Learning from undesired events in the iron and steel industry using machine learning techniques

Zorzi M.*, Stefana E.**, Cocca P.*, Marciano F.*, Tomasoni G.*, Venturetti R.*

** Department of Mechanical and Industrial Engineering, University of Brescia, Via Branze 38, Brescia, Italy (martina.zorzi@unibs.it, paola.cocca@unibs.it, filippo.marciano@unibs.it, giuseppe.tomasoni@unibs.it, r.venturetti001@studenti.unibs.it)*

*** Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Via Eudossiana 18, Rome, Italy (elena.stefana@uniroma1.it)*

Abstract: Safety in the iron and steel industry requires utmost attention: it is one of the most hazardous industries in the world, whose activities may expose workers to a wide range of hazards (e.g., the presence of molten metal). In such a context, it is particularly important to learn from what happened in past incidents and accidents to avoid the same dynamics happening again in the future. Machine Learning (ML) techniques could assist in this task because they permit discovering patterns and correlations from existing data and speeding up the identification of the relevant characteristics of undesired events. The scientific literature offers some contributions dealing with the analysis of incidents and accidents in the iron and steel industry through ML techniques, but they are based only on data collected from one or a few specific steel plants. Therefore, the generalisability of their results is limited. For this reason, this paper aims to learn from safety-related undesired events happened in various iron and steel plants by analysing a broader set of incidents and accidents, collected from relevant international data sources (i.e., Analysis, Research and Information on Accident (ARIA), Chemical Safety Board (CSB), and Occupational Health and Safety Administration (OSHA)). A dataset of incidents and accidents was then built, and a set of features and labels was defined to structure this dataset. We used Orange software to train and test well-known ML classification algorithms (e.g., Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbors (kNN), and Artificial Neural Network (ANN)) to predict the type of occurred incident or accident and its degree. From the analysis of the dataset, we obtained that 45% of incidents were categorised as "caught in", 20% as "struck by", and 9% as "fall". The main triggers associated to these types of incidents were "moving object" in 35% of the cases, and "operating machine" in another 35% of the cases.

Keywords: Occupational safety; supervised learning; incident database; risk management; process safety.

1.Introduction

The iron and steel industry refers to a specific sector of metallurgy, which concerns the treatment of minerals with a high iron content to obtain iron or different types of alloys containing iron, including steel, cast iron, and steels tide up (He and Wang, 2017). The iron and steel industry plays a pivotal role in the global economy, supplying goods to many relevant sectors such as transportation, construction, and automotive. In 2022, the globally iron and steel production reached 1885 million tons, with China emerging as the leading producer, followed by India, Japan, United States, and Russia (World Steel Association, 2023). Safety management in such industry is critical because of the complexity of its processes that involve both high-technology and labour-intensive aspects (Verma et al., 2014). Indeed, the processes may expose workers to a wide range of hazards or conditions that could cause incidents, injuries, death, illnesses, or diseases (ILO, 2005). Within this industry, employees are exposed to the heat of molten metal and slag at high temperature, toxic or corrosive substances, noise, and respirable airborne contaminants (Pickvance, 2011). Moreover, working with heavy machinery and equipment poses the risk of workers getting struck by moving parts or caught in machinery. The absence or failure to adhere to proper lifting techniques, safety measures, or storage procedures increase the risk of objects falling and causing injuries. The scientific literature describes some examples of hazards in the iron and steel industry. For instance, International Labour Office (ILO, 2005) reports the following most common type of hazards:

- physical hazards, such as exposure to noise, exposure to hazardous vibration, heat and cold stress, and exposure to ionizing and non-ionizing radiation;
- chemical hazards, such as exposure to chemicals, exposure to inhalable agents, and exposure to asbestos;
- safety hazards, such as confined spaces, the use of work equipment, falling objects;
- ergonomics hazards, such as musculoskeletal injuries.

To enhance both occupational safety and process safety in the iron and steel industry, "learning from incidents" plays a pivotal role (Cooke and Rohleder, 2006). Indeed, nowadays "learning from incidents" is instrumental for modern safety management (Stefana et al., 2024). By adopting a proactive approach to learning from incidents, organisations can identify root causes and take corrective actions to prevent similar incidents from happening in the future. According to Swedish Centre for Lessons Learned from Incidents and Accidents (NCO), the systematic learning from incidents has the purpose of preventing the recurrence of similar events, mitigating damages, and thereby enhancing occupational safety (Lindberg, Hansson and Rollenhagen, 2010). To achieve this, it is fundamental to extract and analyse knowledge regarding incidents and near-misses, and subsequently to communicate the discovered information to all who are involved. However, the identification of incidents and their characteristics can be a long and laborious process (Verma et al., 2014).

To overcome these drawbacks, a useful support could be provided by Machine Learning (ML) techniques. ML techniques are able to identify patters, correlations, and trends within the data (Alzubi, Nayyar and Kumar, 2018), thereby extracting valuable insights to foster a deeper understanding of a phenomenon. In the scientific literature some contributions dealing with the prediction of undesired events in the iron and steel industry by leveraging on ML techniques are available (Table 1). However, these studies are based on data collected from one or few specific steel plants. Therefore, the generalisability of their results is limited. This calls for a comprehensive analysis of incident and accidents that happened in various iron and steel plants, collected from relevant international data sources, to identify common patterns, correlations, and trends across the whole industry. Such analysis could support the development of effective safety rules, procedures, and programmes. To investigate the undesired events occurred worldwide and reported in several data sources, it is necessary to build a unique database with a proper structure for permitting the application of ML techniques.

This paper has the objective to build such database and use it to learn from safety-related undesired events that happened in the iron and steel industry by adopting well-known ML techniques. In contrast to existing studies in the literature, this research is based on the records collected from different sources with unstructured or semi-structured data: i.e., Analysis, Research and Information on Accident (ARIA), Chemical Safety Board (CSB), and Occupational Health and Safety Administration (OSHA). The remainder of the paper is organised as follows. Section 2 describes the methodology followed in this research. Results are presented and discussed in Section 3. Concluding remarks are provided in the final section.

2. Methodology

To achieve the objective of this paper, we implemented the strategy outlined in Figure 1.

The first step of our research concerned the selection of relevant safety-related data sources containing information on incidents and accidents that occurred in iron and steel companies worldwide. In particular, we selected the websites and archives of OSHA, CSB, and ARIA. We selected these organisations because their data are publicly accessible online, and offer textual descriptions of incident and accident scenarios. We only considered records related to the iron and steel industry by filtering them through the codes specific to the iron and steel activity: for OSHA and CSB, both Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS) codes were considered, while for ARIA we used Nomenclature d'activités française (NAF) code. In Table 2 the relevant SIC, NAICS, and NAF codes for this study are reported. All the records related to incidents and accidents that took place worldwide between 1980 and 2022, recorded in the OSHA, ARIA, and CSB data sources were retrieved. We obtained four different databases. Indeed, from OSHA data source we created two databases: one focused on incidents and accidents that caused severe injuries, and one containing both fatalities and catastrophes.

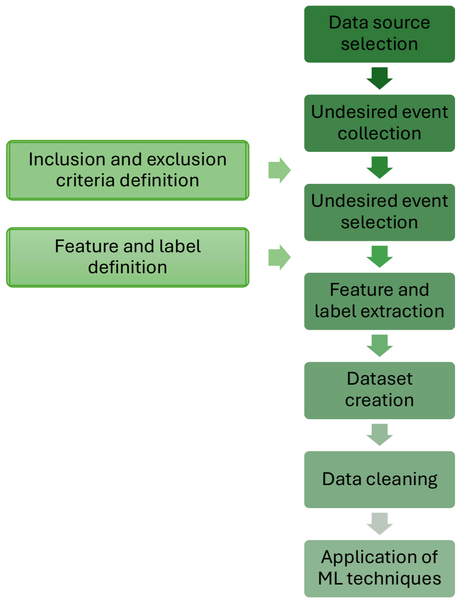

Figure 1: Strategy for the creation of the incident dataset for the ML application

Afterwards, we analysed and reviewed the retrieved records to select only those of interest. For such purpose, we defined a set of inclusion and exclusion criteria. For instance, a record was included if the undesired event occurred in the iron and steel industry and its description was available. We searched for the presence of specific details about the event: the cause(s), the year, the number of involved workers, the consequences (including number of deaths or number of injured workers), the degree or severity of consequences. On the contrary, the incident records were excluded if: (i) the data were unclear or missing, (ii) the event concerned no work-related activities or non-work personnel, or (iii) the event referred to a near miss. After the record selection, we organised information about incidents and accidents in a structured way by defining relevant features and labels. A feature is a "measurable property of an object or event with respect to a set of characteristics" (ISO, 2022), while a label represents a specific value that a feature can take (Nakano, Cerri and Vens, 2020). Since ML techniques operate on datasets characterised by features and labels, the process of feature selection assumes a crucial role in influencing algorithm performance. To select these features, we analysed data and information available in each data source, and we identified a set of interesting dimensions that allow properly describing the incidents and accidents under investigation. Once the four databases were manually structured by means of these features and labels, we merged them into an only dataset. A dataset represents a collection of data that are typically used for analysis, research, or ML purpose (Provost and Fawcett, 2013). Finally, we performed data cleaning to remove the duplicates.

The dataset so developed represented the key input for applying ML techniques and for analysing the factors that caused those undesired events. Specifically, our focus was on identifying relationships between the scenarios of the incidents and accidents and their corresponding consequences. Therefore, we employed some well-known ML algorithms to predict the degree of consequences and the type of incident or accident. We used supervised learning algorithms that learn to make predictions or classify new, unseen data based on the patterns and relationships learned from the training data (Hastie, Tibshirani and Friedman, 2009). Specifically, we implemented Logistic Regression (LR), K-Nearest Neighbors (kNN), Artificial Neural Network (ANN), and Random Forest (RF). We used Orange, which is an open-source data visualisation, ML, and data mining toolkit (Demsar, Curk and Erjavec, 2013). The workflow developed in Orange software for our analyses is depicted in Figure 2. It is composed of the following widgets:

- File, to read the dataset;
- Data sampler, to sample the data by a ratio (we chose 70-30 ratio: 70% of the original dataset was training data and the remaining 30% was test data);
- LR, RF, kNN, NN, i.e., the implemented ML algorithms;
- Test and score, to inspect the accuracy of the model on the test dataset;
- Confusion matrix, to perform additional analysis of cross validation results;
- Data table, to visualise the data selected in the Confusion Matrix;
- Predictions, to display the predictions of the model on the data;
- Save data, to save data on a file.

Table 1: Available approaches for incident analysis in the iron and steel industry through ML techniques

Authors (year)	Tasks	Type of ML techniques
Sarkar et al. (2020)	Identifying the injury severity of accidents in a steel plant in India	Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes (NB), K-Nearest Neighbours (KNN), CART (Classification And Regression Trees), Random Forest (RF)
Sarkar et al. (2019)	Developing a Decision Support System to predict occupational accident in a steel plant	Support Vector Machine (SVM), Random Forest (RF)
Sarkar et al. (2018)	Extracting patterns from unstructured accident text data of a steel plant	Self-Organizing Map (SOM), k-means clustering (K-MEANS), Hierarchical clustering
Song (2018)	Using predictive models to identify specific variables that have high correlations of an unsafe event within a steel plant	Binary Logit Model
Verma and Maiti (2018)	Analysing text data and discovering root causes behind the incidents data of a steel plant	Singular Value Decomposition (SVD), Expectation-Maximization (EM)
Dhalmahapatra et al. (2019)	Analysing near misses in electric overhead travelling crane operations within a steel plant	Multiple Correspondence Analysis (MCA), t-distributed Stochastic Neighbor Embedding (t-SNE), k-means clustering (K-MEANS)

Table 2: Relevant SIC, NAICS, and NAF codes

SIC	NAICS	NAF
3312, 3313, 3315, 3316, 3317, 3321, 3322, 3324, 3325, 3331, 3334, 3339, 3341, 3351, 3353, 3354, 3355, 3356, 3357, 3363, 3364, 3365, 3366, 3369, 3398, 3399, 3462, 3463	324199, 331110, 331210, 331221, 331222, 331313, 331314, 331315, 331318, 331410, 331420, 331491, 331492, 331511, 331512, 331513, 331523, 331524, 331529, 332111, 332112, 332618, 332811, 332813, 335929	B07.10, C19.10, C24.10, C24.20, C24.31, C24.32, C24.33, C24.34, C24.41, C24.42, C24.43, C24.44, C24.51, C24.52, C24.53, C25.11, C25.50, C25.61, C25.62

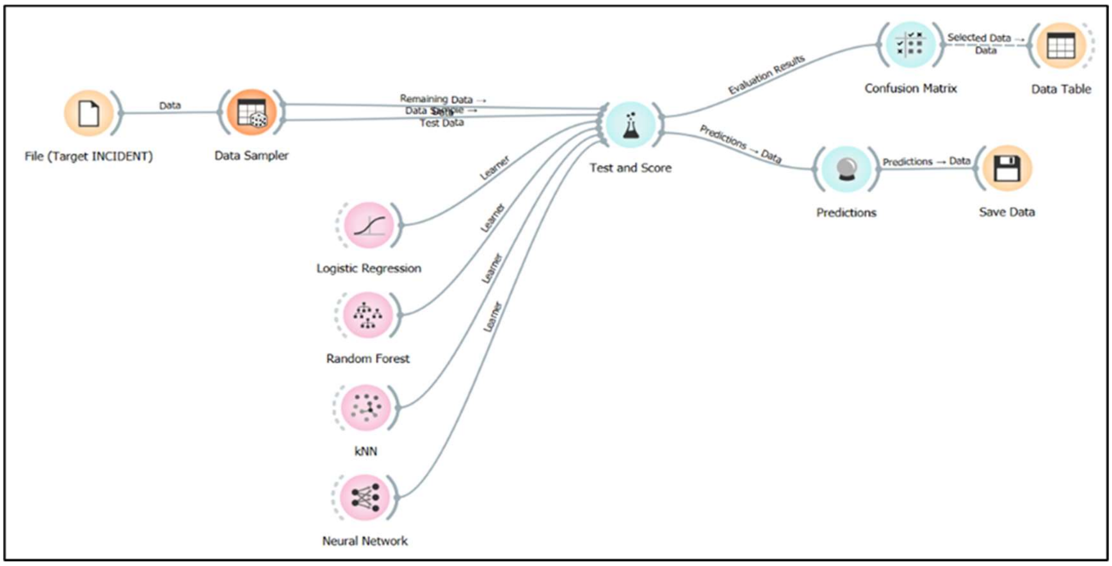

Figure 2: Workflow in Orange for the ML application

3. Results and discussion

3.1 Incident and accident dataset

We collected a total of 5750 records related to the iron and steel industry in the OSHA, CSB, and ARIA data sources. By applying the inclusion and exclusion criteria and deleting the duplicates, we obtained a dataset of 5630 records of fatal and non-fatal incidents occurred in the iron and steel industry worldwide. This number of records was obtained by selecting:

- 5 incidents from CSB;
- 143 incidents from ARIA;
- 5482 incidents from OSHA.

The majority of such events occurred in US: this is because OSHA and CSB contain records of incidents or accidents that happened in the US, while ARIA database inventories undesired events that occurred in France or abroad. Moreover, the time trend of the events highlights that the largest number of incidents and accidents occurred from 2015 to 2022 (31% of the total).

Each row of the dataset represents an incident, while each column a feature of the incident. We defined 14 features to capture essential aspects of the events and peculiarities of the iron and steel industry. Such features are summarised in Table 3. Furthermore, for each feature, we defined a set of labels. For instance, the labels associated with the feature "department" are: (1) casting shop, (2) elevator, (3) finishing shop, (4) furnace area, (5) melting area, (6) moulding area, (7) pre-processing area, (8) quality control area, (9) refining area, (10) storage area. The feature "degree" includes the following labels: (1) fatality, (2) fatality and hospitalised injury, (3) hospitalised injury, (4) non hospitalised injury.

Initially, we identified the most frequent incidents or accidents in the dataset. The analysis revealed that 45% of the events were categorised as "caught in", 20% as "struck by", and 9% as "fall". This result is unsurprising, given the nature of the industry. The presence of moving parts, conveyor belts, presses, and other mechanical systems increases the risk of workers getting caught in or struck by these objects. Moreover, handling heavy and bulky materials, such as sheets, steel beams, or machinery components, is a common task in that industry. The weight and size of these materials make them potential hazards. If not properly controlled, stored, or lifted, they can cause workers to be struck by or caught in. Furthermore, in the iron and steel industry, workers often operate at elevated heights, such as working on scaffolding, platforms, or structures. This increases the risk of falls, especially if proper preventive (i.e., presence of guard rails) and protective (i.e., fall arrest system) safety measures are not in place or used correctly.

We also analysed the triggers correlated to these undesired events. In the "caught in", "struck by" or "fall" related incidents or accidents (accounted for 74% of the total events in the dataset), the main trigger is "moving object" in 35% of the cases, and "operating machine" in another 35% of the cases. In particular, the trigger "moving object" can be related to scenarios where a worker is hit by a rolling or falling object. On the other hand, the trigger "operating machine" may involve situations where a worker gets compressed, pinched, pulled, or buried while operating the machine.

An examination of the severity of all the incidents and accidents revealed that:

- the most prevalent degree is "hospitalised injury" (54%), which was mainly caused by "caught in" (40%), "struck by" (22%), and "fall" (11%);
- fatalities represent the consequences in 23% of the total events, and are caused by "caught in" in 34% of cases, by "struck by" in 25% of cases, and by "fall" in 11% of cases;
- "non hospitalised" injuries result in 21% of the events, whose main causes are "caught in" (75%) and "struck by" (11%).

Furthermore, the department where the most fatalities occurred is the furnace area (Figure 3). This area is quite critical from the safety perspective since it is characterised by the presence of molten metal, moving objects and equipment, and hanging loads.

Table 3: Features of the dataset

Name	Description
ID	Number to uniquely identify a record, based on the ID provided in the data source.
Type of incident or accident	Classification of the occurred incident or accident.
Details	Additional notes regarding the incident or accident.
Trigger	The main condition, action, or circumstance that initiates the incident or accident scenario.
Data source	Indication about which database provides the event details.
Country	Place where the incident or accident occurred.
Department	Area within an organisation where the incident or accident occurred.
Event year	Year when the incident or accident occurred.
SIC/NAICS/NAF code	Code indicating the classification of the industry where the incident or accidents happened.
Number of involved employees	Total number of individuals involved in the incident or accident.
Number of deaths	Total number of fatalities resulting from the incident or accident.
Number of injuries	Total number of individuals who sustained injuries as a result of the incident or accident.
Degree	Severity of the incident or accident on workers.
Nature	Details about incident or accident consequences on workers.

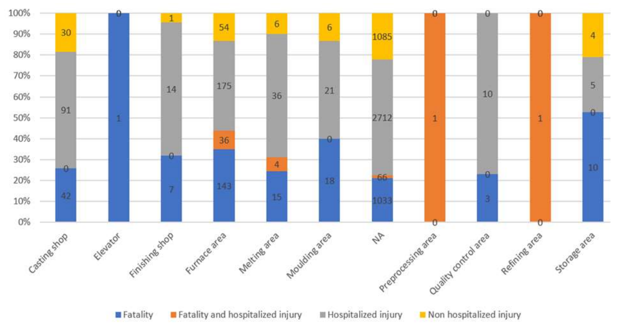

Figure 3: Degree of events in iron and steel departments

3.2 Prediction models by using ML algorithms

By employing Orange software, we investigated two different learning tasks. For this purpose, we defined two different features as target variables: "type of incident or accident" to predict the cause(s) of the event, and "degree" to predict the severity of the event. Subsequently, we investigated the correlation with SIC, NAICS, and NAF codes for both the target variables.

3.2.1 Predictions for the type of incident or accident

To predict the type of incident or accident, we trained LR, RF, NN, and kNN models. Six metrics (i.e., Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score (F1), Precision, Recall, and Mattew Correlation Coefficient (MCC)) were used to evaluate the performance of the algorithms. To calculate the number of misclassified records, we employed a confusion matrix.

Firstly, we considered the entire set of features (Table 3) to predict the type of occurred incident or accident, but the obtained accuracy was not completely satisfactory (e.g., the RF model produced an accuracy equal to 75%). By excluding the features of "ID", "country", and "department", we achieved the results summarised in Table 4. The RF model outperforms the others considering all six metrics.

Table 4: Performance metrics for the prediction of "Type of incident or accident"

Model	AUC	CA	F1	Precision	Recall	MCC
RF	**0.997**	**0.962**	**0.961**	**0.963**	**0.962**	**0.948**
LR	0.993	0.958	0.958	0.960	0.958	0.943
NN	0.996	0.952	0.951	0.951	0.952	0.934
kNN	0.953	0.821	0.813	0.823	0.821	0.751

By focusing on the records misclassified by the RF model, we noticed that the incidents and accidents incorrectly labelled were identified as "caught in" in 40% of cases, and as "struck by" in 28% of cases. To identify possible reasons behind these incorrect predictions, we implemented the widget "feature explanation" in Orange. This revealed that the primary feature employed for predicting the type of event is the "details" feature, as reported in Figure 4. The problem of misclassification is linked to the fact that the labels "caught in" and "struck by" have the same details and the same trigger. The model cannot predict the correct type of incident or accident because, in the description, equal information on different types of incidents and accidents is held.

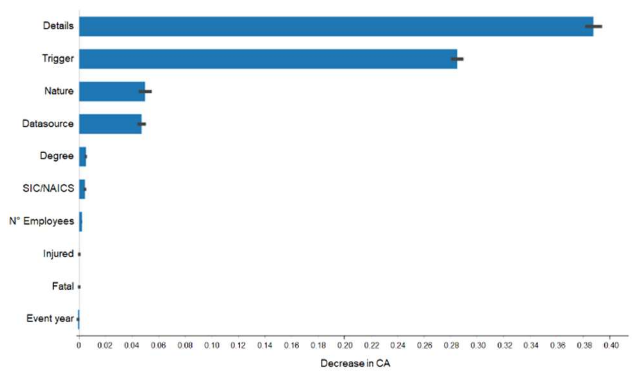

Figure 4: Feature importance for "Type of incident or accident"

An analysis about the correlation between the type of incident or accident and the SIC, NAICS, or NAF codes indicated that the highest frequency of events is linked to the SIC code 3321 (i.e., gray and ductile iron foundries). This code refers to the part of plant related to steelmaking process that involves the handling of hazardous materials, high temperatures, and complex machinery. Indeed, these factors contribute to an increased probability of incidents and accidents.

3.2.2 Predictions for the degree of consequences

For the prediction of the "degree" of consequences caused by the event, we employed LR, RF, NN, and kNN models. To evaluate the performance of these algorithms, we used the same six metrics mentioned in paragraph 3.2.1. As in the previous case, the models did not produce satisfactory results when all the features were taken into account. For this reason, we excluded the features "ID" and "country". The results obtained are shown in Table 5: the LR model outperforms the others considering all the metrics.
By employing the confusion matrix, we were able to find out the records misclassified by the LR model. The results reveal that one record was incorrectly classified as "fatality", nine records as "non hospitalised injury", and 12 records as "hospitalised injury". We correlated the misclassified results of the "degree" with the "type of incident or accident" feature. This highlighted that a significant portion of the data categorised as "non hospitalised injury" were classified as "hospitalised injury" within the LR model. Similarly, the LR model inaccurately designated "fatality" data as "non hospitalised injury". It is worth noting that an inclination towards overestimation within the model is preferable over underestimation. Indeed, overestimating a model entails predicting a more severe outcome than it may truly be. Conversely, underestimating a model could lead to a lack of preparedness and a poor ability to address the actual severity of a scenario.
A further relevant result demonstrates that the incident or accident causing the highest number of fatalities is "caught in". In this case, we can observe that the results of the LR model are almost identical to the predictions.
Finally, we correlated the incident severity with the SIC, NAICS, and NAF codes. As expected, we can observe that, also in this case, the majority of fatal incidents and accidents are associated with the SIC code 3321.

Table 5: Performance metrics for the prediction of "Degree"

Model	AUC	CA	F1	Precision	Recall	MCC
LR	**0.996**	**0.982**	**0.982**	**0.982**	**0.982**	**0.969**
RF	0.997	0.970	0.969	0.970	0.970	0.949
NN	0.974	0.929	0.927	0.926	0.929	0.881
kNN	0.975	0.915	0.913	0.914	0.915	0.857

3.3 Limitations

The overall performance of the ML models cannot be deemed completely satisfactory. It is important to note that ML algorithms can make reliable predictions when trained on a large number of high-quality input data, so the data collection step is crucial. The data sources from which the records are extracted in this work were not completely structured according to the same features. Moreover, several incident and accident descriptions were incomplete and lacked some details, as well as some of them were difficult to interpret and ambiguous. Therefore, the accuracy and reliability of the prediction models based on ML algorithms could have been affected.
Another limitation concerns the exclusion of specific available data sources (i.e., the Italian National Institute for Insurance against Accidents at Work INAIL) due to their complex structure and for the language used to collect incidents and accidents. Indeed, several organisations collect events in national languages, and this limits the possibility to analyse further interesting incidents and accidents.

3.4 Future developments

Possible future developments of this study may concern the definition and selection of further interesting features focusing on other details of undesired events in the iron and steel industry.
Other ML models can be developed to carry out other predictions related to different aspects, e.g., to identify potential safety risks and proactively take corrective actions able to limit the occurrence of accidents.
In addition, the dataset could be extended by extracting further incidents and accidents from other relevant data sources (e.g., INAIL) written in languages other than English, also by implementing automatic language translation tools.

4. Conclusions

This article had the objective to learn from incidents and accidents occurred in the iron and steel industry worldwide between 1980 and 2022. This was carried out by creating a specific dataset and applying ML algorithms to build predictive models. We collected and systematically analysed 5630 records of undesired events from different international data sources (i.e., ARIA, OSHA, and CSB), and we structured them into a unique dataset thanks to the definition of 14 features that represent relevant dimensions for the analysis.
The construction of such a dataset represents one of the main contributions of our study: to the best of our knowledge, currently in the literature there is not a unique database containing data on incidents and accidents occurred in the iron and steel industry organised in a structured way. The development of a single database is useful for analysing correlations among accidents and incidents happened worldwide in this sector.
Furthermore, we trained and tested a set of well-known ML classification algorithms (i.e., LR, RF, k-NN, and ANN) to predict, knowing the scenario, the type of incident or accident occurred and the degree. Afterwards, we analysed the correlations between the incident or accident happened and the associated SIC, NAICS, or NAF codes.
The analysis of the scenarios reveals that the majority of incidents and accidents occurred in the iron and steel industry are classified as "caught in" and "struck by". In these types of incidents, the main causes are attributable to: (1) moving object, or (2) not lockout machine, or (3) wrong

procedures. The prediction yielded the same results as the analysis of the entire dataset.

The results of this analysis can be helpful for safety managers of the iron and steel industry to understand the main causes of incidents and accidents, and to identify measures and strategies able to mitigate risks and improve safety procedures. Some potential safety measures and barriers to prevent the occurrence of incidents or accidents may include incorporating physical barriers, machine guarding, or automated systems that restrict the movement of objects to designated areas.

Acknowledgements

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

References

Alzubi, J., Nayyar, A., and Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142.

Cooke, D.L. and Rohleder, T.R. (2006). Learning from incidents: from normal accidents to high reliability. *System Dynamics Review*, 22(3), 213-239.

Demsar, J., Curk, T., and Erjavec, A. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14, 2349-2353.

Dhalmahapatra, K., Shingade, R., Mahajan, H., Verma, A., and Maiti, J. (2019). Decision support system for safety improvement: An approach using multiple correspondence analysis, t-SNE algorithm and K-means clustering. *Computers Industrial Engineering*, 128, 277-289.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

He, K. and Wang, L. (2017). A review of energy use and energy-efficient technologies for the iron and steel industry. *Renewable and Sustainable Energy Reviews*, 70, 1022-1039.

ILO (International Labour Organization) (2005). *Code of practice on safety and health in the iron and steel industry*. ILO, Geneva. (No MEISI/2005/8)

INAIL (2024). 2024 Infor.Mo scheda 2024: Dinamiche infortunistiche nell'industria metalmeccanica. INAIL, Italy

ISO (International Organization for Standardization). (2022). *ISO/IED 23053:2022. Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).*

Jain, A.K. and Dubes R. (1978). Feature definition in pattern recognition with small sample size. *Pattern Recognition*, 10(2), 85-97.

Lindberg, A.-K., Hansson, S.O., Rollenhagen C. (2010). Learning from accidents - What more do we need to know?. *Safety Science* 48: 714-721.Nakano, F.K., Cerri, R., and Vens, C. (2020). Active Learning for Hierarchical Multi-Label Classification. *Data Mining and Knowledge Discovery*, 34, 1496-1530.

Pickvance, S. (2011). Health and Safety problems and patterns. *Encyclopaedia of Occupational Health and Safety* (Part XI, Chapter 73)

Provost, F. and Fawcett T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. *O'Reilly Media*, 2013.

Sarkar, S., Ejaz, N., and Maiti, J. (2018). Application of hybrid clustering technique for pattern extraction of accident at work: A case study of a steel industry. *Proceedings of the 4th IEEE International Conference on Recent Advances in Information Technology, RAIT 2018*. Dhanbad, 1-6.

Sarkar, S., Mainak, C., Sohit, N., and Maiti, J. (2019). Decision support system for prediction of occupational accident: A case study from a steel plant. Emerging Technologies in Data Mining and Information Security. *Advances in Intelligent Systems and Computing*, 813, 787-796.

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*, 125, (104616).

Song, S. (2018). Steel Manufacturing Incident Analysis and Prediction. *Journal of Safety, Health and Environmental Research*, 14(1), 331-336.

Stefana, E., Marciano, F., Paltrinieri, N., Cocca, P. (2024). 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.

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

World Steel Association (2023). *2023 World Steel in Figures*. World Steel Association, Belgium.