Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen

doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P5729-cd

Clustering for Learning from Safety-Related Undesired Events: Application to the Iron and Steel Industry

Paola Cocca, Martina Zorzi, Filippo Marciano, Giuseppe Tomasoni

Department of Mechanical and Industrial Engineering, University of Brescia, Italy. E-mail: paola.cocca@unibs.it, martina.zorzi@unibs.it, filippo.marciano@unibs.it, giuseppe.tomasoni@unibs.it

Massimo Guarascio, Bernardo Valente, Francesco Sergio Pisani

Institute for High Performance Computing and Networking (ICAR-CNR), Rende, Italy. E-mail: massimo.guarascio@icar.cnr.it, francescosergio.pisani@icar.cnr.it, Bernardo.valente@icar.cnr.it

Ettore Ritacco

Department of Mathematics, Computer Science and Physics, University of Udine, Italy. E-mail: ettore.ritacco@uniud.it

Elena Stefana

Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Italy. E-mail: elena.stefana@uniroma1.it

Safety-related undesired events can cause different kinds of workers' injuries and fatalities. *Learning from incidents* is a key step in safety risk management, which guides the exploitation of information for implementing effective safety-related decision-making processes. To accelerate the overall process and mitigate the impact of potential human *biases*, Machine Learning (ML) techniques may be adopted. However, available sources of safety incident reports frequently collect brief unstructured narratives with significant missing data, which are also phrased in a no standardised structure and language. In such a context, relying only on outcomes provided by ML techniques is risky, highlighting the need for human intervention to ensure meaningful results. For such reason, this paper proposes a multi-step approach integrating a hierarchical clustering and subject matter expert evaluations for *learning from incidents*. The proposed approach has been applied to examine undesired events happened in the iron and steel industry, i.e., one of the most hazardous industries in the world, where a multitude of risks can potentially give rise to a wide range of accidental scenarios. A set of 24 clusters were identified, providing insights into relationships among consequences, number of events, and operating conditions.

Keywords: Learning from accidents, safety management, human-in-the-loop, Natural Language Processing, unsupervised learning, Occupational Safety and Health Administration.

1. Introduction

According to the International Labour Organization (ILO), nearly 400 million workers worldwide sustained a non-fatal work injury, and three million workers died due to work-related accidents and diseases in 2019 (ILO 2023). Thorough investigations of past incidents and near-misses allow identifying causes and factors responsible for the creation of hazardous scenarios and stimulating the identification of measures to enhance safety performance. Accordingly, the *learning from incidents* process has emerged as a pivotal component of modern safety management supporting the development of proactive Occupational Safety and Health (OSH) and operational risk management (Stefana et al. 2024b). The aim of *learning from incidents* is to prevent the recurrence of undesired events, mitigate damages, and enhance safety performance (Lindberg et al. 2010). By leveraging information about past events, organizations cannot only improve safety performance, but also contribute to the improvement of standards and best practices (Campari et al. 2023).

The adoption of Machine Learning (ML) techniques can offer a significant contribution to the *learning from incidents* process: they may support the investigation of an extensive number of records by mitigating the impact of potential human *biases* and accelerating the overall process (Stefana et al. 2024c). The use of ML techniques permits dealing with large dimensional data, enhancing knowledge about hazardous scenarios, detecting patterns and risk factors that human assessors may not be able to immediately recognize, and predicting incident outcomes, injury risk and/or severity (Paltrinieri et al. 2019, Sarkar and Maiti 2020). In scientific literature on safety topics, a certain number of publications address the prediction of undesired events by applying ML algorithms across various sectors: e.g., in the construction industry (e.g., Ayhan and Tokdemir 2019, Ghodrati et al. 2018), in the mining sector (e.g., Sanmiquel et al. 2015), in the iron and steel industry (e.g., Sarakar et al. 2020, Verma and Maiti 2018).

However, the application of ML techniques for the *learning from incidents* process may pose several challenges. Indeed, available event narratives are often uncertain, unstructured, or lacking some details, and this results in a lower prediction power of algorithms (Sarkar et al. 2023). The quality of incident data is a critical factor for the effective application of the ML techniques, and it is dependent on the understanding and expertise of the subjects responsible for reporting incidents in the database (Verma et al. 2020). Moreover, the structure and language of incidents are typically lacking in standardization and uniformity (Verma et al. 2023). Therefore, relying only on the results produced by ML algorithms is risky, emphasizing the necessity for human intervention to ensure the reliability and meaningfulness of the outcomes. The involvement of subject matter experts permits capturing any potential biases introduced by automated processes, providing a deeper understanding of the context, and increasing confidence about the appropriateness of the obtained results (Stefana et al. 2024c).

In such a context, this paper aims to propose a multi-step approach integrating hierarchical clustering with subject matter experts' evaluations to enhance the *learning from incidents* process in the safety domain. Clustering identifies related data points in large datasets: it permits grouping data into collections (i.e., *clusters*) according to the similarities of data point features and characteristics (Ezugwu et al. 2022, Sarker 2021). Hierarchical clustering organizes data into a tree-based representation, not requiring the specification of the number of clusters (Cabezas et al. 2023). The evaluations performed by subject matter experts refine the clustering application and review the results.

The proposed approach is applied to analyze incidents that occurred in the iron and steel industry. Such industry is one of the most hazardous sectors: it is a complex socio-technical system with all components of operational safety and OSH (Verma et al. 2014), where workers are exposed to numerous physical, chemical, and mechanical hazards (ILO 2012). Moreover, the use of heavy machinery can cause fatal incidents or serious injuries to employees who are struck by moving parts or become caught in equipment (Zorzi et al. 2024).

The remainder of the paper is organized as follows. Section 2 describes the methodology, whereas its application to the case study is presented in Section 3. The discussion about the practical implications and concluding remarks are provided in the final section.

2. Methodology

The methodology is composed of three main steps: (i) safety-related undesired event identification, (ii) ML approach, and (iii) *learning from incidents*. It is shown in Fig. 1, and described in the next paragraphs.

2.1. *Safety-Related Undesired Event Identification*

The safety-related undesired event identification comprises the three activities in the following.

- Define the purpose of the analysis: it is necessary to specify the study objectives and the investigated undesired events regarding OSH and/or operational perspectives. A set of inclusion and exclusion criteria (including any data quality requirements) should be established to precisely delineate the boundaries of the research.

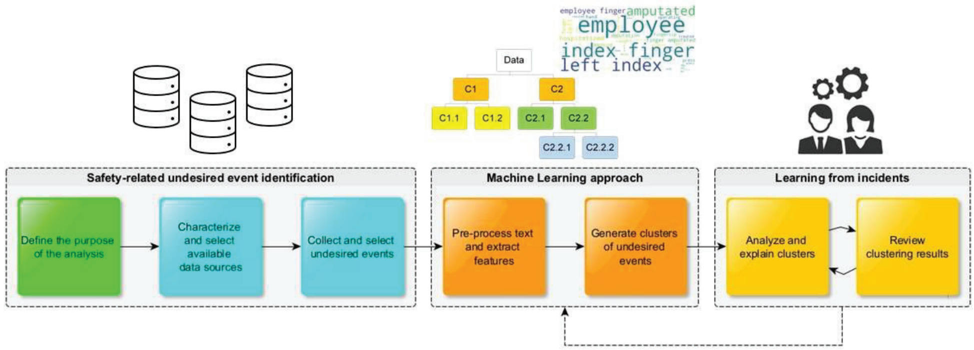

Fig. 1. Multi-step approach integrating hierarchical clustering with subject matter experts' evaluations.

- Characterize and select available data sources, in accordance with the defined purpose. Interesting reports can be gathered through national and international data sources that are specifically related to the phenomenon investigated, as well as generic sources collecting different kinds of safety-related undesired events.
- Collect and select undesired events to easily visualize data and implement any pre-processing techniques. Subject matter experts should examine the records according to the inclusion and exclusion criteria to select only those relevant for the analysis purpose.

2.2. *Machine Learning Approach*

The ML approach aims at pre-processing texts contained in specific record fields (e.g., title) and performing the clustering of the incidents.

Different text pre-processing techniques are applied to prepare the incident narratives for the analysis:

- tokenization: the text is separated into individual words, with any numerical, punctuation, or special characters removed;
- stop word removal: common words are eliminated, as they add little value in terms of distinguishing categories;
- stemming: to facilitate the process of grouping related words, the basic form of words is shortened;
- lemming: words are reduced to their fundamental form, considering grammatical and contextual factors for meaningful conversion;
- N-gram generation: all possible sequences of N contiguous words contained in the text are extracted for developing a structured propositional representation.

This process enables the standardization of word forms and the elimination of noise from the data. To optimize the effectiveness of clustering, a feature selection process is integrated into the processing stage. Subject matter experts identify pertinent features of interest (e.g., degree of consequences), after which an information gain-based selection is conducted, utilizing the identified feature as the target variable.

The pre-processed data are used to feed a clustering algorithm to identify recurring incident patterns. Hierarchical clustering is adopted in this study since it can group data with different levels of granularity according to the characteristics of the given data (Yang and Lin 2024). Specifically, the *X-means* algorithm (Pelleg and Moore 2000) is selected for its ability to determine the optimal number of clusters. The algorithm is an extension of the *K-means* clustering algorithm, and determines the optimal number of clusters iteratively, starting with *K-means* using an initial, lower-bound guess for clusters. It evaluates each cluster for potential splitting by performing a local *k-means* with k=2, using the Bayesian Information Criterion (BIC) (Kass and Wasserman 1995) to compare the original cluster with the split ones. If the split clusters have a lower BIC (indicating a better fit), the split is accepted. This process continues until no further BIC improvement is observed. The Euclidean distance serves as the metric for evaluating intra-cluster distances.

During this step of the approach, subject matter experts are involved in refining the stop word list and suggesting stopping criteria (e.g., minimum number of records per each cluster).

2.3 *Learning from Incidents*

Subject matter experts also play a key role during the actual step of *learning from incidents*: they examine the clustering results, analyze and explain clusters, suggest the merging of those that can be considered similar from a safety point of view. The activities of *analyzing and explaining clusters*, and *reviewing clustering results* are performed iteratively until a reasonable number of relevant clusters are obtained. To support such evaluations, the ML approach provides the following information:

- a *Term Frequency-Inverse Document Frequency* (TF-IDF) score for each cluster to identify terms frequent in a cluster but not frequent in all the clusters: TF captures how frequently a term occurs in a document, while IDF diminishes the weight of terms that appear in many documents;
- two word clouds for each cluster: one where the size of each word is proportional to its TF-IDF score, and another one where the size of each word is proportional to its frequency in selected fields (e.g., abstracts) of the records;
- a *t-Distributed Stochastic Neighbor Embedding* (t-SNE) graph (van der Maaten and Hinton 2008) for each iteration of the clustering process, to visually analyze the clustering quality and provide insights into the actual separation between the clusters.

3. Case Study

The methodology summarized in Section 2 was applied to investigate the incidents that occurred in the iron and steel industry in US from 2000 to 2023, whose records were collected by Occupational Safety and Health Administration (OSHA) available at: https://www.osha.gov/ords/imis/accidentsearch.html. Such records represent the inspections performed by OSHA in response to a fatality, catastrophe, or employer-reported referral. They can be searched for by keywords, text in the description or abstract, event date, and industry types through *Standard Industrial Classification* (SIC) or *North American Industry Classification System* (NAICS) codes.

3.1. *Safety-Related Undesired Event Identification*

We were interested in learning from incidents happened in the iron and steel industry. To consider only those incidents related to this industry, we searched for undesired events characterized by the specific SIC and/or NAICS codes; some examples are given in Table 1.

Table 1. Included and excluded SIC and NAICS codes (examples).

	Included codes	Excluded codes
SIC	Iron and steel foundries: 3320, 3321, 3322, 3324, 3325	Primary smelting and refining of nonferrous metals: 3330, 3331, 3334, 3339, 3341
NAICS	Iron and steel forging: 332111	Copper rolling drawing extruding: 331421

We obtained 4412 records, whose details were downloaded in a .csv file. Accordingly, the resulting database was composed of 4412 rows and 24 columns referring to the various types of data directly provided in the investigation summary (e.g., abstract, nature of injury).

3.2. *Machine Learning Approach*

In the step related to the ML approach, we applied tokenization, stop word removal, stemming, lemming, and N-gram generation as text-preprocessing techniques. We considered N-gram equals to 3. We adopted the *X-means* algorithm by taking into account the title and abstract fields of the records in the OSHA database.

We obtained 39 clusters from the ML algorithm by splitting the set of incidents through a different number of iterations (i.e., minimum number of iterations equals to 4, maximum number of iterations equals to 7).

3.3. *Learning from Incidents*

The obtained 39 clusters were examined independently by five subject matter experts to explain them and suggest any merging of specific clusters. Indeed, clusters that appeared similar in terms of the severity of workers' consequences (e.g., fatalities or non-hospitalized injuries) and of incident scenarios were combined. Brainstorming sessions were organized among the experts to discuss any discrepancies, and reach a consensus on the final set of relevant clusters.

Table 2 summarizes the 24 final clusters representing the undesired events that occurred in the iron and steel industry and collected by OSHA. Some cluster titles contain the term *injury* for referring to different consequences on workers due to incidents, while the word *non-fatal* for identifying set of events with either hospitalized or non-hospitalized adverse outcomes. The achievement of these clusters allowed for reducing the average entropy, whose value is equal to 0.56. The entropy is a measure related to the disorder, randomness, or uncertainty within a set of data: it increases as the classification of objects in a cluster becomes more varied (Ezugwu et al. 2022).

In the 24 final clusters, nine are associated with fatal incidents, ten with hospitalized injuries, one with non-hospitalized consequences, one with both fatalities and hospitalized effects, and three with non-fatal events.

Fatalities were caused by events when employees: (i) were struck, crushed, caught by equipment (e.g., crane, lathe, press, shaft), falling or flying objects (e.g., metal plate, steel coil), or vehicles (e.g., forklift, truck) (i.e., C7, C8, C11, and C16), (ii) were injured in fires or explosions (e.g., of furnaces, vessels, gas cylinders, or boilers) (i.e., C9 and C13), (iii) fell from height (e.g., scaffolding, roof, platform), through openings, or into industrial equipment or containers (e.g., furnace, tank) (i.e., C10), (iv) were exposed to hazardous substances (e.g., carbon monoxide) or inert gases (e.g., nitrogen, argon), or contacted molten metal (i.e., C12), (v) experienced infectious (i.e., COVID-19) or cardiac diseases (i.e., C14 and C15).

These undesired events mainly occurred during traditional production processes, while employees operating industrial equipment, or when maintenance activities were carried out. Furthermore, C7, C11, and C9 represent the clusters related to fatalities with a high number of records: this highlights the relevance of incident scenarios involving human activities in proximity of rotating equipment, suspended loads, moving vehicles. The incidents were frequently triggered by the complete absence of safety measures, the presence of ineffective ones, and/or the non-adoption of the existing ones. This regarded at least one of engineering controls (e.g., guarded machine), administrative controls (e.g., adoption of precise procedures for de-energizing equipment), and/or Personal Protective Equipment (PPE), including fall protection systems. These outcomes indicate the necessity for a thorough examination of the actual need and the reasons for the presence of workers during the conduction of such operations. Furthermore, the implementation of effective and various safety barriers is recommended, as well as the identification of alternative (also novel) modes of carrying out the activities.

Hospitalized injuries were mostly responsible for amputations, fractures, lacerations, burns related to chemical substances or heat, electric shocks, or heat stress. The main injured body parts were head, forehead, hands, fingers, fingertips. Such consequences resulted from different kinds of scenarios, e.g., when the employees: (i) were exposed to hazardous substances (e.g., carbon monoxide or acids), were injured in fires or explosions (e.g., of furnaces or tanks, or when there is interaction between the molten material and water), contacted molten metal or caustic fluids, or inhaled dust (i.e., C1 and C4); (ii) were struck or caught by falling objects or equipment, including conveyors, molding machines, rollers, and lathes (i.e., C2, C3, C5, and C16); (iii) fell from height (e.g., roof, ladder), into existing openings or confined spaces (e.g., pit, tank) (i.e., C6), (iv) operated process equipment (i.e. C20, C22, C23, and C24). In particular, the C20, C22, C23, and C24 clusters account for 22% of the incidents within the entire database analyzed, highlighting the relevance of events resulting in finger amputations and hand injuries during the machine operations in the iron and steel industry. Consequently, relevant efforts should be devoted to prevent, in every conceivable way, all possible interactions between humans and equipment during its functioning.

The operations of equipment (e.g., press, saw, grinding wheel, auger, roller) also caused non-hospitalized fingertip amputations (i.e., C19). It represents a minority of undesired events: about 3% of the incidents in the database were classified in this cluster. In addition to these clusters addressing a precise severity of incidents related to the machine operations, the C17, C18, and C21 clusters revealed a combination of non-fatal adverse consequences on workers, including amputations, fractures, and lacerations. Such clusters contain about 15.9% of the total incidents in the database.

Table 2. Final clusters of incidents in the iron and steel industry, collected by OSHA.

ID	Title	Record number	Description
C1	Heterogeneous hospitalized injuries	233	Events caused by falling objects, falling from platforms or into tanks, exposure to hazardous substances, fires, explosions. They were responsible for amputation, electric shock, heat stress, burns.
C2	Hospitalized fractures and lacerations	88	Events caused by flying objects, striking objects or equipment, use of presses. The main injured body parts were head and fingers.
C3	Hospitalized injuries for being caught by machine	173	Events caused by operating machines or tasks performed without safety measures. They were responsible for fractures or amputations.
C4	Hospitalized burns	174	Events caused by fires, explosions, chemical exposures, or contact with molten metal. Some occurred during welding operations.
C5	Hospitalized fractures for striking items	433	Events caused by falling objects or moving machinery. Multiple injuries were caused by forklifts, cranes, and trucks.
C6	Hospitalized fractures for falls	212	Events caused by fall from ladders, roofs, or platforms, often due the lack of fall protection systems or the presence of inadequate ones.
C7	Fatalities for striking and crushing	249	Events caused by falling or flying objects, falling from platforms, vehicles, or forklifts, often linked to insufficient safety measures.
C8	Fatalities for being caught by machine	63	Events caused by machine operations, where employees were caught in or between moving parts.
C9	Heterogeneous fatal incidents during operations	148	Events caused by explosions, fires, electrocutions, vehicle rollovers, or pinning from vehicles during the operations of equipment.
C10	Fatal falls	79	Events caused by falling from height, through openings, or into furnace, often related to the lack of fall protection measures.
C11	Fatalities for crushing	155	Events caused by being crushed by or between machines, falling objects, forklifts, typically during operations or maintenance tasks.
C12	Fatalities after hospitalization	130	Events caused by falling from platforms or into containers, exposure to hazardous substance, explosions, or contacts with molten metal.
C13	Heterogeneous fatal incidents	74	Events caused by respiratory issues, burns, asphyxiation, sometimes leading to unresponsiveness or unconsciousness of workers.
C14	Fatal infectious diseases	27	Events caused by COVID-19 infection.
C15	Fatal cardiac diseases	53	Events caused by heart attacks, cardiac arrests, cardiorespiratory arrests, or cardiovascular system diseases.
C16	Heterogeneous fatal and hospitalized injuries	315	Events caused by being struck, crushed, caught by machine, also responsible for amputations or fractures.
C17	Non-fatal finger amputations	420	Events caused by machine operations, often linked to lack or non-adoption of guards, de-energization or lockout procedures.
C18	Non-fatal finger amputations for being caught	214	Events caused by being caught by or in machines, sometimes due to the use of unguarded machines or the removal of machine guards.
C19	Non-hospitalized fingertip amputations	134	Events caused by machine operations, sometimes due to unguarded machine use, machine guard removal, guard failure.
C20	Hospitalized amputations and fractures	265	Events caused by machine operations, often due to lack or non-adoption of lockout procedures, or unguarded machine use.
C21	Non-fatal consequences for machine operations	67	Events caused by machine operations, responsible for fractures, lacerations, and amputations.
C22	Hospitalized hand injuries for machine operations	130	Events caused by machine operations, sometimes due to the lack of machine guards or the presence of ineffective guards.
C23	Hospitalized hand and finger injuries	320	Events caused by machine operations, often due to the lack or non-use of machine guards, lack or non-adoption of lockout procedures.
C24	Hospitalized finger amputations	256	Events caused by machine operations, often related to the use of unguarded equipment.

4. Discussion and Conclusions

This paper describes a multi-step approach integrating hierarchical clustering with subject matter experts' evaluations to support the *learning from incidents* process in the safety domain. Its application to undesired events in the iron and steel industry permitted identifying 24 clusters of fatal and non-fatal incidents linked to a wide range of interactions between social and technical elements typically operating in such companies. Indeed, the highlighted critical and risky scenarios mainly refer to equipment operations, falling objects, falls from height, contacts with molten material.

In addition to the factors contributing to incidents for each cluster, the analysis encompassing the entire set of events highlights the fundamental adverse role of missing or inadequate safety measures in the different hazardous scenarios. This stresses the necessity of not only implementing effective and specific safety measures at all levels of hierarchy of controls, but also monitoring their proper functioning and stimulating an appropriate application by the workers. Accordingly, any individual in a company, based on his/her role and responsibilities, is called upon to contribute to the development of a safety management system able to promote proactive OSH and operational risk management. It is one of the main lessons learned pointed out by the examination of the collected incidents, which is by no means insignificant. Note that the primary objective of incident investigations is not the apportioning of blame to the individuals involved in an undesired event, but rather the acquisition of a comprehensive understanding of the contributing factors, with particular attention to human factors (Stefana et al. 2024a). However, the majority of the records examined in this study do not permit fully individuating the several shortcomings and critical interactions among the elements involved in the events. This calls for extending our analysis to other databases of safety-related undesired events (e.g., Analysis, Research and Information on Accidents - ARIA) by employing our approach for preliminarily classifying the incidents into homogeneous groups.

In general, the step of ML approach proposed in this study could be further enhanced since the applied clustering algorithm: (i) split the set of data into only two distinct clusters, (ii) produced few clusters containing incident scenarios quite different from each other (i.e., C1, C9, C13, and C16), and (iii) was largely based on both the exact formulation of the sentences and the words used (e.g., the ML approach produced the two different clusters C9 and C13 for record descriptions containing the terms *killed* and *dies*, respectively). These limitations have been managed in this study by integrating the ML approach with human evaluations. This allowed for recognizing those clusters to be merged from the safety point of view because of similar incident outcomes and contributing factors.

Future research activities could investigate the modification of input parameters of the algorithm, the usage of other clustering and classification techniques, and the adoption of Large Language Models (LLMs) to boost the generalization capability. Furthermore, specific LLMs could be developed for enhancing the current versions of the incident databases: on the one hand, they could help in completing any missing data in the records that may have a role in the understanding of the events (e.g., severity, SIC and NAICS codes), on the other they could support the definition of structured features and attributes by extracting data from fully textual narratives. In turn, this could represent a fundamental phase towards the creation of a unique dataset collecting incidents that occurred in relevant contexts (e.g., the iron and steel industry) in several countries and indexed in various relevant data sources. Their in-depth (data) analysis could ultimately suggest additional features that should be systematically collected and made available to actually improve the *learning from incidents* process for safety management in industrial plants.

Acknowledgement

This paper has been supported from the European Union's Next Generation EU - Italian Ministerial grant PRIN 2022 - No 20222YRP2F - CUP: D53D23003010006.

References

Ayhan, B.U. and O.B. Tokdemir (2019). Predicting the outcome of construction incidents. *Safety Science 113*, 91-104.

Cabezas, L.M.C., R. Izbicki, and R.B Stern (2023). Hierarchical clustering: Visualization, feature importance and model selection. *Applied Soft Computing 141*, 110303.

Campari, A., A.J. Nakhal Akel, F. Ustolin, A. Alvaro, A. Ledda, P. Agnello, P. Moretto, R. Patriarca, and N. Paltrinieri (2023). Lessons learned from HIAD 2.0: Inspection and maintenance to avoid hydrogen-induced material failures. *Computers & Chemical Engineering 173*, 108199.

Ezugwu, A.E., A.M. Ikotun, O.O. Oyelade, L. Abualigah, J.O. Agushaka, C.I. Eke, and A.A. Akinyelu (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence 110*, 104743.

Ghodrati, N., T.W. You, S. Wilkinson, and M. Shahbazpour (2018). A new approach to predict safety outcomes in the construction industry. *Safety Science 109*, 86-94.

ILO (International Labour Organization) (2012). Industries Based on Natural Resources. Iron and Steel. *Encyclopaedia of Occupational Health and Safety*, Part XI, Chapter 73. ILO, Geneva.

ILO (International Labour Organization) (2023). *A call for safer and healthier working environments*. ILO, Geneva.

Kass, R.E. and L. Wasserman (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association 90*, 431, 928-934.

Lindberg, A.-K., S.O. Hansson, and C. Rollenhagen (2010). Learning from accidents - What more do we need to know?. *Safety Science 48*, 714-721.

Paltrinieri, N., L. Comfort, and G. Reniers (2019). Learning about risk: machine learning for risk assessment. *Safety Science* 118, 475-486.

Pelleg, D. and A.W. Moore (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, 727-734, San Francisco, CA (USA).

Sanmiquel, L., J.M. Rossell, and C. Vintrò (2015). Study of Spanish mining accidents using data mining techniques. *Safety Science 75*, 49-55.

Sarkar, S. and J. Maiti (2020). Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis. *Safety Science 131*, 104900.

Sarkar, S., A. Pramanik, and J. Maiti (2023). An integrated approach using rough set theory, ANFIS, and Z-number in occupational risk prediction. *Engineering Applications of Artificial Intelligence 117*, Part A, 105515.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science 2*, 160.

Stefana, E., E. De Paola, C.S. Snaiderbaur Bono, F. Bianchini, T. Vagheggi, and R. Patriarca (2024a). Beyond blame: A systemic accident analysis through a neutralized human factors taxonomy. *Human Factors and Ergonomics in Manufacturing & Service Industries 34*, 5, 450-465.

Stefana, E., F. Marciano, N. Paltrinieri, and P. Cocca (2024b). A systematic approach to develop safety-related undesired event databases for Machine Learning analyses: Application to confined space incidents. *Process Safety and Environmental Protection 182*, 279-297.

Stefana, E., M. Ramos, and N. Paltrinieri (2024c). Machine learning-based literature review on the concept of safety barriers against hazardous events. *Journal of Loss Prevention in the Process Industries 92*, 105470.

van der Maaten, L. and G. Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*, 2579-2605.

Verma, A. and J. Maiti (2018). Text-document clustering-based cause and effect analysis methodology for steel plant incident data. *International Journal of Injury Control and Safety Promotion* 25, 4, 416-426.

Verma, A., J. Maiti, and G. Boustras (2020). Analysis of categorical incident data and design for safety interventions using axiomatic design framework. *Safety Science 123*, 104557.

Verma, A., K. Dhalmahapatra, and J. Maiti (2023). Forecasting occupational safety performance and mining text-based association rules for incident occurrences. *Safety Science 159*, 106014.

Verma, A., S. Das Khan, J. Maiti, and O.B. Krishna (2014). Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Safety Science 70*, 89-98.

Yang, J. and C.T. Lin (2024). Enhanced Adjacency-Constrained Hierarchical Clustering Using Fined-Grained Pseudo Labels. *IEEE Transactions on Emerging Topics in Computational Intelligence 8*, 3, 2481-2492.

Zorzi, M., E. Stefana, P. Cocca, F. Marciano, G. Tomasoni, and R. Venturetti (2024). Learning from undesired events in the iron and steel industry using machine learning techniques. *Proceedings of the XXIX Summer School "Francesco Turco"*, Otranto (Italy).