example, the oil industry is considered a high-risk industry due to the hazardous nature of chemicals and complex operating conditions (Nolan, 2019; Perrow, 1999).

Specifically, accidents that occur during the processing and transportation of chemicals are frequent, and catastrophic (Khan & Abbasi, 1999). Historically, accident data analysis was overshadowed by delays due to lack of skilled workforce, human bias, and time-consuming (Das *et al*., 2021; Cameron *et al*., 2017). Other related issues include poor decision making by management teams, and the inability to effectively analyze large, complex and varied datasets (Hovden *et al*., 2010). More recently, oil companies are increasingly investing in emerging technologies, such as artificial intelligence, and big data analytics (Alsaadoun, 2019) to generate meaningful insights and learn from previous accidents (Suh, 2021; Ranjan & Foropon, 2021). ML techniques, a subset of AI applications, have shown potential for extracting valuable information from large complex accident datasets (Madeira *et al*., 2021; Badri *et al*., 2018). ML techniques outperform traditional accident data analysis methods in terms of reducing the amount of time, money, and effort required to analyse the data, as well as handling a variety of data issues such as class imbalance (Das *et al*., 2021; Sarkar *et al*., 2019), punctuation correction, outlier removal, sentence correction, and missing values imputation (Tan *et al*., 2020; Kumar & Toshniwal, 2015).

Recent studies have shown that ML can be used to identify the pattern of events (Suh, 2021), find causative factors (Yazdi *et al*., 2020), solve complex and dynamic problems (Sarkar *et al*., 2019), predict failure events (Pramanik *et al*., 2021; Singh *et al*., 2019) provide alternative solutions (Ao *et al*., 2019). Analyzing accident data enables organizations to better understand the various set of accident attributes (Kumar & Toshniwal, 2015; Wang *et al*., 2018), as well as design and implement preventative measures (e.g., training programmes) to reduce the hazards and risks that cause accidents (Sarkar *et al*., 2020; Wiegmann & Shappell, 2017). India is one of the world's largest oil importers (IBEF, 2021) and has a history of catastrophic oil accidents, including LPG tank explosion at Hindustan Petroleum,

Visakhapatnam (Wasewar, & Kumar, 2010), Jaipur oil depot fire (Abbasi *et al*., 2014), GAIL pipeline fire (Lakshmi & Kumar, 2015), Baghjan oil well fire, Assam (Dutta, 2020). Due to the rate of serious accidents in the oil industry, analysing accident data has become an important activity for oil companies.

Recent studies highlight the value of using ML techniques in exploration, operations, and maintenance activities. Despite oil companies being quick to adopt emerging technologies (Koroteev & Tekic, 2021), the application of ML techniques in accident data analyzes has not been successfully adopted in the oil industry (Pandey *et al*., 2021). It is suggested that an analysis of the barriers to the adoption of ML could support companies to develop more effective accident prevention strategies (Misuri *et al*., 2021). This study aims to address this gap by answering the following research questions (RQ):

RQ.1 What are the barriers to the implementation of ML techniques in accident data analysis in the Indian oil industry?

RQ.2 What is the relationship between these barriers?

The remainder of this paper is structured as follows. First, a systematic literature review of the reported barriers to ML adoption is presented. Next, the research methodology and analytical techniques used in this study are provided. Then, findings and analysis are presented. This is followed by a discussion, implications, and limitations. The paper ends with a conclusion.

2. Systematic Review of Barriers to ML Implementation

This section summarizes the key challenges that were identified using the PRISMA literature review process of ML in the context of the Indian oil industry.

Data collection and inconsistent data formats: Following an accident, an investigation team analyse multiple data sources and prepare a report that is then sent to concerned stakeholders (i.e., internal departments, government agencies, regulatory bodies). A major flaw with this process is the use of different data taxonomies which have inconsistent data formats (Kaisler *et al*., 2013; Kim *et al*., 2003) , as well as curating and validating data (Ansaldi *et al*., 2021). Further, the reliance on traditional accident analysis techniques are vulnerable to human bias, human error, due to time constraints and not reporting minor incidents (Janssen *et al*., 2017; Hovden *et al*., 2010) resulting in the loss of critical information (Ahmed *et al*., 2019).

Human skills: A study by the World Economic Forum, (2018) revealed that 36% of respondents reported that a shortage of skilled labour was the top barrier for implementing data analytics in the domain. Subsequent studies also report that the lack of human skills a key barrier to implementing ML approaches (Moueddene *et al*., 2021; Guo *et al*., 2020; Paltrinieri *et al*., 2019; Angrave *et al*., 2016) as it requires advanced statistical skills; understanding hardware and software compatibility, and data architecture (Sarkar *et al*., 2017).

Data privacy, security and access: Access to accident datasets has been reported as a barrier for implementing ML to analyze accident data and develop predictive models (George & Renjith, 2021; Dixit *et al*., 2021). Data privacy and security issues have also been reported in numerous studies (Ansaldi *et al*., 2021; Khatri *et al*., 2020; Wang *et al*., 2018). Due to these issues, oil companies are highly susceptible to cyber-attacks (Whitworth & Suthaharan, 2014; Agrafiotis *et al*., 2018).

Limited understanding in ML techniques: Although there is awareness about the benefits of ML, the lack of understanding about how to implement and utilize ML techniques remains a barrier (Moueddene *et al*., 2021; Qazi *et al*., 2013). A lack of tailored ML training programs is attributed to a lack of understanding in this context (Dogruyol & Sekeroglu, 2019; Jidiga &

Sammulal, 2013). While the algorithms may be faster in terms of calculation time (analysis), ML techniques require independent training, inference, and cooperation activities (Yokoyama, 2019).

Operations management: As each ML technique has a different set of accuracy, precision, scoring, users must know how to select relevant measures for their models (Powers, 2020; Gharib & Bondavalli, 2019). Some of the operational issues identified in analyzing the safety data are automating the ML processes, tuning the data per the model requirement, scaling up the models, and measuring the performance (Paltrinieri *et al*., 2019; Rezapour & Ksaibati, 2021). The lack of tailored guidelines on how to embed ML techniques within operations is inhibiting its implementation (Agrafiotis *et al*., 2018; Jharkharia & Shankar, 2005).

IT and data infrastructure: Although the Indian oil industry has previously implemented technologies to reduce accidents in its operations and maintenance activities (Nianyin *et al*., 2021; Wanasinghe *et al*., 2020; Selcuk, 2016). There are concerns that basic requirements (e.g., data storage, servers, parallel computing, high-speed connectivity) are impeding the successful implementation of ML (Kumar & Goudar, 2012; Lin & Chen, 2012). Embedding ML techniques would enhance infrastructure, improve communication, reduce reliance on human efforts, and enable real-time time data analysis (Gohel *et al*., 2020; Ribes & Poth, 2014).

Resistance to change: The implementation of new ML initiatives is frequently blocked by organizational resistance to change (Ansaldi *et al*., 2021; Scholkmann, 2021; Schuetz & Venkatesh, 2020; Côrte-Real *et al*., 2019). Organizational resistance to change occurs when organizations lack readiness and willingness to change their current technologies and underlying processes (Dennehy *et al*., 2021; Dubey *et al*., 2015; (Raut *et al*., 2021; Alharthi *et al*., 2017). As a result, the implementation of ML techniques in the accident analysis process is being slowed due to a resistance to change (Mannering & Bhat, 2014) .

Lack of management commitment: The role of top management in driving the implementation of ML techniques is critical to its successful implementation (Mannering & Bhat, 2014; Yang & Wu, 2021; Müller *et al*., 2016). The empirical findings in this study, based on insights from industry experts revealed that companies are initially planning to implement ML initiatives in the maintenance activities because they presume that they can save huge costs by optimising maintenance activities. IOCL has already started implementing ML approaches in two oil processing plants in Northern India. During the data collection, a contradictory finding was observed that organizations' top management is committed to spending budgets on new technologies, but the driving force to implement these technologies is missing.

Lack of trust in ML techniques: Employees' lack of awareness always limits the research capabilities of the firms (Ajimoko, 2018; Ritala *et al*., 2015). During the Delphi process, we have observed that there are no plans for developing the data analytics skills of the employees. Organisations might hire experienced staff to draw extracts from the data, but hiring external agencies brings new issues such as data sharing, security, and mis-utilisation (Hausladen & Schosser, 2020). Lack of trust is an un-dimensional barrier that might create many obstacles for the organization in the long run (Kwon *et al*., 2014).

Financial constraints: Despite high profits, management are less keen on budg*et al*location to non-profit areas while prioritising investment decisions (Russom, 2013). In some cases, safety has never been given equal priority compared to other business processes (Fyffe *et al*., 2016). Organizations are slow to allocate the necessary budget for the recruitment of domains experts such as data scientists and ML engineers (Yang & Wu, 2021). There are many other costs associated with the implementation of ML including specialized training, upgrading IT infrastructure and software, and management of big data (Davenport, 2014).

3. Research methodology

3.1 Prisma technique

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) technique was used to conduct the review of the literature to identify studies that focused on the implementation of ML in the context of accident data analysis. PRISMA is suited to this study as it is a widely used method for conducting systematic literature reviews (Pasayat *et al*., 2020). The databases used in this study include Scopus, Web of Science, Science Direct, and EBSCO are they are the most widely used databases in academic research.

The keywords including but not limited to 'data analytics', 'accident data', 'incident data', ‘machine learning, 'accident analysis', 'injury analysis' and 'adoption', 'challenge', 'barrier' are used in different combinations (Boolean operations) to find the relevant papers. We have considered only full research articles and conference articles for the analysis. In the initial search total of 239 papers were collected. In the filtration process, 59 duplicates are eliminated from the analysis. After reading the title, abstract, and keywords of the collected papers to determine whether they aligned with the research topic, 17 were removed, not meeting the criteria. Then the second level of filtration was performed to see if the main content of those papers met the criteria and matched the keywords listed above. After this step, 107 articles were included for further categorizing. The final 107 articles chosen based on the bibliographic research are carefully examined to identify the potential challenges of applying machine learning to accident data analysis. After a detailed examination of findings in these selected papers, ten influencing barriers were identified. Figure 1 illustrates the process of the PRISMA method followed.

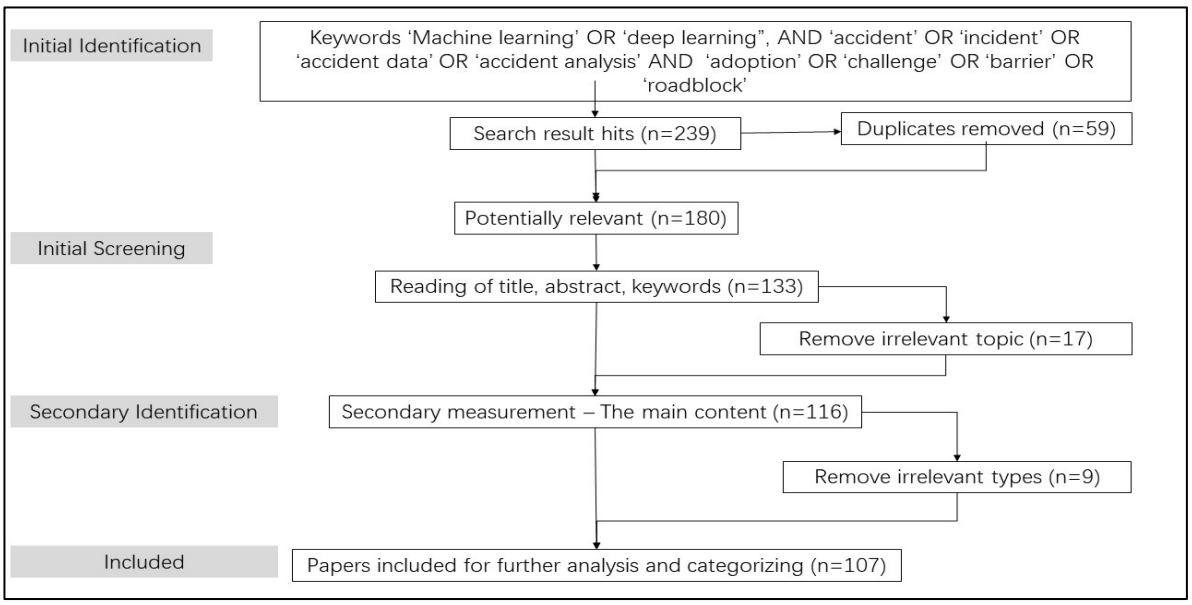

Figure 1. PRISMA process for selecting relevant literature for review

Figure 2 illustrates the complete methodology used in this research. In the first phase, we have used the PRISMA method for finding relevant papers using a set of keywords (see Figure 1). After a careful review of the contents of the final 107 articles, ten barriers were identified which hinder the implementation of ML in accident data analysis.

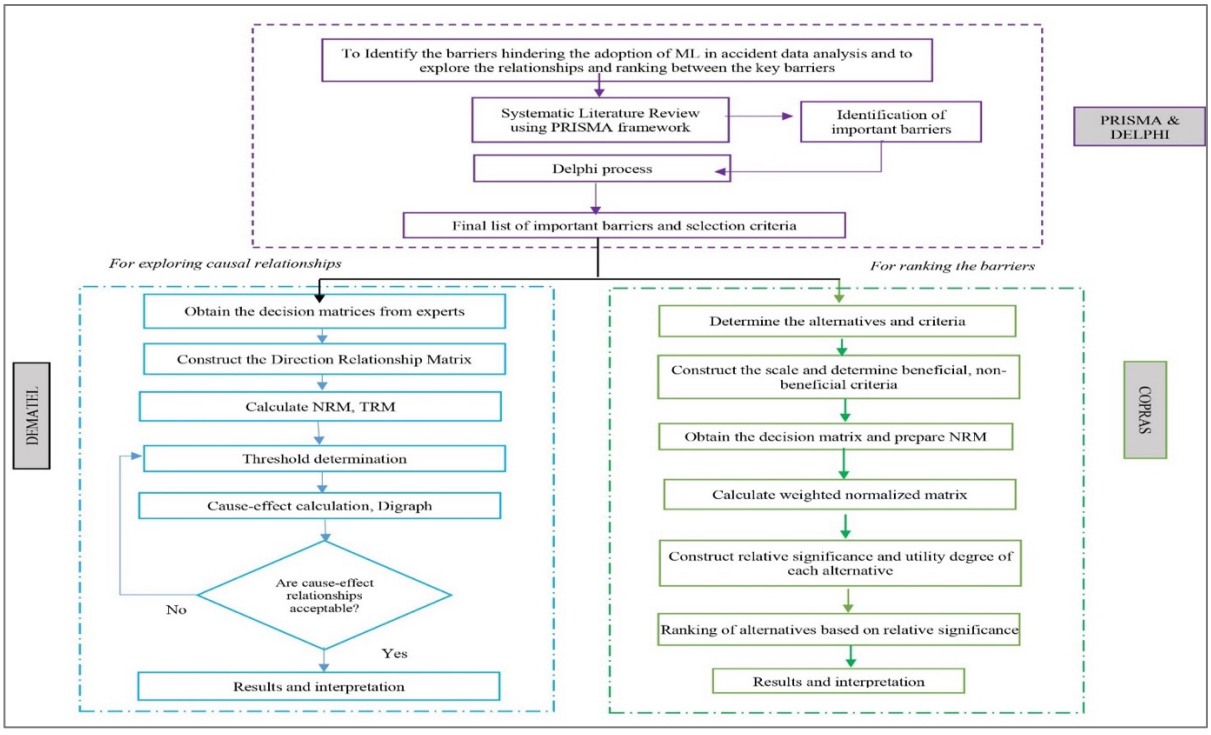

Figure 2. Research framework (Delphi-DEMATEL-COPRAS)

3.2 Delphi interview process

The Delphi method was used to acquire opinions on the barriers listed previously from 10 experts in the oil industry (see Table I). The value of using the Delphi method is to achieve consensus from domain experts and for collecting complex decision-making problems (Amirghodsi *et al*., 2020; Joshi *et al*., 2011) . Semi-structured interviews were conducted with the domain experts to assess the relevance of the barriers to the implementation of ML in accident data analysis.

****Insert Table I about here****

During the interview process, the experts suggested merging a barrier named "Trust in ML techniques" with "lack of awareness in employees" since they are covering similar aspects. A new barrier called "lack of information sharing between the firms" emerged from this stage of the study. Thus, ten barriers were finalised after the Delphi process. Table II lists the changes in the titles of barriers before and after the Delphi process.

****Insert Table II about here****

3.3 The DEMATEL method

A Decision-Making Trial and Evaluation Laboratory (DEMATEL) method was used to evaluate the importance of each barrier and the interrelationships between the barriers. The DEMATEL technique is useful for visualising the structure of complicated interdependent relationships (Amirghodsi *et al*., 2020; Joshi *et al*., 2011). DEMATEL helps to determine the causal relationship between given parameters/variables. DEMATEL confirms interdependence among factors, aids in developing a map to reflect relative relationships within them and can be used to investigate and solve complicated and intertwined problems (Amirghodsi *et al*.,

2020). This method uses matrices to convert interdependency relationships into a cause-and-effect group, and an impact relation diagram to find the critical factors of a complex structure system (Digraph). Network Relationship Matrix (NRM) helps to signify the interrelationships between the barriers (Amirghodsi *et al*., 2020). A five point rating scale ranging from 1 to 5 has been used to collect the data, where *one* indicates no influence, *five* indicate high influence.

3.4 The COPRAS method

The final ranking of the barriers was conducted using the Complex Proportional Assessment (COPRAS) method proposed by Zavadskas *et al*., (1994). This method is useful to evaluate the maximising and minimising index values, and the impact of maximising and minimising attribute indexes on the evaluation of the results is looked at separately (Organ & Yalçın, 2016). Three criteria, namely, time, cost, relative importance, were given by experts for ranking the barriers. Since ranking techniques require weights for criteria, the experts have given cost (C_1=weight =0.5), time (C_2=weight=03), relative importance (C_3=weight =0.2). To avoid any bias in the criteria weights, we also calculated the ranks using equal weights (C_1=C_2=C_3=0.333) for all three criteria. The data collection sheet for ranking the barriers is listed in table III. As part of this exploratory phase of the study, three different criteria were considered based on expert advice to analyse the hierarchy of the proposed barriers.

****INSERT TABLE III ABOUT HERE****

4 Findings and Analysis

The findings are presented as per the three steps required to analyze the relationship between the barriers to implementing ML techniques in the analysis of accident data in the Indian oil industry.

Step 1: Formation of Direct Relationship Matrix (DRM)

Using the matrix provided in equation.1, the causality effect of one variable (barrier) on the other variable is calculated.

$$A_P = \begin{bmatrix} 0 & a12 & a13 & a14 & a15 & \dots & a1n \\ a21 & 0 & a23 & a24 & a25 & \dots & a2n \\ a31 & a32 & 0 & a34 & a35 & \dots & a3n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a(n-1)1 & a(n-2)2 & a(n-3)3 & \dots & \dots & 0 & a(n-1)n \\ an1 & an2 & an3 & \dots & \dots & an(n-1) & 0 \end{bmatrix} \quad \dots\dots(1)$$

Here A_p sample format is used for data collection, N represents the no. of barriers, and no. of respondents are represented by P. For example, to determine the influence between "human skills (B1)" on "Inertia to change to new systems (B2)," all the ten experts' inputs from the data collection table are extracted, and the values are 3,4,3,3,4,3,3,4,2,3 respectively. All scores of experts' was obtained and we took the average values, we get $\frac{3+4+3+3+4+3+3+4+2+3}{10} = \frac{32}{10} = 3.2$, and this value has been confined to a cell (1, 2) of direct relation matrix (DRM) represented in table.IV. The value 2.3 indicates a significant effect of human skills on inertia to change to new systems. Another computation gives the influence between "Data collection and inconsistent data formats (B4)" and "Human skills (B1)", we get, $\frac{3+0+0+3+3+3+1+0+0+0}{10} = \frac{13}{10} = 1.3$, specified in cell (4, 1) of DRM, indicates a low or moderate relationship between the variables. As the same barriers cannot affect themselves, all diagonal elements in the DRM are set to zero. All cells of DRM are calculated using the above procedure, and the results are entered in Table IV.

****INSERT TABLE IV ABOUT HERE****

Step 2: Formation of Normalised Direct Relation Matrix (NRM)

The data from the original direct-relation matrix is used to create the normalised direct-relation matrix "M2" (see table.V). The sum of all cell values in the corresponding row is divided by each cell. For example, see the matrix in NRM in row.1 the summation of all values 0, 3.2, 1.7, 1.6, 3.1, 3.4, 1.2, 3.8, 2.9, 1.6 equals 22.5. Each cell in row1 of table.IV is divided by this corresponding row total of 22.5. This step is repeated for all remaining with their respective row summation values, and then we obtain normalized values. The NRM is represented in Table V.

****INSERT TABLE V ABOUT HERE****

Step 3: Calculation of Total Relation Matrix (TRM)

The total-relation matrix (TRM) is calculated using the following equation.

$$T = X\,(I - X)^{-1} \qquad \text{……..(2)}$$

Where T stands for TRM, X indiacates NRM, and *I* used to denote the identity matrix. Further, using the procedure mentioned in Chen *et al*. (2021), threshold values (α), sum fo the values in a row (D), and Sum of values in columns (R) are calculated. Equations (3), (4), (5) are used to calculate the values D, R, α values

$$D = \left[d_{i_i}\right]_{n*1} = \left[\sum_{j=1}^{n} d_{i_j}\right]_{n*1} \qquad \text{…….. (3)}$$

$$R = [r]_{1*n} = \left[\sum_{i=1}^{n} r_{i_j}\right]_{1*n} \qquad \text{……..(4)}$$

$$\alpha = \frac{\sum_{j=1}^{n} \cdot \sum_{i=1}^{n} r_{i_j}}{n^2} \qquad \text{……..(5)}$$

Where i, j are values for respected rows and columns in the TRM, and n is the number of barriers in the study. The threshold value (α) is used to determine important and insignificant barriers.The α score is obtained as 0.2319, and the cell values in the TRM matrix that are smaller than the threshold α value (0.2319) are replaced with the value "zero (0)" (Table VI).

****INSERT TABLE VI ABOUT HERE****

The values from table.VI are used to draw the Network relationship map (di-graph) as illustrated in Figure 3. The arrows in Figure 3 indicate the relationship between the barriers. For example, the "Awareness of ML models" has many incoming arrows, which means other barriers highly influence it. Similarly, the "human skills" barrier has many outgoing arrows from it, which means it is influencing other barriers. The findings of R and D corroborate the degree of relationship effect among each critical challenge. While the sum D+R represents the significance of a particular barrier, the difference D-R shows the net influence of the given barrier. For instance, computations of D+R and D-R are for challenge B1; the D score is 2.3437, and the R score is 0.8723, so adding them together D+R is 3.2161 whereas subtracting them D-R is 1.4714. The barriers having high D-R values will have high importance and are named as "cause' group or causal barriers, the barriers with low D-R scores are less significant, and they are influenced by cause group barriers and named as "effect" group barriers.

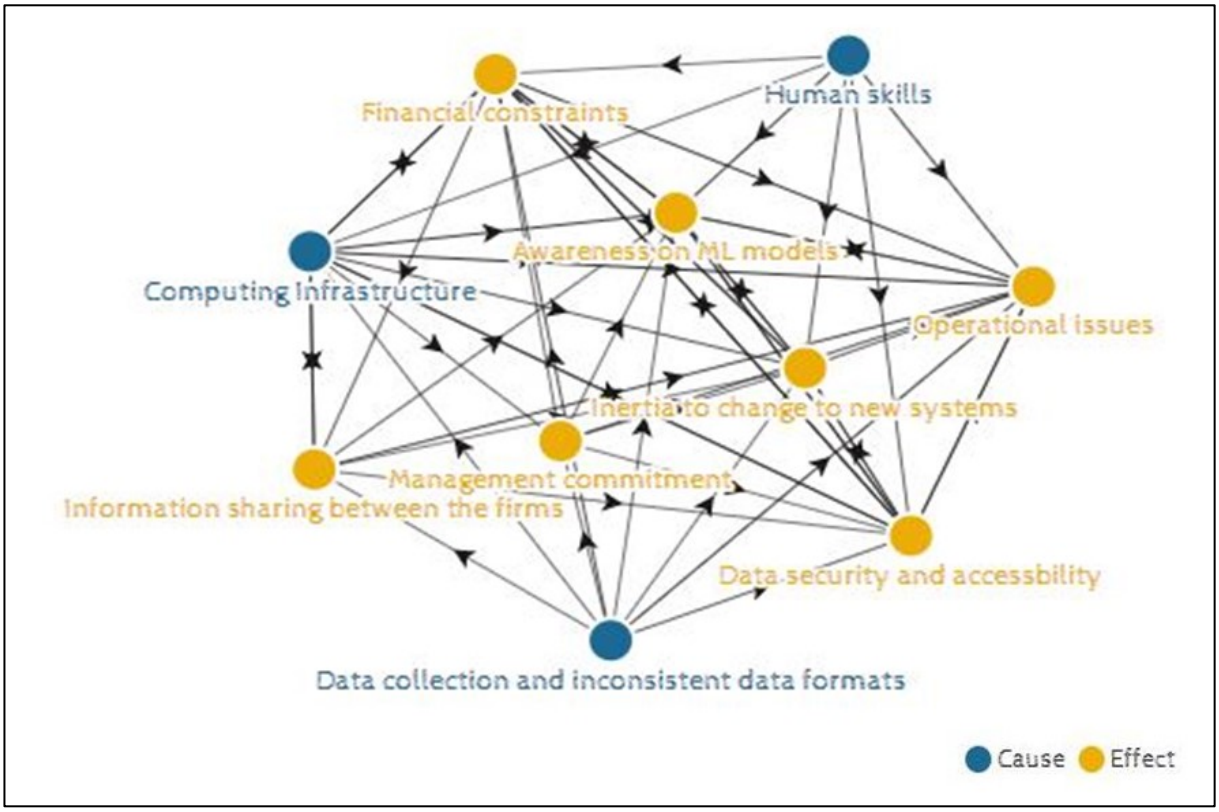

Figure 3. Directed graph for the barriers relationship

The final results obtained using the DEMTAL method are listed in Table VII. The challenges, Human skills (B1), Data collection and inconsistent data formats (B4), and computing infrastructure (B5), are found to follow under cause group. The remaining challenges, i.e., "Inertia to change to new systems (B2), Financial constraints (B3), Operational barriers (B6), Management commitment (B7), Awareness of ML models (B8), Data security and accessibility (B9), and Information sharing between the firms (B10)" are classified as effect group. Although the effect group barriers have less dominance over the other barriers, they are highly influenced by "cause" group barriers.

****INSERT TABLE VII ABOUT HERE****

The most important barrier to implementing ML for accident data analysis is "human skill availability (B1)," with the greatest D-R score of 1.471, implying that B1 should emphasize the entire system ML implementation in accident data analysis. Our findings match with results of existing studies that mention the unavailability of ML domain experts as a major constraint for ML implementation initiatives (Maurice, 2021; Tixier *et al*., 2016). Furthermore, Table V reveals that B4 is the 2nd most important factor. Barrier B4 indicates a lack of data collection and inconsistency in accident reports' reporting formats, making it difficult for ML models to process the data. This result also has roots in the published literature (Jesmeen *et al*., 2018). With the second greatest D-R value, "lack of data collection and inconsistent data formats (B4)" substantially influences other barriers. Without structured data, it becomes difficult for ML models to extract useful insights. If the data is not structured or in a unified format, it consumes many manhours to make it structured. Also, the efficiency of ML models depends on the richness of the data. The "computing infrastructure (B5)" with a D-R value of 0.6636, which also holds in the context of ML implementation. The interconnection between the systems and speed of the internet and the availability of high-speed computers help to analyze accident data quickly and efficiently.

If the value of D-R is negative, the variables (barrier) belong to the impact group (effects) and, the cause group variables heavily influence them. The di-graph (see Figure 4) is an illustration of the cause-effect relationships between the barriers. Furthermore, in terms of notable effect degree, awareness of ML models (B8) and operational barriers (B6) is highly influenced by the cause group variables. Similarly, the remaining barriers—inertia to change to new systems (B2), data security, and accessibility (B9) have low D-R values, implying significantly low importance for implementing ML in accident data analysis. Furthermore, from the cause-effect diagram it can be observed that the barriers, management commitment (B7) and financial constraints (B3) are very close to the cause-effect distinguishing line.

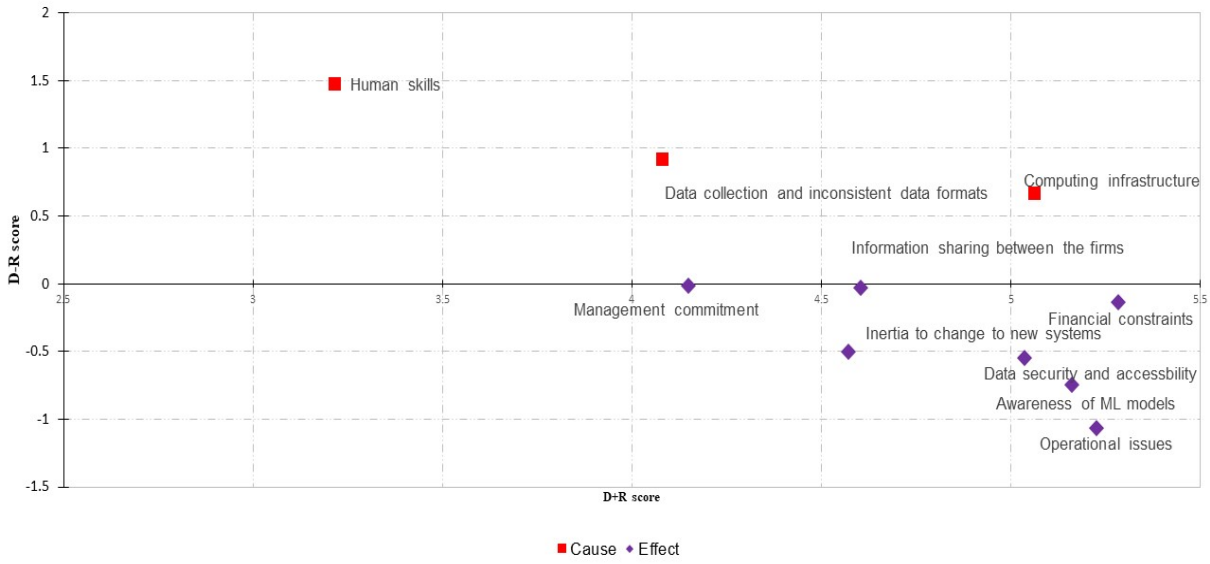

Figure 4. Cause-effect diagram

4.3 Weighting and ranking of barriers using COPRAS

The steps involved to rank different alternatives using the COPRAS method is as follows:

Step1: Calculation of normalized decision matrix (X_{ij})

$$R = X_{ij} = \frac{a_{ij}}{\sqrt{\sum_{k=1}^{m} a_{ij}^2}} \qquad \text{…….(9)}$$

Where a_{ij} is the performance of value of alternative A_i based on criterion C_j. Where a_{ij} is a decision matrix.

****Insert Table VIII about here****

Step 2: Calculation of weighted normalize decision matrix (W_{ij}):

$$W_{ij} = w_j * x_{ij} \quad \text{…… (10)}$$

****Insert table.IX about here****

Step 3: Calculation of S^+ and S^-:

S^+ and S^- are the sum of the weighted normalized values computed (table.IX, X) from the benefit and non-benefit criteria. In this study C_1, C_2 criterion is non-beneficial type, and C_3 is beneficial.

$$S^+_i = \sum_{j=1}^{n} w_{ij} \quad (i=1,2,3\ldots m) \quad \text{……..…….(11)}$$

$$S^-_i = \sum_{j=1}^{n} w_{ij} \quad (i=1,2,3\ldots m) \quad \text{……(12)}$$

Where W_{ij} is the weighted normalised elements

Step 4: Calculation of relative weight of each alternative Q_i:

$$Q_i = (S_i^+) + \left(\frac{\sum_{i=1}^{m} S_i^-}{S_i^- \sum_{i=1}^{m} \frac{1}{S_i^-}}\right) \quad \text{..……..(13)}$$

Step 5: Priority order determination (P_{ri}):

$$P_{ri} = \frac{Q_i}{\max Q_i} \quad \text{……..……..(14)}$$

****Insert Table X about here****

****Insert Table XI about here****

The highest value of importance (K) is considered the best alternative. From Table XI and XII we can infer that the barrier B5- computing infrastructure get first rank (for two different weight scenarios), B1- human skills, B9-data security, and accessibility received top 2 and 3 ranks with different weights. During the analysis, we have not noticed a change of rank between the top 4 barriers w.r.t to different weights, which signifies the change of criteria weights have not shown any significant effect on the ranking of the barriers. Therefore, organizations and researchers who want to start implementing the ML in accident data analysis should first focus on B1, B5, B9, and B4 barriers, and the other barriers can be given less priority than these. Complementary to COPRAS, the MOORA technique was used to cross-check the results of any bais with the proposed technique. Table.X presents the summary of the ranking of alternatives based on COPRAS, MOORA under two-weight scenarios.

****Insert Table XII about here****

It was observed that the change of criteria weights and change of ranking technique does not have influence an overall ranking of barriers. "computing infrastructure (B5)", lack of "human skills (B1)", "data security and accessibility (B9)" are the prime dominant factors affecting the implementation of ML for accident data analysis.

5. Discussion, Limitations, and Future Research

Machine learning techniques have proven effective in analysing huge data to draw insightful patterns and effective data analysis in less time. Despite having access to a reservoir of financial resources, the Indian oil industry is facing concerning issues to the successful implemention of ML in accident data analysis. Ten key barriers were identified using a systematic literature review process (PRISMA) and engagement with industry experts using the Delphi method. The DEMATEL method is used to analyze the relationship between the barriers and the cause-

effect relationships using visualisations and statistical evidence. To obtain ranking among the barriers, we have used COPRAS, MOORA techniques. The ranking results revealed that "B5-computing infrastructure", "B1-Human skills", "B9-Data security and accessibility" are the top barriers that require companies' attention while planning for implementation of ML for accident data analysis. We have used two scenarios with two criteria weights scenarios, but it was observed that change of criteria had affected the ranking of other barriers, not affecting the top four barriers. The organizations should focus on the identified causal barriers and the important paths in Di-graph to identify the interrelationships among key barriers. The ranking of barriers helps prioritize resources while planning to implement the ML techniques for accident data analysis.

Activities associated with the oil industry are hazardous and accidents occur due to unintentional activities in the chain of events. ML techniques have the capabilities in revealing patterns, interpolating association rules, and predicting the weak zones for accidents. However, the implementation of ML in accident data analysis in Indian oil companies remains stubbornly poor. This study do however, provide a comprehensive decision-making framework for identifying the important barriers and plots the relationship between the associated barriers. A hybrid three-phase methodology involving Delphi-DEMATEL-COPRAS was used to establish the relationship among the barriers using the data from industry experts. Researchers and industry professionals can use the study's findings to overcome challenges in leveraging the benefits of ML in accident data analysis.

As with all research, however, we acknowledge this study has a limitation, which also offers directions for future research. In this study, we have considered only the top ten barriers for the analysis. In the DEMATEL process, we have used a five-point scale adopted from Amirghodsi *et al*., (2020). Future studies in this domain could use a fuzzy scale that can help improve the

accuracy of results by removing the ambiguity in the responses. Future research could also identify additional criteria variables with different weight instances and the key factors hindering the implementation of ML techniques. Moreover, future research could formulate the decisions based on the causal diagram, statistics and compare their similarities and differences. In the future, the proposed methodology can be used with more barriers and a larger sample size for getting more crisp results. Despite these limitations, it provides direction for future research in the accident data analysis of the Indian oil industry, which has a poor accident safety record.

6. Conclusion

This study draws on many quantitative analytical techniques to study the relationships between the barriers to the implementation of ML in accident data analysis. The findings demonstrate the importance of understanding the relationship between barriers to ML implementation, as well as its weighting in terms of importance. In doing so, this study demonstrates that strategies for ML implementation in accident data analysis require consideration of not just the technical characteristics of ML, but also the real-world context in which such techniques are intended to be used. Concluding, regardless of the context of the ML initiative, organizations needs to balance the technical with the social aspects of ML.

References

Abbasi, M, Benhelal, E, and Ahmad, A. (2014), "Designing an Optimal safe layout for a fuel storage tanks farm: Case Study of Jaipur Oil Depot," *International Journal of Chemical, Molecular, Nuclear, Materials and Metallurgical Engineering*, vol. 8, pp. 147-155.

Agrafiotis, I., Nurse, J.R., Goldsmith, M., Creese, S. and Upton, D. (2018), "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate", *Journal of Cybersecurity*, Vol. 4 No. 1, available at:http://doi.org/10.1093/cybsec/tyy006.

Ahmed, A., Sadullah, A.F. and Yahya, A.S. (2019), "Errors in accident data, its types, causes and methods of rectification-analysis of the literature", *Accident Analysis & Prevention*, Vol. 130, pp. 3–21.

Ajimoko. O. J. (2018), "Considerations for the Adoption of Cloud-based Big Data Analytics in Small Business Enterprises," *Electronic Journal of Information Systems Evaluation*, vol. 21, pp. 63-79.

Akbari, M. and Do, T.N. (2021), "A systematic review of machine learning in Logistics and Supply Chain Management: Current trends and future directions", *Benchmarking: An International Journal*, Vol. 28 No. 10, pp. 2977–3005.

Alharthi, A., Krotov, V. and Bowman, M. (2017), "Addressing barriers to big data", *Business Horizons*, Vol. 60 No. 3, pp. 285–292.

Alsaadoun, O. (2019), "A cybersecurity prospective on industry 4.0: Enabler role of identity and access management", *International Petroleum Technology Conference*, available at:http://doi.org/10.2523/19072-ms.

Amirghodsi, S., Naeini, A.B. and Makui, A. (2020), "An integrated Delphi-DEMATEL-Electre method on Gray numbers to rank technology providers", *IEEE Transactions on Engineering Management*, pp. 1–17.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M. and Stuart, M. (2016), "HR and analytics: Why hr is set to fail the Big Data Challenge", *Human Resource Management Journal*, Vol. 26 No. 1, pp. 1–11.

Ansaldi, S.M., Agnello, P., Pirone, A. and Vallerotonda, M.R. (2021), "Near miss archive: A Challenge to share knowledge among inspectors and improve Seveso inspections", *Sustainability*, available at:http://doi.org/10.20944/preprints202106.0042.v1.

Ao, Y., Li, H., Zhu, L., Ali, S. and Yang, Z. (2019), "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling", *Journal of Petroleum Science and Engineering*, Vol. 174, pp. 776–789.

Badri, A., Boudreau-Trudel, B. and Souissi, A.S. (2018), "Occupational health and safety in the Industry 4.0 era: A cause for major concern?", *Safety Science*, Vol. 109, pp. 403–411.

Cai, L. and Zhu, Y. (2015), "The challenges of data quality and data quality assessment in the Big Data Era", *Data Science Journal*, Vol. 14, p. 2.

Cameron, I., Mannan, S., Németh, E., Park, S., Pasman, H., Rogers, W. and Seligmann, B. (2017), "Process hazard analysis, hazard identification and scenario definition: Are the conventional tools sufficient, or should and can we do much better?", *Process Safety and Environmental Protection*, Vol. 110, pp. 53–70.

Côrte-Real, N., Ruivo, P., Oliveira, T. and Popovič, A. (2019), "Unlocking the drivers of big data analytics value in firms", *Journal of Business Research*, Vol. 97, pp. 160–173.

Das, S., Datta, S., Zubaidi, H.A. and Obaid, I.A. (2021), "Applying interpretable machine learning to classify tree and utility pole related crash injury types", *IATSS Research*, available at:http://doi.org/10.1016/j.iatssr.2021.01.001.

Davenport, H. T. (2014), "How strategists use 'big data' to support internal business decisions, discovery and production", *Strategy & Leadership*, Vol. 42 No. 4, pp. 45–50.

Dennehy, D., Oredo, J., Spanaki, K., Despoudi, S. and Fitzgibbon, M. (2021), "Supply Chain Resilience in mindful humanitarian aid organizations: The role of Big Data Analytics", *International Journal of Operations & Production Management*, Vol. 41 No. 9, pp. 1417–1441.

Dixit, M., Deshmukh, S., Dongaonkar, M. and Jadhav, S. (2021), "Road accident analysis using Random Forest algorithm", *Recent Trends in Communication and Electronics*, pp. 502–506.

Dogruyol, K. and Sekeroglu, B. (2019), "Absenteeism prediction: A comparative study using Machine Learning Models", *Advances in Intelligent Systems and Computing*, pp. 728–734.

Dubey, R., Gunasekaran, A., Childe, S.J., Wamba, S.F. and Papadopoulos, T. (2015), "The impact of Big Data on world-class sustainable manufacturing", *The International Journal of Advanced Manufacturing Technology*, Vol. 84 No. 1-4, pp. 631–645.

Dutta, P.K. (2020), "Baghjan oil well fire in Assam still raging after five months, the longest in India", *India Today*, 4 November, available at: https://www.indiatoday.in/india/story/assam-baghjan-oil-well-fire-ngt-report-1737938-2020-11-04 (accessed 15 February 2022).

Fyffe, L., Krahn, S., Clarke, J., Kosson, D. and Hutton, J. (2016), "A preliminary analysis of key issues in chemical industry accident reports", *Safety Science*, Vol. 82, pp. 368–373.

George, P.G. and Renjith, V.R. (2021), "Evolution of safety and security risk assessment methodologies towards the use of Bayesian networks in Process Industries", *Process Safety and Environmental Protection*, Vol. 149, pp. 758–775.

Gharib, M. and Bondavalli, A. (2019), "On the evaluation measures for machine learning algorithms for safety-critical systems", *2019 15th European Dependable Computing Conference (EDCC)*, available at:http://doi.org/10.1109/edcc.2019.00035.

Gohel, H.A., Upadhyay, H., Lagos, L., Cooper, K. and Sanzetenea, A. (2020), "Predictive maintenance architecture development for nuclear infrastructure using machine learning", *Nuclear Engineering and Technology*, Vol. 52 No. 7, pp. 1436–1442.

Guo, S., Cui, J., Zhao, Y., Wang, Y., Ma, Y., Gao, W., Mao, G., *et al*. (2020), "Machine learning–based Operation Skills Assessment with Vascular Difficulty Index for vascular intervention surgery", *Medical & Biological Engineering & Computing*, Vol. 58 No. 8, pp. 1707–1721.

Hausladen, I. and Schosser, M. (2020), "Towards a maturity model for big data analytics in Airline Network Planning", *Journal of Air Transport Management*, Vol. 82, p. 101721.

Hovden, J., Albrechtsen, E. and Herrera, I.A. (2010), "Is there a need for new theories, models and approaches to occupational accident prevention?", *Safety Science*, Vol. 48 No. 8, pp. 950–956.

I.B.E.F. (2021), "Oil & Gas Industry in India", *IBEF*, 22 September, available at: https://www.ibef.org/industry/oil-gas-india.aspx (accessed 25 September 2021).

Janssen, M., Konopnicki, D., Snowdon, J.L. and Ojo, A. (2017), "Driving public sector innovation using big and open linked data (BOLD)", *Information Systems Frontiers*, Vol. 19 No. 2, pp. 189–195.

Jesmeen, Z.H.M., Hossen, J., Sayeed, S., Ho, C.K., K, T., Rahman, A. and Arif, E.M.H. (2018), "A survey on cleaning dirty data using machine learning paradigm for Big Data Analytics", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 10 No. 3, p. 1234.

Jharkharia, S. and Shankar, R. (2005), "It-enablement of Supply Chains: Understanding the barriers", *Journal of Enterprise Information Management*, Vol. 18 No. 1, pp. 11–27.

Joshi, R., Banwet, D.K. and Shankar, R. (2011), "A Delphi-AHP-Topsis based benchmarking framework for performance improvement of a cold chain", *Expert Systems with Applications*, Vol. 38 No. 8, pp. 10170–10182.

Kaisler, S., Armour, F., Espinosa, J.A. and Money, W. (2013), "Big data: Issues and challenges moving forward", *2013 46th Hawaii International Conference on System Sciences*, pp. 995–1004.

Khan, F.I. and Abbasi, S.A. (1999), "Major accidents in process industries and an analysis of causes and consequences", *Journal of Loss Prevention in the Process Industries*, Vol. 12 No. 5, pp. 361–378.

Khatri, S., Vachhani, H., Shah, S., Bhatia, J., Chaturvedi, M., Tanwar, S. and Kumar, N. (2020), "Machine learning models and techniques for VANET based traffic management: Implementation issues and challenges", *Peer-to-Peer Networking and Applications*, Vol. 14 No. 3, pp. 1778–1805.

Kim. W, Choi. B, Hong. A, Kim, and Lee.D. (2003), "A taxonomy of dirty data," *Data mining and knowledge discovery*, vol. 7, pp. 81-99.

Koroteev, D. and Tekic, Z. (2021), "Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future", *Energy and AI*, Vol. 3, p. 100041.

Kumar, S. and Goudar, R.H. (2012), "Cloud computing – research issues, challenges, architecture, platforms and applications: A survey", *International Journal of Future Computer and Communication*, pp. 356–360.

Kumar, S. and Toshniwal, D. (2015), "Analysing road accident data using association rule mining", *2015 International Conference on Computing, Communication and Security (ICCCS)*, pp. 1–6.

Kwon, O., Lee, N. and Shin, B. (2014), "Data quality management, data usage experience and acquisition intention of Big Data Analytics", *International Journal of Information Management*, Vol. 34 No. 3, pp. 387–394.

Lakshmi, M.R. and Kumar, V.D. (2015), "Anthropogenic hazard and disaster relief operations: A case study of gail pipeline blaze in East Godavari of A.P", *Procedia - Social and Behavioral Sciences*, Vol. 189, pp. 198–207.

Lin, A. and Chen, N.-C. (2012), "Cloud computing as an innovation: Percepetion, attitude, and adoption", *International Journal of Information Management*, Vol. 32 No. 6, pp. 533–540.

Madeira, T., Melício, R., Valério, D. and Santos, L. (2021), "Machine learning and natural language processing for prediction of human factors in aviation incident reports", *Aerospace*, Vol. 8 No. 2, p. 47.

Mannering, F.L. and Bhat, C.R. (2014), "Analytic methods in accident research: Methodological Frontier and Future Directions", *Analytic Methods in Accident Research*, Vol. 1, pp. 1–22.

Maurice, S. (2021), "Overcoming challenges to ML Adoption", *Transactional Machine Learning with Data Streams and AutoML*, pp. 61–76.

Misuri, A., Landucci, G. and Cozzani, V. (2021), "Assessment of safety BARRIER performance in the mitigation of Domino scenarios caused by Natech events", *Reliability Engineering & System Safety*, Vol. 205, p. 107278.

Moktadir, M.A., Ali, S.M., Rajesh, R. and Paul, S.K. (2018), "Modeling the interrelationships among barriers to sustainable supply chain management in Leather Industry", *Journal of Cleaner Production*, Vol. 181, pp. 631–651.

Moueddene, K., Coppola, M., Wauters, P., Ivanova, M., Paquette, J. and Ansaloni, V. (2021), "Expected skills needs for the future of work - deloitte us", *Https://www2.Deloitte.com/*, available at: https://www2.deloitte.com/content/dam/insights/us/articles/22923_expected-skills-needs-for-the-future-of-work/DI_Expected-skills-needs-for-the-future-of-work.pdf (accessed 15 February 2022).

MPNG. (2021), "Home: Ministry of petroleum and natural Gas: Government of India", *Home | Ministry of Petroleum and Natural Gas | Government of India*, available at: https://mopng.gov.in/en (accessed 25 September 2021).

Nianyin, L., Chao, W., Suiwang, Z., Jiajie, Y., Kang, J., Wang, Y. and Yinhong, D. (2021), "Recent advances in waterless fracturing technology for the petroleum industry: An overview", *Journal of Natural Gas Science and Engineering*, Vol. 92, p. 103999.

Nolan, D.P. (2019), *Handbook of Fire and Explosion Protection Engineering Principles for Oil, Gas, Chemical, and Related Facilities*, Gulf Professional Publishing is an imprint of Elsevier.

Organ. A, and Yalçın. E. (2016), "Performance evaluation of research assistants by COPRAS method," *European Scientific Journal*, vol. 12, pp. 102-109.

Paltrinieri, N., Comfort, L. and Reniers, G. (2019), "Learning about risk: Machine Learning for Risk Assessment", *Safety Science*, Vol. 118, pp. 475–486.

Pandey, R., Gautam, V., Pal, R., Bandhey, H., Dhingra, L.S., Misra, V., Sharma, H., *et al*. (2021), "A machine learning application for raising wash awareness in the times of COVID-19 pandemic", *ArXiv*, pp. 1–14.

Pasayat, A.K., Bhowmick, B. and Roy, R. (2020), "Factors responsible for the success of a start-up: A meta-analytic approach", *IEEE Transactions on Engineering Management*, pp. 1–11.

Perrow, C. (1999), *Normal Accidents: Living with High-Risk Technologies*, Princeton University Press, Princeton, NJ.

Powers. D. M. (2020), "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint* arXiv:2010.16061, pp. 1-27.

Pramanik, A., Sarkar, S. and Maiti, J. (2021), "A real-time video surveillance system for traffic pre-events detection", *Accident Analysis & Prevention*, Vol. 154, p. 106019.

Qazi, Z.A., Lee, J., Jin, T., Bellala, G., Arndt, M. and Noubir, G. (2013), "Application-awareness in SDN", *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, pp. 487–492.

Qureshi, O.M., Hafeez, A. and Kazmi, S.S. (2020), "Ahmedpur Sharqia oil tanker TRAGEDY: Lessons learnt from one of the biggest road accidents in history", *Journal of Loss Prevention in the Process Industries*, Vol. 67, p. 104243.

Ranjan, J. and Foropon, C. (2021), "Big data analytics in building the competitive intelligence of organizations", *International Journal of Information Management*, Vol. 56, p. 102231.

Raut, R., Narwane, V., Kumar Mangla, S., Yadav, V.S., Narkhede, B.E. and Luthra, S. (2021), "Unlocking causal relations of barriers to big data analytics in manufacturing firms", *Industrial Management & Data Systems*, Vol. 121 No. 9, pp. 1939–1968.

Rezapour, M. and Ksaibati, K. (2021), "Application of machine learning technique for optimizing roadside design to decrease barrier crash costs, a quantile regression model approach", *Journal of Safety Research*, Vol. 78, pp. 19–27.

Ribes, D. and Poth, J. (2014), "Flexibility relative to what? change to research infrastructure", *Journal of the Association for Information Systems*, Vol. 15 No. 5, pp. 287–305.

Ritala, P., Olander, H., Michailova, S. and Husted, K. (2015), "Knowledge sharing, knowledge leaking and relative innovation performance: An empirical study", *Technovation*, Vol. 35, pp. 22–31.

Russom. P. (2013), "Managing big data," *TDWI Best Practices Report, TDWI Research*, pp. 1-40.

Sarkar, D., Bali, R. and Sharma, T. (2017), "The Python Machine Learning Ecosystem", *Practical Machine Learning with Python*, pp. 67–118.

Sarkar, S., Khatedi, N., Pramanik, A. and Maiti, J. (2019), "An ensemble learning-based undersampling technique for handling class-imbalance problem", *Proceedings of ICETIT 2019*, pp. 586–595.

Sarkar, S., Pramanik, A., Maiti, J. and Reniers, G. (2020), "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data", *Safety Science*, Vol. 125, p. 104616.

Scholkmann, A.B. (2021), "Resistance to (Digital) change", *Digital Transformation of Learning Organizations*, pp. 219–236.

Schuetz, S. and Venkatesh, V. (2020), "'research perspectives: The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction "", *Journal of the Association for Information Systems*, pp. 460–482.

Selcuk, S. (2016), "Predictive maintenance, its implementation and latest trends", *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, Vol. 231 No. 9, pp. 1670–1679.

Singh, C., Singh, D. and Khamba, J.S. (2020), "Analyzing barriers of green lean practices in manufacturing industries by DEMATEL approach", *Journal of Manufacturing Technology Management*, Vol. 32 No. 1, pp. 176–198.

Singh, K., Maiti, J. and Dhalmahapatra, K. (2019), "Chain of events model for safety management: Data analytics approach", *Safety Science*, Vol. 118, pp. 568–582.

Skakun, S., Roger, J.-C., Vermote, E.F., Masek, J.G. and Justice, C.O. (2017), "Automatic sub-pixel co-registration of landsat-8 operational land imager and sentinel-2a multi-spectral instrument images using phase correlation and machine learning based mapping", *International Journal of Digital Earth*, Vol. 10 No. 12, pp. 1253–1269.

Suh, Y. (2021), "Sectoral patterns of ACCIDENT process for occupational safety using NARRATIVE texts of OSHA DATABASE", *Safety Science*, Vol. 142, p. 105363.

Tan, P.-N., Steinbach, M., Karpatne, A. and Kumar, V. (2020), *Introduction to Data Mining*, Pearson, New York, NY.

Tixier, A.J.-P., Hallowell, M.R., Rajagopalan, B. and Bowman, D. (2016), "Application of machine learning to construction injury prediction", *Automation in Construction*, Vol. 69, pp. 102–114.

Wanasinghe, T.R., Wroblewski, L., Petersen, B.K., Gosine, R.G., James, L.A., De Silva, O., Mann, G.K., *et al*. (2020), "Digital twin for the oil and gas industry: Overview, research trends, opportunities, and challenges", *IEEE Access*, Vol. 8, pp. 104175–104197.

Wang, L., Cao, Q. and Zhou, L. (2018), "Research on the influencing factors in coal mine production safety based on the combination of DEMATEL and ism", *Safety Science*, Vol. 103, pp. 51–61.

Wasewar, K.L. and Kumar M.S. (2010), "Quantitative Risk Assessment (QRA) of a Petroleum Refinery," *IUP Journal of Chemistry*, vol. 3, pp.1-7.

Wiegmann, D.A. and Shappell, S.A. (2017), "The human Factors analysis and classification System (hfacs)", *A Human Error Approach to Aviation Accident Analysis*, pp. 45–71.

Yang, Y. and Wu, L. (2021), "Machine learning approaches to the unit commitment problem: Current trends, emerging challenges, and new strategies", *The Electricity Journal*, Vol. 34 No. 1, p. 106889.

Yazdi, M., Nedjati, A., Zarei, E. and Abbassi, R. (2020), "A novel extension of DEMATEL approach for probabilistic safety analysis in Process Systems", *Safety Science*, Vol. 121, pp. 119–136.

Yokoyama, H. (2019), "Machine learning system architectural pattern for improving operational stability", *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*, available at:http://doi.org/10.1109/icsa-c.2019.00055.

Zarei, E., Yazdi, M., Abbassi, R. and Khan, F. (2019), "A hybrid model for Human Factor Analysis in process accidents: FBN-HFACS", *Journal of Loss Prevention in the Process Industries*, Vol. 57, pp. 142–155.

Zavadskas. E, Kaklauskas. A, and Sarka. V. (1994), "The new method of multicriteria complex proportional assessment of projects," *Technological and economic development of economy*, vol. 1, pp. 131-139.