ISSN 1980-5411 (On-line version)
Production, 31, e20210048, 2021
DOI: 10.1590/0103-6513.20210048

Systematic Review

Data mining in occupational safety and health: a systematic mapping and roadmap

Beatriz Lavezo dos Reis[a]*, Ana Caroline Francisco da Rosa[a], Ageu de Araujo Machado[a], Simone Luzia Santana Sambugaro Wencel[a], Gislaine Camila Lapasini Leal[a], Edwin Vladimir Cardoza Galdamez[a], Rodrigo Clemente Thom de Souza[a,b]

[a]Universidade Estadual de Maringá, Maringá, PR, Brasil

[b]Universidade Federal do Paraná, Jandaia do Sul, PR, Brasil

*bia.lavezo@gmail.com

Abstract

Paper aims: This research presents a literature overview in relation to data mining and machine learning applications in the area of occupational health and safety.

Originality: A summary of main insights obtained from the analysis of systematic mapping is presented at the end, as well as a roadmap with recommendations for directing future research on the topic.

Research method: This article carries out a thorough descriptive research of the scientific literature on the topic through a systematic mapping covering the period between the years 2008 and 2019 and 12 scientific databases, which at the end presents 68 selected records.

Main findings: Around 84% of the selected records were of total significance for the research, with the majority of them being classified in the areas of civil construction and steel industry.

Implications for theory and practice: Through this study it is possible to understand the way research has been developed on this theme, as well as point to the guidelines for future studies. Other contribution is the indication of studies in OSH 4.0 concept, based on monitoring workers full-time.

Keywords

Machine learning. Safety and health at work. Occupational accidents.

How to cite this article: Reis, B. L., Rosa, A. C. F., Machado, A. A., Wencel, S. L. S. S., Leal, G. C. L., Galdamez, E. V. C., & Souza, R. C. T. (2021). Data mining in occupational safety and health: a systematic mapping and roadmap. *Production*, *31*, e20210048. https://doi.org/10.1590/0103-6513.20210048.

Received: May 28, 2021; Accepted: Sept. 21, 2021.

1. Introduction

Market developments from industrial revolutions have provided numerous changes inside companies in general, as well as in labor environments, conditions and legislation. At the beginning of industrialization, companies were not required to have commitment to the safety and health of workers. Yet, this issue came to be considered employers' obligation and a competitive strategy for organizations (Ciarapica & Giacchetta, 2009). Therefore, it is essential to study and understand the concepts related to occupational health and safety, presenting an effective and simplified approach for industrial applications (Sanni-Anibire et al., 2020).

The term occupational safety and health (OSH) is related to mitigation and prevention of accidents and diseases that affect individuals considering the work they perform. This area requires attention since, according to the International Labor Organization (ILO), every second about 10 workers have an accident and each year 2.34 million employees die around the world due to accidents and occupational or professional diseases. Thus, health and safety are directly related to the social development of organizations and countries (Chen et al., 2020).

There are several approaches to discuss OSH-related factors. Yoon et al. (2013) link the emergence of the OSH management system to the United Kingdom in 1991, when a guide was developed to assist employees in improving health and safety in organizations. Hicks et al. (2016) present a study on the influence that occupational stress has on organizations' security environment, and Sanni-Anibire et al. (2020) assess risks in civil construction to improve workers' performance in relation to safety. Kang & Ryu (2019) use data mining (DM) applied to accident data to predict and classify these episodes, associating accidents to climatic conditions.

Given the volume of data generated from accidents, diseases, deaths and occurrences related to workers' health and safety, DM and machine learning (ML) are fundamental resources to take actions in this area. The concept of DM is described as part of the KDD (Knowledge Discovery in Databases) process and is responsible for extracting patterns from the data (Fayyad et al., 1996). This definition is directly linked to ML and sometimes the terms are mixed up. Yet, ML is usually more related to learning the algorithm, which occurs from the data used for its training (Buczak & Guven, 2016).

Every day more data are generated related to health and safety at work, and in the literature there are studies that make the interaction of these data with mining techniques and machine learning; Kakhki et al. (2019) compared the performance of four techniques in DM on labor claims; Baghdadi (2018) used sensors to collect kinematic data from workers and evaluate them with DM techniques; and Siddula et al. (2016) described an analysis of construction sites safety using images of the place.

The application of mining methods to OSH data is also a tool to assist organizations managers. The monitoring of employees by a coordinator is a possibility to reduce accidents (Antwi-Afari et al., 2018; Yanar et al., 2019). Furthermore, evaluating results and correlations presented by data mining serves as a subsidy for management's strategic decision making (Bevilacqua et al., 2008; Del Pozo-Antúnez et al., 2018) or can be used to establish new policies aiming at workers' health (Comberti et al., 2018; Liao & Perng, 2008).

This study seeks to answer the following question:

"How does data mining support decision-making in OSH?"

The article aims to provide an overview of the literature through systematic mapping, highlighting primary and quality studies involving the combination of DM and OSH themes, and defining future directions of research based on the gaps found.

2. Materials and methods

The Systematic Literature Mapping presents the purpose of carrying out the classification and analysis of the literature in relation to a topic, which provides a general view of the studies and their respective results. In this way, Systematic Literature Mapping guides the researcher to seek a holistic view instead of just answering a question in detail (Kitchenham et al., 2011).

In terms of structuring, Systematic Literature Mapping is arranged in three phases: Planning (Input), Execution (Processing) and Discussion of Results (Output). The Planning stage included the elaboration of the research protocol which covered the research questions, the search string and search sources, selection strategy, extraction strategy as well as quality assessment. The processing stage consisted of performing searches in the databases, classification, ordering and quality assessment, while the stage named discussion presented the synthesis of the results obtained.

The research protocol was defined based on the guidelines presented by Dybå et al. (2007), Paternoster et al. (2014) and Petersen et al. (2015). Moreover, the protocol was evaluated by three experts who suggested arrangements related to search string and insertion of databases. To conduct the systematic mapping, Microsoft Excel and Mendeley software programs were used as support tools.

Five questions related to the objective were defined to direct information about each selected article, which were:

(i) What kind of OSH data are explored?

(ii) What types of DM tasks, techniques and tools are used?

(iii) What industrial activity sector is explored in the research?

(iv) Which OSH database was used?

(v) Does the study use OSH data in a way related to other information?

Search strategies involve the definition of the sources that will be used, listing 12 sources for searches as shown in Table 1, with their respective results after the first search. As the study is described by a multidisciplinary theme, the data sources chosen are linked to different areas, some focusing on health and others more related to research in the areas of computing and information technology.

Table 1. Relationship between researched data sources and their results.

Data source	Search results	Selected articles
SpringerLink ™	509	7
ACM Digital Library ™	296	0
Scopus ™	131	22
ProQuest ™	123	4
Cambridge Journals ™	122	0
IEEE Xplore ™	89	12
Engineering Village ™	39	0
PubMed ™	36	2
ScienceDirect ™	32	20
EBSCO HOST ™	24	0
IngentaConnect ™	13	1
MathSciNet ™	10	0
Total	1424	68

Three groups of words were fixed, one related to health and safety; the other related to work; and the last one considering DM. From the combination of the key terms, the developed string is: (injury* OR health* OR safety* OR accident*) AND (work* OR labour* OR labor* OR occupational) AND ("data mining"* OR "machine learning"*).

The inclusion criteria adopted met the following restrictions: Published from 2008; English-language publications; and Evidence OSH in DM application. Exclusion criteria were considered: Not peer reviewed; No evidence of OSH in DM application; Does not respond satisfactorily to research questions; In case of duplicate articles, keep only the most complete one. Figure 1 highlights the results obtained in conducting systematic mapping.

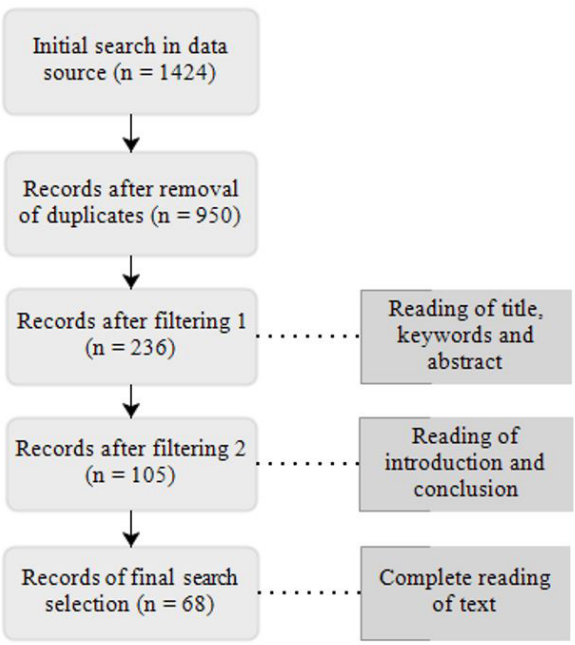

Figure 1. Flowchart of the systematic mapping execution process.

The search resulted in 1424 articles, from which 474 duplicate papers were excluded, reducing the sample to approximately 67% of the initial amount. After the first filtering, 236 records remained; with the second filtering, the number dropped to 105; and after the final selection there were 68 remaining articles. These selected articles are related to their original search sources in Table 1. In cases where there was duplicity of articles in different sources, only one of them was considered.

To classify the research four aspects were created, each one presenting its respective categories, as shown in Table 2, based on the study by Paternoster et al. (2014).

Table 2. Aspects and categories to classify the studies.

Aspect	Category
Data	Typical work acidentes
	Occupational diseases
	Fatalities
Learning	Supervised learning
	Unsupervised learning
Focus	Discovery process
	Compliance check
	Variance analysis of processes
	Predictive monitoring
Significance	Total
	Partial
	Marginal

According to Kitchenham (2004), evaluate the quality of the chosen records makes it possible to assess the importance of each article as well as facilitate the interpretation of their result. Therefore, to guide the quality assessment in this research, 11 questions were used through a binary scale ("yes" or "no") (Dybå et al., 2007).

Each question is related to a category and two classes are presented to adjust the articles: rigor and relevance. The rigor category, represented by eight questions, is related to the research methods used, answering whether the approach used was complete and covered all important aspects of the research. Relevance, represented by two questions, evaluates whether the results are clearly described and significant for the research. In the last question, it is also considered if the research is important in the academic and industrial scenario (Dybå et al., 2007).

3. Results

Based on the 68 articles selected for the research, it is possible to carry out a descriptive analysis of the records and all the selected articles are detailed in Appendix A. The first analysis, represented by Figure 2, is related to the year the articles on this topic were published.

It is possible to observe an increase in the number of studies published over the years. Their publication, which was stable until 2014, had a growth in the following years and in 2018 the number was more than twice that from the previous years. It may be associated with the growth of research in DM. In this way, the last two years are representative and responsible for approximately 40% of the total sample of selected records, demonstrating the topicality, importance and relevance of this theme. Accidents are the main focus of the studies chosen, except for the years 2012 and 2013, when diseases represented the main interest.

3.1. Information extraction by the research questions

Regarding the data used in the chosen research studies, represent by the first question most of them represent typical accidents that occurred in the industry, with emphasis on specific accidents such as slipping, stumbling and falling (SSF) (Nenonen, 2013; Sarkar et al., 2019b), diseases (Krishna et al., 2015) and occupational injuries (Ciarapica & Giacchetta, 2009), or studies that focus on cases of death due to occupational accidents (Ruso & Stojanović, 2012; Shin et al., 2018; Shirali et al., 2018).

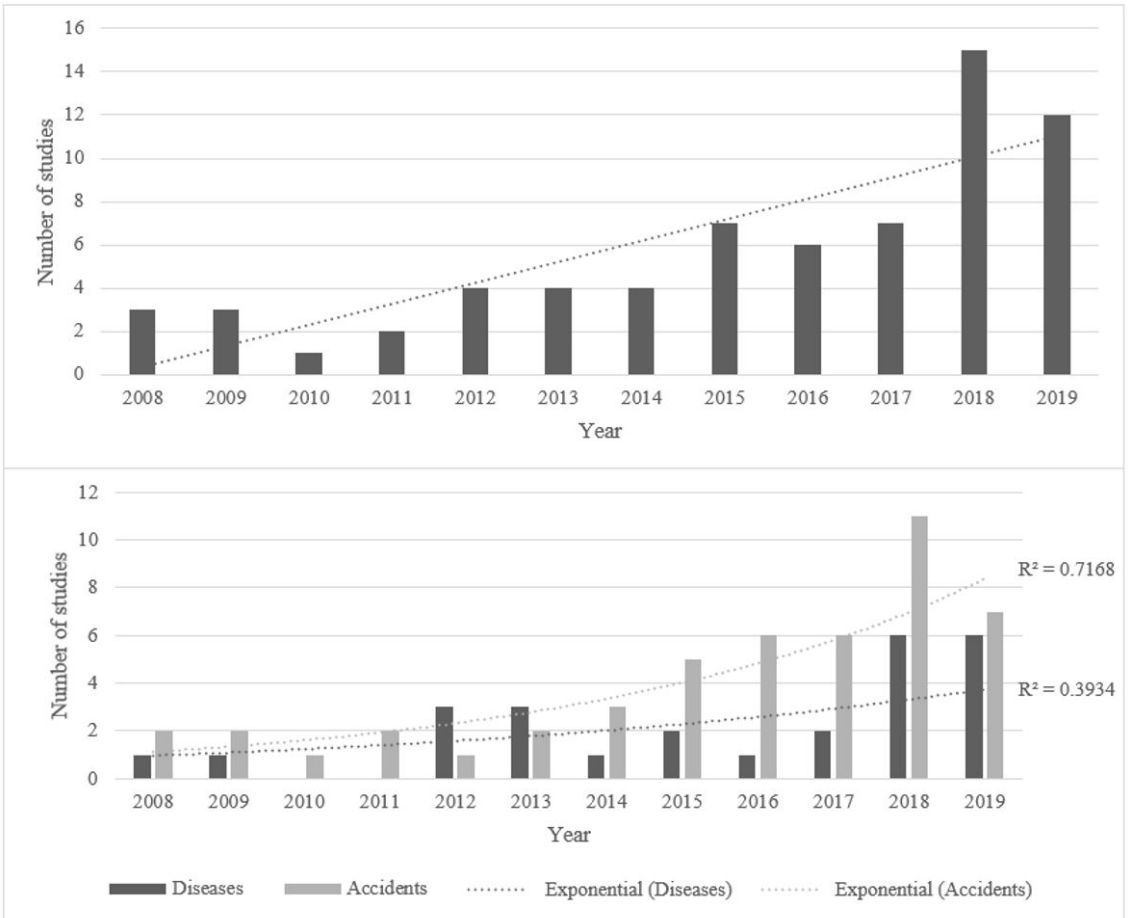

Figure 2. Temporal distribution and interest of the selected articles.

Other studies use benefit claim data sets to analyze absences from work (Bertke et al., 2012; Kakhki et al., 2019). Some of them use interviews with employees (Del Pozo-Antúnez et al., 2018; Zhao et al., 2019) or ergonomic tests to analyze performed activities (Baghdadi, 2018; Zhao et al., 2019). In the machine learning area, there are articles that use sensors to assess the worker (Xie & Chang, 2018), as well as photos and videos to detect the use of protective equipment and work postures (Rubaiyat et al., 2016; Shein et al., 2015; Siddula et al., 2016).

Related to second question, to characterize DM in OSH, the use of tasks, techniques and tools of the selected records were analyzed. To list the tasks used in DM, four types of tasks were considered; three associated with supervised learning - association, classification and regression - and one - clustering - linked to unsupervised learning.

Some studies describe more than one task in their scope and are classified according to the individual occurrences of each task, also according to the joining of two tasks in the same research. However, the data general analysis shows that most studies encompass the classification task, characterizing 78% of the chosen records. Clustering and association tasks are described in 15 and 8 studies, respectively. The least used task is regression, found in only five works, what represents about 7% of the selected sample. A single study uses it as the only task. Figure 3 shows the distribution of the time used in each task:

As for the techniques, those used in a higher number of studies were: decision tree, Support Vector Machine (SVM), naive Bayes and neural networks. Algorithms associated with decision tree, the most used technique, were observed in 20 studies, followed by SVM, applied in 17 articles. The use of algorithms associated to naïve Bayes and neural networks were observed in 14 studies, each. Other studies, in turn, presented the use of more

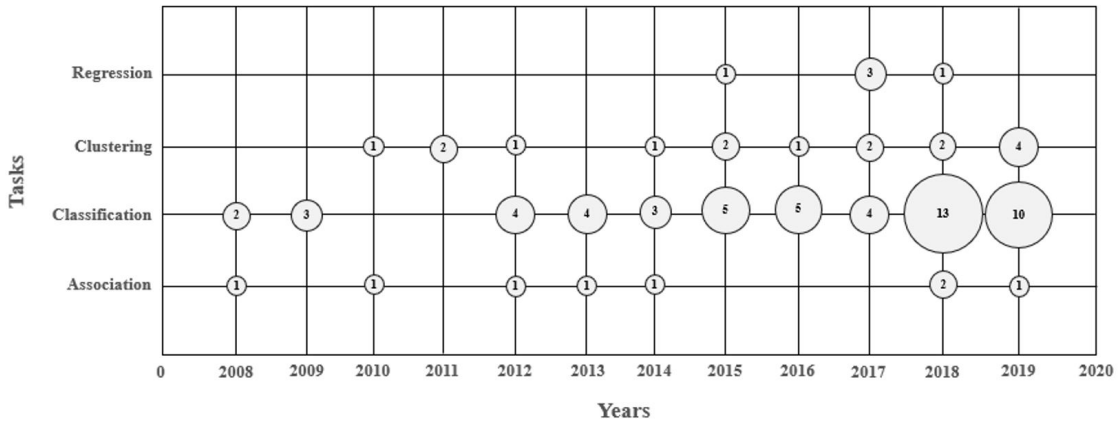

Figure 3. Temporal distribution of tasks used in the articles.

unusual techniques, occurred only once. For instance: EM algorithm, k-Bayesian (k-BNC), Hidden Markov Model (HMM), Maximum Entropy (MaxEnt), among others.

Considering the tools, their use was not clearly described in 31 of the articles studied. Only eight tools were used in more than one study (MATLAB, Weka, Clementine, Statistica Data Miner, Python, R, SAS e TextMiner), whereas the others were used in only one study. The most of the records chosen for systematic mapping had undefined software programs, or this tool was used only once.

In relation to the third question, industrial sectors, most of the selected studies were on the areas of civil construction and steel industry. Civil construction as the most researched sector, present in 19 articles and steel sector are explored by eight articles. Also representative are the healthcare sector (five studies), mining and petrochemicals (four), the administrative and timber sectors (three) and the agriculture category (two). In 14 studies there was not a categorization of the sector to which they belong, since the sector used was not specified, or even described as a general data set.

About the databases (fourth question), the analysis of the records selected for the systematic mapping of the literature, some of them publicly available, showed that the highest occurrence of the database was on the use of Occupational Safety and Health Administration (OSHA), presented in five studies, with information from the United Studies and South Korea.

The Council of Labor Affairs (Executive Yuan) of Taiwan and the Istituto Nazionale Assicurazione Infortuni sul Lavoro (INAIL), in Italy, come in second place, both presented in three studies. The Spanish Ministry of Employment and Social Safety and the Occupational Health Center of Presidente Prudente data sets are also noteworthy, for they were used in two studies. The other data sets are considered in only one study. Half of the studies represent internal data sets of the companies, tests carried out specifically for the article or data whose origin was not informed or specified.

To answer the last question of the research protocol, which considers the relationship between the selected studies and other information, some articles presenting other aspects related to OSH can be highlighted. Some of them can associate the occurrence of accidents with climatic conditions in the workplace (Bohanec & Delibašić, 2015), mainly for events in the construction industry (Kang & Ryu, 2019; Liao & Perng, 2008).

The costs that companies or government have with workers' health and safety are demonstrated in five of the articles selected (Cheng et al., 2012; Kakhki et al., 2019; Meyers et al., 2018; Olsen et al., 2009; Shin et al., 2018). This difficulty in finding research that addresses OSH costs represents a gap in the literature. It means more studies need to be developed.

3.2. Record classification and quality evaluation

The selected articles were stratified according to the aspects and categories defined in section 2 and the results of the classification are shown in Figure 4. Rigor represents the precision of the study in its research

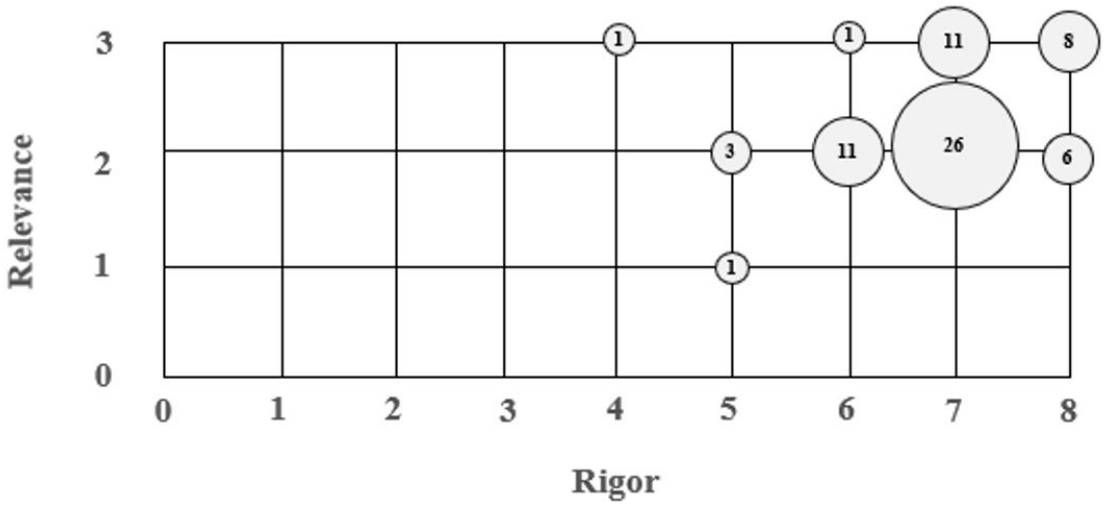

Figure 4. Evaluation of the rigor and relevance criteria.

method and the way the study is presented. Relevance represents the value of the study to the research community and industry.

This figure shows the incidence of studies on its upper right corner. It indicates they are characterized by high rigor and relevance/credibility. The higher results are represented by twenty-six studies that got marks 7 and 2 respectively in rigor and relevance and eleven studies that got 7-3. Low results are represented by one study with marks 4-3; one with 5-1 and three with 5-2. Finally, eight studies got the highest marks in both aspects (8-3).

The eight studies that got the maximum marks, thus considered the most rigorous with the most relevant themes, are represented by Abad et al. (2019), Antwi-Afari et al. (2018), Di Noia et al. (2019), Sarkar et al. (2018, 2019a, b), Xie & Chang (2018) and Zhao et al. (2019).

4. Discussion

4.1. Insights identified in the systematic mapping on DM in OSH

Some studies present publicly available data sets, such as those from the Council of Labor Affairs (Cheng et al., 2010, 2012, 2013) or INAIL (Comberti et al., 2015, 2018; Palamara et al., 2011). The mapping also found research that used tests for data generation (Antwi-Afari et al., 2018; Baghdadi, 2018; Olsen et al., 2009). Others used interviews as a data source (Del Pozo-Antúnez et al., 2018; Zhao et al., 2019) or medical examinations, such as audiometric analyzes (Tomiazzi et al., 2019; Zhao et al., 2019), radiographs (Heo et al., 2019; Waghmare & Pai, 2013) and blood tests (Abad et al., 2019; Waghmare & Pai, 2013).

The data used may have different characterizations, as some correspond to claims for benefits after an accident or diseases at work (Bertke et al., 2012; Kakhki et al., 2019; Gross et al., 2013; Meyers et al., 2018). Other studies focus on the accident data set in general (Bevilacqua et al., 2008; Sarkar et al., 2019a), specific accidents such as slipping, stumbling and falling (SSF) (Nenonen, 2013; Sarkar et al., 2019b), roof fall (Mistikoglu et al., 2015), or major accidents, which injure at least three people or cause one or more deaths (Cheng et al., 2013).

Regarding cases of death, some data presented sets of fatal and non-fatal accident cases (Jocelyn et al., 2018; Mistikoglu et al., 2015; Shin et al., 2018; Shirali et al., 2018), but there are also cases in which the set presents only accident events with death (Ruso & Stojanović, 2012). There are also studies that show severity levels of occurrences (Shirali et al., 2018), considering more and less serious accidents, or only less serious cases of occupational diseases (Krishna et al., 2015).

Another analysis that can be performed has to do with the industrial sector of workers who are affected by accidents or occupational diseases. Most research studies are concentrated in the construction industry and at different levels of this scenario, such as in landfills preparation steps to start the work (Gerassis et al., 2017), in construction dockyard (Kang & Ryu, 2019) or construction sites (Shin et al., 2018).

The steel industry also presents relevant research for systematic mapping (Krishna et al., 2015; Sarkar et al., 2017, 2018, 2019a, b, c; Shirali et al., 2018), as well as applications in the area of healthcare (Di Noia et al., 2019; Kao et al., 2018; Olsen et al., 2009; Saâdaoui et al., 2015; Ueno et al., 2008), mining (Luo et al., 2016; Qu, 2009; Sanmiquel et al., 2015, 2018) and petrochemical (Bevilacqua et al., 2008; Cheng et al., 2013; Sanchez-Pi et al., 2014; Waghmare & Pai, 2013).

Considering the DM techniques used, some studies present similar data using different techniques. For instance, two studies used data sets of accidents in civil construction, but with different techniques: decision tree and association rules, respectively. Related to this, there is a large number of articles that present the decision tree technique and its specific algorithms, such as C 4.5 (Gross et al., 2013; Sanmiquel et al., 2015; Shein et al., 2015) and C 5.0 (Hajakbari & Minaei-Bidgoli, 2014; Mistikoglu et al., 2015; Sarkar et al., 2018, 2019c).

Another widely used technique is Support Vector Machine (SVM), present in 17 studies of the selected sample, some applying only this technique (Rubaiyat et al., 2016; Siddula et al., 2016; Xie & Chang, 2018) or comparing with results of others (Goh & Ubeynarayana, 2017; Lee & Kim, 2018; Tomiazzi et al., 2018). Related to the clustering task, the k-means technique is also representative in the research (Bohanec & Delibašić, 2015; Chokor et al., 2016; Palamara et al., 2011). There are also applications of Bayesian networks (Jiang et al., 2018; Marucci-Wellman et al., 2017; Nanda et al., 2016) and neural networks (Ciarapica & Giacchetta, 2009; Del Pozo-Antúnez et al., 2018).

Concerning the application of techniques, there are many tools which perform and support DM methods. Some tools are easy to handle, with ready internal packages and part of their programming already developed, such as Weka (Gerassis et al., 2017; Jocelyn et al., 2018; Pekel et al., 2018; Sanmiquel et al., 2015; Waghmare & Pai, 2013) and (Heo et al., 2019; Sanmiquel et al., 2018). Other studies, in their turn, use Python language to elaborate the code for DM (Goh & Ubeynarayana, 2017; Heo et al., 2019; Marucci-Wellman et al., 2017) and present a higher complexity of running. Some software programs have their specific functions for a type of information, such as TextMiner for textual data sets (Nanda et al., 2016; Taylor et al., 2014).

As for the types of data that can go through the mining process, this sample includes research studies that use images (Rubaiyat et al., 2016; Siddula et al., 2016) and videos (Paliyawan et al., 2014; Ueno et al., 2008), being directly associated with the concept of machine learning and automation in the worker's environment. Other studies present textual data sets in formats of injuries (Tixier et al., 2017) and accidents reports (Liao & Perng, 2008; Sarkar et al., 2016) and medical examinations (Bonneterre et al., 2012).

Some research studies also use accident data related to other information to generate associations or enrich the models developed. Some seek to associate accidents with a certain location (Rashid et al., 2017; Valêncio et al., 2011), others associate the sources of injuries, accidents and deaths with climatic conditions (Bohanec & Delibašić, 2015; Kang & Ryu, 2019; Liao & Perng, 2008). Figure 5 shows the relationship between data, focus and type of learning.

As can be seen, 36 (52.94%) of the 68 articles studied referred to Typical Work Accidents, 20 (29.41%) were classified as Occupational Diseases, 6 (8.82%) were categorized as Accidents of Typical Work and Fatalities. Of the studies in question, 4 (5.88%) were simultaneously shown as Typical Work Accidents and Occupational Diseases and 2 (2.95%) could be classified as Typical Work Accidents, Occupational Diseases and Fatalities.

Regarding the Focus of the 36 studies classified as Typical Occupational Accident, 18 (50.00%) were shown as Process Analysis of Variance, 11 (30.55%) as Predictive Monitoring, 6 (16,67%) as Compliance Check and 1 (2.78%) as Variance Analysis of Processes and Predictive Monitoring.

4.2. Roadmap for recommending future research

From the systematic mapping of the literature in context of DM and OSH, it was possible to build a roadmap for future research in the area. Future research is based on data published on accidents at work by national and international agencies responsible for monitoring the financial, economic and productive impacts of such events. They are also built by the type of method used for DM and the type of results that the research will present to society. Based on these paths, it is possible for researchers and managers in the area to identify the studies that have already been carried out and what are the possibilities and combinations of new research.

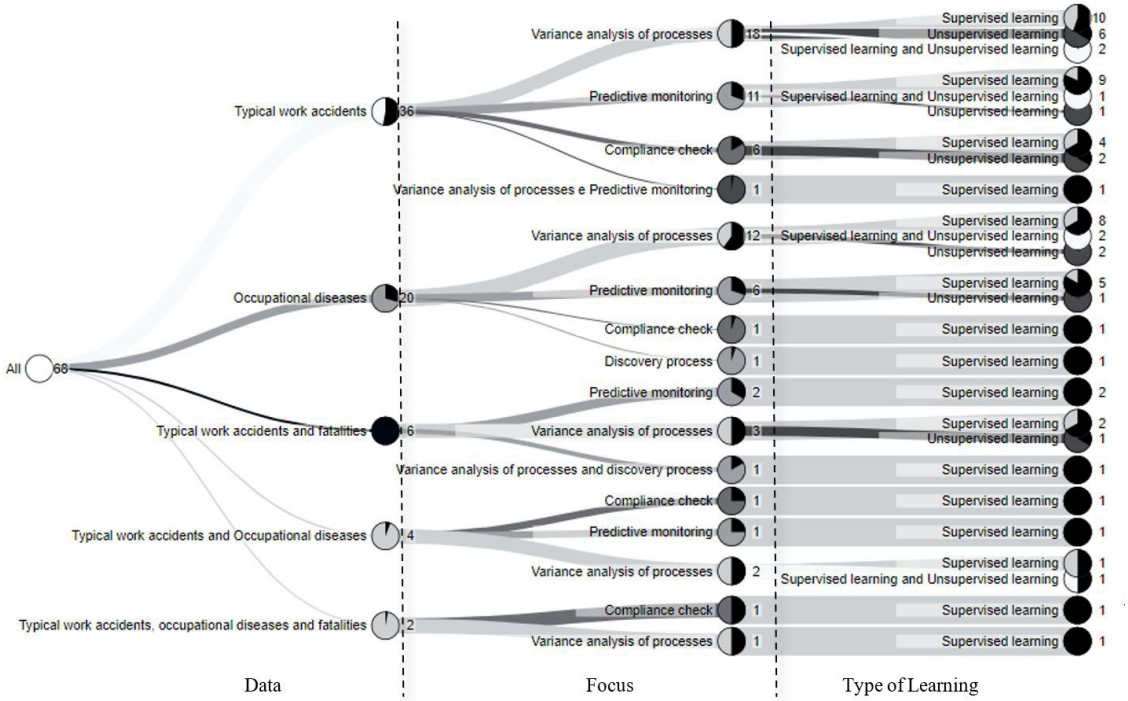

Figure 5. Relationship between Data, Focus and Type of Learning.

The characteristics of the roadmap (data, method and results) are validated based on the results described in the previous sections and presented, as well as their subdivisions, by Figure 6, which allow classifying the studies identified in the systematic mapping conducted by the researchers.

Quality, continuity and reliability of the databases selected for future research are critical attributes for research in the area, since they allow increasing the impact, the evaluation and the improvement of preventive actions or measures that can be adopted for Health and Safety policies or programs at work. There is a possibility to separate public, private and test data. Public data are those made available through organizations such as the Taiwan Labor Affairs Council, INAIL in Italy and Spanish Ministry of Employment and Social Security of Spain.

The use of public data made available by agencies in research in the area can result in benefits for operational safety and health. These advances are related to the creation of policies and the implementation of prevention programs, as well as the development of indexes that allow the monitoring of the accident scenario and assist in decision making. Some of these benefits can be found in the research results found by Cheng et al. (2012) and Marucci-Wellman et al. (2017) through proposed actions, or in Del Pozo-Antúnez et al. (2018) with a suggestion of organizational changes. Another advantage of using public data is the application in different contexts, because when using private data, in most cases, we are talking about specific industrial sectors, differently from what happens with public data, that are more comprehensive.

Private data are data provided by a specific company, such as, for example, data from an insurance company (Kakhki et al., 2019), petrochemical industry (Bevilacqua et al., 2008) or accidents at a ski resort (Bohanec & Delibašić, 2015). Using these data allows industrial sectors with lower rates to be studied in depth and in detail. Besides, private data sets can contain more specific information than public data in general. However, the data selection process can be a difficult stage, both for public and private organizations. Sometimes the data are not updated, they take time to be published, or there are still obstacles on the part of organizations, which are apprehensive about the exposure of internal and negative information.

Tests are specific observations for the study, such as ergonomic simulations (Antwi-Afari et al., 2018; Olsen et al., 2009) and risk situations of workers in their workplace (Rashid et al., 2017). There are also specific cuts that can be made in these data, such as, for example, temporal cuts, specific types of accidents, prioritized

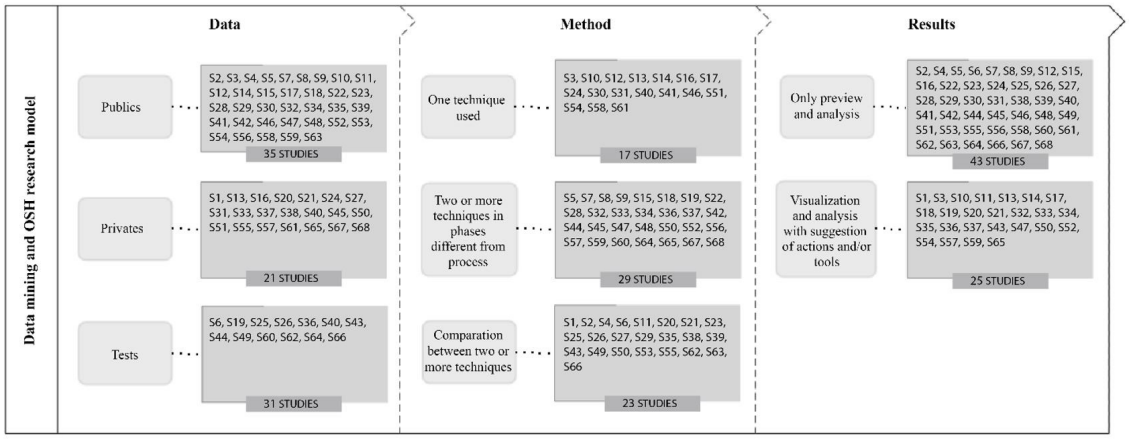

Figure 6. Roadmap for data mining in OSH.

industrial sector or other divisions that the researcher wishes to investigate in order to define preventive actions and specific programs for companies (Paliyawan et al., 2014; Sarkar et al., 2019b; Taylor et al., 2014).

As for the method used to apply data mining, the researcher may decide to choose a single technique to mine his data, as it happened in the research by Cheng et al. (2013), in which only the decision tree technique was used. In addition, two or more techniques can be used at different stages of the data mining process, as described in the research by Sarkar et al. (2019c) with the random forest technique for data preprocessing, SVM and artificial neural networks (ANN) in a second step and decision tree to complete data mining. Another possibility of methodology for conducting the research is to make a comparison between two or more techniques, analyzing which one performs the best result for the chosen data set, as in Kakhki et al. (2019) comparing the performance of five classification algorithms (linear and quadratic SVM, RBF kernels, Boosted Trees and Naïve Bayes).

Selecting different techniques and comparing the results of their applications allows the researcher to make decisions about the best mining strategy based on performance of algorithms, which makes their choice more reliable. The choice of using several methods at different stages of the process makes data mining more robust and also more reliable, since the model will be composed of different algorithms with specific focuses. In both cases, the methods allow for a more accurate result, reflecting in better performance in systems of accident prevention and mitigation. On the other hand, because they are more complex processes and with more steps involved, they can result in more time for planning and execution by researchers, demanding more investments and human resources to analyze research problems.

Regarding the results generated by the studies conducted, they can be defined by two complementary approaches. The first option is to present a visualization and analysis of the data, and the second is to complement the analytical step with preventive actions or tools that are proposed based on the results found during the research. The first case is commonly found in studies with only the presentation of the analyzes carried out from the mining results, as it is presented, for example, in studies by Ciarapica & Giacchetta (2009), Bohanec & Delibašić (2015) and Zhao et al. (2019).

The presentation and analysis of data mining allows us to understand and evaluate the method used, its results and the research scenario, but it is possible to go further. Some authors, besides presenting the analytical stage, also highlight actions that can be taken and present tools that can be used in this context, as it is the case of the study by Cheng et al. (2013), who, based on the associations made by algorithm, presented actions and areas of attention to reduce accidents in the industrial sector. Hajakbari & Minaei-Bidgoli (2014), in addition to presenting a guide and locations for future inspections based on the data used, also present a scoring system to prioritize the workplaces to be inspected, showing all stages, in a way that other researchers can use this tool in other contexts.

5. Threats to validity

Threats to validity exist in all empirical studies (Petersen et al., 2015). The construction and definition of the search string is one of the main challenges in defining the protocol. To minimize threats, the search string

for the mapping was developed based on consulting other studies and specialists in the area. Furthermore, the protocol was evaluated by three experts, and their suggestions were accepted and included.

The threats related to the selection of studies were mitigated through inclusion and exclusion criteria (Keele, 2007; Petersen et al., 2015) plus the fact that the research was carried out by four researchers, with supervision of three more specialists, which minimizes bias in selection due to personal opinions. Besides, the studies that generated uncertainty were discussed with the group of researchers. To reduce threats related to quality assessment, a binary scale was adopted instead of assigning any punctuation to the studies (Afzal et al., 2009).

6. Final considerations

In order to provide an overview of the literature on issues of data mining and occupational health and safety, this study presented a systematic mapping, which evaluated about 12 databases of research studies related to theme. A total of 68 records were selected, seeking to understand how data mining can help decision making in OSH. For this, the articles were classified considering their data, type of learning, focus and significance. Furthermore, they were evaluated for quality taking into account rigor and relevance of each selected research.

With the execution of a systematic mapping of the literature, it was possible to identify that most studies describe typical accidents (71%), as well as the number of occurrences related to sectors of civil construction (28%) and steel industry (12%). The techniques associated with classification tasks were the most chosen ones for application of data mining. Decision trees, SVM, Bayesian networks and neural networks were the most recurrent. Most of these studies (46%) do not indicate which tool was used in their data mining, but as for the ones that do it, the most found tools were MATLAB and Weka.

Research using DM, interacting with OSH, is an opportunity for academia and industry. The use of computational techniques reduces time for analyzing information and, consequently, reduces resources, facilitating studies in OSH. In addition to assisting in evaluation of information, results of data mining applications are subsidies for decision-making. Governmental and regulatory institutions can use this information to create new regional and global policies, and private institutions can develop internal standards.

The results can also foster discussions to encourage the emergence of new practices to reduce the risk of accidents and diseases. In addition, the use of sensors, videos and images can help monitor the workplace and conditions that workers are submitted to, even from a distance, or block some dangerous act, preventing the occurrence of accidents. Monitoring costs with accidents can also be a viable action for managers, who can use these financial expenses as indicators of efficiency in new OSH policies, or to justify implementing new techniques to reduce the possibility of accidents and diseases.

Considering norms and policies that indicate the tolerance levels for work environment, such as noise, light, temperature, among others, machine learning and data mining could also be used to monitor these levels, avoiding discomfort for the worker. The concept of OSH 4.0 is related to the idea of full-time monitoring of workers and their jobs. This practice is concerned with the way health and safety management will follow the changes of the fourth industrial revolution (Badri et al., 2018), which should also be a focus of further studies in this area.

References

Abad, A., Gerassis, S., Saavedra, Á., Giráldez, E., García, J. F., & Taboada, J. (2019). A Bayesian assessment of occupational health surveillance in workers exposed to silica in the energy and construction industry. *Environmental Science and Pollution Research International*, *26*(29), 29560-29569. http://dx.doi.org/10.1007/s11356-018-2962-6. PMid:30121763.

Afzal, W., Torkar, R., & Feldt, R. (2009). A systematic review of search-based testing for non-functional system properties. *Information and Software Technology*, *51*(6), 957-976. http://dx.doi.org/10.1016/j.infsof.2008.12.005.

Akboğa, Ö., & Baradan, S. (2017). Safety in ready mixed concrete industry: descriptive analysis of injuries and development of preventive measures. *Industrial Health*, *55*(1), 54-66. http://dx.doi.org/10.2486/indhealth.2016-0083. PMid:27524105.

Antwi-Afari, M. F., Li, H., Yu, Y., & Kong, L. (2018). Wearable insole pressure system for automated detection and classification of awkward working postures in construction workers. *Automation in Construction*, *96*, 433-441. http://dx.doi.org/10.1016/j.autcon.2018.10.004.

Badri, A., Boudreau-Trudel, B., & Souissi, A. S. (2018). Occupational health and safety in the industry 4.0 era: a cause for major concern? *Safety Science*, *109*, 403-411. http://dx.doi.org/10.1016/j.ssci.2018.06.012.

Baghdadi, A. (2018). Application of inertial measurement units for advanced safety surveillance system using individualized sensor technology (ASSIST): a data fusion and machine learning approach. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 450-451). New York: IEEE. http://dx.doi.org/10.1109/ICHI.2018.00097.

Bertke, S. J., Meyers, A. R., Wurzelbacher, S. J., Bell, J., Lampl, M. L., & Robins, D. (2012). Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. *Journal of Safety Research*, *43*(5-6), 327-332. http://dx.doi.org/10.1016/j.jsr.2012.10.012. PMid:23206504.

Bevilacqua, M., Ciarapica, F. E., & Giacchetta, G. (2008). Industrial and occupational ergonomics in the petrochemical process industry: a regression trees approach. *Accident; Analysis and Prevention*, *40*(4), 1468-1479. http://dx.doi.org/10.1016/j.aap.2008.03.012. PMid:18606280.

Bohanec, M., & Delibašić, B. (2015). Data-mining and expert models for predicting injury risk in ski resorts. *Lecture Notes in Business Information Processing*, *216*, 46-60. http://dx.doi.org/10.1007/978-3-319-18533-0_5.

Bonneterre, V., Bicout, D. J., & De Gaudemaris, R. (2012). Application of pharmacovigilance methods in occupational health surveillance: comparison of seven disproportionality metrics. *Safety and Health at Work*, *3*(2), 92-100. http://dx.doi.org/10.5491/SHAW.2012.3.2.92. PMid:22993712.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys and Tutorials*, *18*(2), 1153-1176. http://dx.doi.org/10.1109/COMST.2015.2494502.

Chen, H., Hou, C., Zhang, L., & Li, S. (2020). Comparative study on the strands of research on the governance model of international occupational safety and health issues. *Safety Science*, *122*, 104513. http://dx.doi.org/10.1016/j.ssci.2019.104513.

Cheng, C.-W., Leu, S.-S., Cheng, Y.-M., Wu, T.-C., & Lin, C.-C. (2012). Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. *Accident; Analysis and Prevention*, *48*, 214-222. http://dx.doi.org/10.1016/j.aap.2011.04.014. PMid:22664684.

Cheng, C.-W., Lin, C.-C., & Leu, S.-S. (2010). Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science*, *48*(4), 436-444. http://dx.doi.org/10.1016/j.ssci.2009.12.005.

Cheng, C.-W., Yao, H.-Q., & Wu, T.-C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, *26*(6), 1269-1278. http://dx.doi.org/10.1016/j.jlp.2013.07.002.

Chokor, A., Naganathan, H., Chong, W. K., & Asmar, M. E. (2016). Analyzing Arizona OSHA injury reports using unsupervised machine learning. *Procedia Engineering*, *145*, 1588-1593. http://dx.doi.org/10.1016/j.proeng.2016.04.200.

Ciarapica, F. E., & Giacchetta, G. (2009). Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. *Safety Science*, *47*(1), 36-49. http://dx.doi.org/10.1016/j.ssci.2008.01.006.

Comberti, L., Baldissone, G., & Demichela, M. (2015). Workplace accidents analysis with a coupled clustering methods: S.O.M. and K-means algorithms. *Chemical Engineering Transactions*, *43*, 1261-1266. http://dx.doi.org/10.3303/CET1543211.

Comberti, L., Demichela, M., & Baldissone, G. (2018). A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making. *Safety Science*, *106*, 191-202. http://dx.doi.org/10.1016/j.ssci.2018.03.014.

Del Pozo-Antúnez, J. J., Ariza-Montes, A., Fernández-Navarro, F., & Molina-Sánchez, H. (2018). Effect of a job demand-control-social support model on accounting professionals' health perception. *International Journal of Environmental Research and Public Health*, *15*(11), 2437. http://dx.doi.org/10.3390/ijerph15112437. PMid:30388812.

Di Noia, A., Martino, A., Montanari, P., & Rizzi, A. (2019). Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing*, *24*, 4393-4406. http://dx.doi.org/10.1007/s00500-019-04200-2.

Dybå, T., Dingsøyr, T., & Hanssen, G. K. (2007). Applying systematic reviews to diverse study types: an experience report. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)* (pp. 225-234). New York: IEEE. http://dx.doi.org/10.1109/ESEM.2007.59.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-54. http://dx.doi.org/10.1609/aimag.v17i3.1230.

Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. *Journal of Construction Engineering and Management*, *143*(2), 04016093. http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0001225.

Goh, Y. M., & Ubeynarayana, C. U. (2017). Construction accident narrative classification: an evaluation of text mining techniques. *Accident; Analysis and Prevention*, *108*, 122-130. http://dx.doi.org/10.1016/j.aap.2017.08.026. PMid:28865927.

Gross, D. P., Zhang, J., Steenstra, I., Barnsley, S., Haws, C., Amell, T., McIntosh, G., Cooper, J., & Zaiane, O. (2013). Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers. *Journal of Occupational Rehabilitation*, *23*(4), 597-609. http://dx.doi.org/10.1007/s10926-013-9430-4. PMid:23468410.

Hajakbari, M. S., & Minaei-Bidgoli, B. (2014). A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with Iran's Ministry of Labor data. *Journal of Loss Prevention in the Process Industries*, *32*, 443-453. http://dx.doi.org/10.1016/j.jlp.2014.10.013.

Heo, S.-J., Kim, Y., Yun, S., Lim, S.-S., Kim, J., Nam, C.-M., Park, E.-C., Jung, I., & Yoon, J.-H. (2019). Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. *International Journal of Environmental Research and Public Health*, *16*(2), 250. http://dx.doi.org/10.3390/ijerph16020250. PMid:30654560.

Hicks, G., Buttigieg, D., & De Cieri, H. (2016). Safety climate, strain and safety outcomes. *Journal of Management & Organization*, *22*(1), 19-31. http://dx.doi.org/10.1017/jmo.2015.45.

Jiang, H., Cai, Y., Zeng, X., & Huang, M. (2018). Does background really matter? Worker activity recognition in unconstrained construction environment. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (pp. 50-53). New York: IEEE. http://dx.doi.org/10.1109/BSN.2018.8329656.

Jocelyn, S., Ouali, M.-S., & Chinniah, Y. (2018). Estimation of probability of harm in safety of machinery using an investigation systemic approach and Logical Analysis of Data. *Safety Science*, *105*, 32-45. http://dx.doi.org/10.1016/j.ssci.2018.01.018.

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, *117*, 257-262. http://dx.doi.org/10.1016/j.ssci.2019.04.026.

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. *Safety Science*, *120*, 226-236. http://dx.doi.org/10.1016/j.ssci.2019.06.034.

Kao, H., Hosseinmardi, H., Yan, S., Hasan, M., Narayanan, S., Lerman, K., & Ferrara, E. (2018). Discovering latent psychological structures from self-report assessments of hospital workers. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)* (pp. 156-161). New York: IEEE. http://dx.doi.org/10.1109/BESC.2018.8697325.

Keele, S. (2007). *Guidelines for performing systematic literature reviews in software engineering: technical report, version 2.3* (EBSE Technical Report). Durham: EBSE.

Khosrowabadi, N., & Ghousi, R. (2019). Decision support approach to occupational safety using data mining. *International Journal of Industrial Engineering & Production Research*, *30*(2), 149-164. http://dx.doi.org/10.22068/ijiepr.30.2.149.

Kitchenham, B. (2004). *Procedures for performing systematic reviews* (Vol. *33*, pp. 1-26). Keele: Keele University.

Kitchenham, B., Budgen, D., & Brereton, P. (2011). Using mapping studies as the basis for further research: a participant-observer case study. *Information and Software Technology*, *53*(6), 638-651. http://dx.doi.org/10.1016/j.infsof.2010.12.011.

Krishna, O. B., Maiti, J., Ray, P. K., & Mandal, S. (2015). Assessment of risk of musculoskeletal disorders among crane operators in a steel plant: a data mining-based analysis. *Human Factors and Ergonomics in Manufacturing*, *25*(5), 559-572. http://dx.doi.org/10.1002/hfm.20575.

Lee, J., & Kim, H.-R. (2018). Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques. *Journal of Korean Medical Science*, *33*(19), e144. http://dx.doi.org/10.3346/jkms.2018.33.e144. PMid:29736160.

Liao, C.-W., & Perng, Y.-H. (2008). Data mining for occupational injuries in the Taiwan construction industry. *Safety Science*, *46*(7), 1091-1102. http://dx.doi.org/10.1016/j.ssci.2007.04.007.

Luo, X., Yang, X., Wang, W., Chang, X., Wang, X., & Zhao, Z. (2016). A novel hidden danger prediction method in cloud-based intelligent industrial production management using timeliness managing extreme learning machine. *China Communications*, *13*(7), 74-82. http://dx.doi.org/10.1109/CC.2016.7559078.

Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance: a practical approach combining machine learning ensembles and human review. *Accident; Analysis and Prevention*, *98*, 359-371. http://dx.doi.org/10.1016/j.aap.2016.10.014. PMid:27863339.

Meyers, A. R., Al-Tarawneh, I. S., Wurzelbacher, S. J., Bushnell, P. T., Lampl, M. P., Bell, J. L., Bertke, S. J., Robins, D. C., Tseng, C.-Y., Wei, C., Raudabaugh, J. A., & Schnorr, T. M. (2018). Applying machine learning to workers' compensation data to identify industry-specific ergonomic and safety prevention priorities: Ohio, 2001 to 2011. *Journal of Occupational and Environmental Medicine*, *60*(1), 55-73. http://dx.doi.org/10.1097/JOM.0000000000001162. PMid:28953071.

Mistikoglu, G., Gerek, I. H., Erdis, E., Mumtaz Usmen, P. E., Cakan, H., & Kazan, E. E. (2015). Decision tree analysis of construction fall accidents involving roofers. *Expert Systems with Applications*, *42*(4), 2256-2263. http://dx.doi.org/10.1016/j.eswa.2014.10.009.

Nanda, G., Grattan, K. M., Chu, M. T., Davis, L. K., & Lehto, M. R. (2016). Bayesian decision support for coding occupational injury data. *Journal of Safety Research*, *57*, 71-82. http://dx.doi.org/10.1016/j.jsr.2016.03.001. PMid:27178082.

Nenonen, N. (2013). Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. *Applied Ergonomics*, *44*(2), 215-224. http://dx.doi.org/10.1016/j.apergo.2012.07.001. PMid:22877702.

Olsen, G. F., Brilliant, S. S., Primeaux, D., & Najarian, K. (2009). Signal processing and machine learning for real-time classification of ergonomic posture with unobtrusive on-body sensors; application in dental practice. In *2009 ICME International Conference on Complex Medical Engineering* (pp. 1-11). New York: IEEE. http://dx.doi.org/10.1109/ICCME.2009.4906675.

Palamara, F., Piglione, F., & Piccinini, N. (2011). Self-organizing map and clustering algorithms for the analysis of occupational accident databases. *Safety Science*, *49*(8-9), 1215-1230. http://dx.doi.org/10.1016/j.ssci.2011.04.003.

Paliyawan, P., Nukoolkit, C., & Mongkolnam, P. (2014). Office workers syndrome monitoring using kinect. In *The 20th Asia-Pacific Conference on Communication (APCC2014)* (pp. 58-63). New York: IEEE. http://dx.doi.org/10.1109/APCC.2014.7091605.

Paternoster, N., Giardino, C., Unterkalmsteiner, M., Gorschek, T., & Abrahamsson, P. (2014). Software development in startup companies: a systematic mapping study. *Information and Software Technology*, *56*(10), 1200-1218. http://dx.doi.org/10.1016/j.infsof.2014.04.014.

Pekel, E., Akschir, Z. D., Meto, B., Akleylek, S., & Kilic, E. (2018). A Bayesian network application in occupational health and safety. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 239-243). New York: IEEE. http://dx.doi.org/10.1109/UBMK.2018.8566568.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: an update. *Information and Software Technology*, *64*, 1-18. http://dx.doi.org/10.1016/j.infsof.2015.03.007.

Qu, Z. (2009). Application of data mining in classification analysis of safety accidents based on alternate covering neural network. In *2009 International Conference on Future BioMedical Information Engineering (FBIE)* (pp. 144-147). New York: IEEE. http://dx.doi.org/10.1109/FBIE.2009.5405861.

Rashid, K. M., Datta, S., & Behzadan, A. H. (2017). Coupling risk attitude and motion data mining in a preemtive construction safety framework. In *2017 Winter Simulation Conference (WSC)* (pp. 2413-2424). New York: IEEE. http://dx.doi.org/10.1109/WSC.2017.8247971.

Rubaiyat, A. H. M., Toma, T. T., Kalantari-Khandani, M., Rahman, S. A., Chen, L., Ye, Y., & Pan, C. S. (2016). Automatic detection of helmet uses for construction safety. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)* (pp. 135-142). New York: IEEE. http://dx.doi.org/10.1109/WIW.2016.045.

Ruso, J., & Stojanović, V. (2012). Occupational health and safety using data mining. *International Journal of Qualitative Research*, *6*(4), 168-194.

Saâdaoui, F., Bertrand, P. R., Boudet, G., Rouffiac, K., Dutheil, F., & Chamoux, A. (2015). A dimensionally reduced clustering methodology for heterogeneous occupational medicine data mining. *IEEE Transactions on Nanobioscience*, *14*(7), 707-715. http://dx.doi.org/10.1109/TNB.2015.2477407. PMid:26357403.

Sanchez-Pi, N., Marti, L., Molina, J. M., & Garcia, A. C. B. (2014). An information fusion framework for context-based accidents prevention. In *17th International Conference on Information Fusion (FUSION)* (pp. 1-8). New York: IEEE.

Sanmiquel, L., Bascompta, M., Rossell, J. M., Anticoi, H. F., & Guash, E. (2018). Analysis of occupational accidents in underground and surface mining in Spain using data-mining techniques. *International Journal of Environmental Research and Public Health*, *15*(3), 462. http://dx.doi.org/10.3390/ijerph15030462. PMid:29518921.

Sanmiquel, L., Rossell, J. M., & Vintró, C. (2015). Study of Spanish mining accidents using data mining techniques. *Safety Science*, *75*, 49-55. http://dx.doi.org/10.1016/j.ssci.2015.01.016.

Sanni-Anibire, M. O., Mahmoud, A. S., Hassanain, M. A., & Salami, B. A. (2020). A risk assessment approach for enhancing construction safety performance. *Safety Science*, *121*, 15-29. http://dx.doi.org/10.1016/j.ssci.2019.08.044.

Sarkar, S., Lodhi, V., & Maiti, J. (2019a). Text-clustering based deep neural network for prediction of occupational accident risk: a case study. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). New York: IEEE. https://doi.org/10.1109/iSAI-NLP.2018.8692881.

Sarkar, S., Pateshwari, V., & Maiti, J. (2017). Predictive model for incident occurrences in steel plant in India. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). New York: IEEE. http://dx.doi.org/10.1109/ICCCNT.2017.8204077.

Sarkar, S., Raj, R., Vinay, S., Maiti, J., & Pratihar, D. K. (2019b). An optimization-based decision tree approach for predicting slip-trip-fall accidents at work. *Safety Science*, *118*, 57-69. http://dx.doi.org/10.1016/j.ssci.2019.05.009.

Sarkar, S., Verma, A., & Maiti, J. (2018). Prediction of occupational incidents using proactive and reactive data: a data mining approach. In J. Maiti & P. K. Ray (Eds.), *Industrial safety management* (pp. 65-79). Singapore: Springer. http://dx.doi.org/10.1007/978-981-10-6328-2_6.

Sarkar, S., Vinay, S., & Maiti, J. (2016). Text mining based safety risk assessment and prediction of occupational accidents in a steel plant. In *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)* (pp. 439-444). New York: IEEE. http://dx.doi.org/10.1109/ICCTICT.2016.7514621.

Sarkar, S., Vinay, S., Raj, R., Maiti, J., & Mitra, P. (2019c). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, *106*, 210-224. http://dx.doi.org/10.1016/j.cor.2018.02.021.

Shein, M. M., Hamilton-Wright, A., Black, N., Samson, M., & Lecanelier, M. (2015). Assessing ergonomic and postural data for pain and fatigue markers using machine learning techniques. In *2015 International Conference and Workshop on Computing and Communication (IEMCON)* (pp. 1-6). New York: IEEE. http://dx.doi.org/10.1109/IEMCON.2015.7344435.

Shin, D.-P., Park, Y.-J., Seo, J., & Lee, D.-E. (2018). Association rules mined from construction accident data. *KSCE Journal of Civil Engineering*, *22*(4), 1027-1039. http://dx.doi.org/10.1007/s12205-017-0537-6.

Shirali, G. A., Noroozi, M. V., & Malehi, A. S. (2018). Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. *Journal of Public Health Research*, *7*(2), 1361. http://dx.doi.org/10.4081/jphr.2018.1361. PMid:30581805.

Siddula, M., Dai, F., Ye, Y., & Fan, J. (2016). Classifying construction site photos for roof detection. *Construction Innovation*, *16*(3), 368-389. http://dx.doi.org/10.1108/CI-10-2015-0052.

Taylor, J. A., Lacovara, A. V., Smith, G. S., Pandian, R., & Lehto, M. (2014). Near-miss narratives from the fire service: a Bayesian analysis. *Accident; Analysis and Prevention*, *62*, 119-129. http://dx.doi.org/10.1016/j.aap.2013.09.012. PMid:24144497.

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2017). Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining. *Automation in Construction*, *74*, 39-54. http://dx.doi.org/10.1016/j.autcon.2016.11.001.

Tomiazzi, J. S., Judai, M. A., Nai, G. A., Pereira, D. R., Antunes, P. A., & Favareto, A. P. A. (2018). Evaluation of genotoxic effects in Brazilian agricultural workers exposed to pesticides and cigarette smoke using machine-learning algorithms. *Environmental Science and Pollution Research International*, *25*(2), 1259-1269. http://dx.doi.org/10.1007/s11356-017-0496-y. PMid:29086360.

Tomiazzi, J. S., Pereira, D. R., Judai, M. A., Antunes, P. A., & Favareto, A. P. A. (2019). Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. Environ. *Environmental Science and Pollution Research International*, *26*(7), 6481-6491. http://dx.doi.org/10.1007/s11356-018-04106-w. PMid:30623325.

Ueno, K., Hayashi, T., Iwata, K., Honda, N., Kitahara, Y., & Paul, T. K. (2008). Prioritizing health promotion plans with k-bayesian network classifier. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 10-15). New York: IEEE. http://dx.doi.org/10.1109/ICMLA.2008.117.

Valêncio, C. R., Ichiba, F. T., Medeiros, C. A., & Souza, R. C. G. (2011). Spatial clustering applied to health area. In *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies* (pp. 427-432). New York: IEEE. http://dx.doi.org/10.1109/PDCAT.2011.76.

Waghmare, K., & Pai, A. R. (2013). Analytical study using data mining for periodical medical examination of employees. In *Proceedings of International Conference on Advances in Computing* (pp. 221-227). New Delhi: Springer Verlag. https://doi.org/10.1007/978-81-322-0740-5_27.

Xie, X., & Chang, Z. (2018). Intelligent wearable occupational health safety assurance system of power operation. *Journal of Medical Systems*, *43*(1), 16. http://dx.doi.org/10.1007/s10916-018-1122-3. PMid:30542831.

Yanar, B., Lay, M., & Smith, P. M. (2019). The interplay between supervisor safety support and occupational health and safety vulnerability on work injury. *Safety and Health at Work*, *10*(2), 172-179. http://dx.doi.org/10.1016/j.shaw.2018.11.001. PMid:31297279.

Yoon, S. J., Lin, H. K., Chen, G., Yi, S., Choi, J., & Rui, Z. (2013). Effect of occupational health and safety management system on work-related accident rate and differences of occupational health and safety management system awareness between managers in South Korea's construction industry. *Safety and Health at Work*, *4*(4), 201-209. http://dx.doi.org/10.1016/j.shaw.2013.10.002. PMid:24422176.

Zhao, Y., Li, J., Zhang, M., Lu, Y., Xie, H., Tian, Y., & Qiu, W. (2019). Machine learning models for the hearing impairment prediction in workers exposed to complex industrial noise. *Ear and Hearing*, *40*(3), 690-699. http://dx.doi.org/10.1097/AUD.0000000000000649. PMid:30142102.

Appendix A. Table with the articles chosen in systematic mapping (identification and citation codes).

Study ID	Citation	Dataset input	DM technic	Industrial activity sector
S1	Bevilacqua et al. (2008)	200	Decision Trees	Petrochemical
S2	Liao & Perng (2008)	309	Association Rules	Civil construction
S3	Ueno et al. (2008)		K- Bayesian Network Classifier	Healthcare
S4	Ciarapica & Giacchetta (2009)	36960	ANN; Fuzzy; Holt-Winters	
S5	Qu (2009)	1500	ANN	Mining
S6	Olsen et al. (2009)	11		Healthcare
S7	Cheng et al. (2010)	1347	Association Rules	Civil construction
S8	Palamara et al. (2011)	1700	K-means; SOM; Hill Climbing	Timber industry
S9	Valêncio et al. (2011)		K-means; Clarans	
S10	Cheng et al. (2012)	1542	Decision Trees	Civil construction
S11	Bonneterre et al. (2012)	81132		
S12	Bertke et al. (2012)		Naïve Bayes	
S13	Ruso & Stojanović (2012)	309	K-means	Civil construction
S14	Cheng et al. (2013)	349	Decision Trees	Petrochemical
S15	Nenonen (2013)	48869	Decision Trees; Association Rules	
S16	Waghmare & Pai (2013)	303	Decision Trees	Petrochemical
S17	Gross et al. (2013)		Decision Trees; Naïve Bayes; RIPPER	
S18	Hajakbari & Minaei-Bidgoli (2014)	8100	Decision Trees; K-means	
S19	Paliyawan et al. (2014)			Administrative
S20	Taylor et al. (2014)			
S21	Sanchez-Pi et al. (2014)		FP-Growth	Petrochemical
S22	Sanmiquel et al. (2015)	69869	Decision Trees	Mining
S23	Mistikoglu et al. (2015)	1413	Decision Trees	Civil construction
S24	Saâdaoui et al. (2015)	813	Expectation Maximization	Healthcare
S25	Shein et al. (2015)	10	Decision Trees	Administrative
S26	Krishna et al. (2015)	76	Decision Trees	Steel industry
S27	Bohanec & Delibašić (2015)		Decision Trees; Naïve Bayes; SVM; K-means	Tourism
S28	Comberti et al. (2015)	1247	K-means; SOM	Timber industry
S29	Nanda et al. (2016)	50000	Naïve Bayes	
S30	Chokor et al. (2016)	1044	K-means	Civil construction
S31	Luo et al. (2016)		ANN	Mining
S32	Rubaiyat et al. (2016)		SVM	Civil construction
S33	Sarkar et al. (2016)		Naïve Bayes; Fault Tree	Steel industry
S34	Siddula et al. (2016)		SVM	Civil construction
S35	Marucci-Wellman et al. (2017)	30000	Naïve Bayes; SVM; Logistic Regression	
S36	Rashid et al. (2017)		Hidden Markov Model	Civil construction
S37	Tixier et al. (2017)	5298	Hill Climbing	Civil construction
S38	Sarkar et al. (2017)	9488	Random Forest; SVM	Steel industry
S39	Goh & Ubeynarayana (2017)	1000	Decision Trees; KNN; Naïve Bayes; Logistic Regression; Random Forest; SVM	Civil construction
S40	Gerassis et al. (2017)		Naïve Bayes	Civil construction
S41	Akboğa & Baradan (2017)	2024		Civil construction
S42	Comberti et al. (2018)	1247	K-means; SOM	Timber industry
S43	Antwi-Afari et al. (2018)			Civil construction
S44	Baghdadi (2018)	20	SVM	
S45	Jiang et al. (2018)	1170	Naïve Bayes	Civil construction

Appendix A. Continued...

Study ID	Citation	Dataset input	DM technic	Industrial activity sector
S46	Jocelyn et al. (2018)	23	Hierarchical Agglomerative Clustering	Transport
S47	Shin et al. (2018)	98189	Decision Trees	Civil construction
S48	Abad et al. (2019)		BNN	Civil construction; Energy
S49	Tomiazzi et al. (2018)		KNN; NN; Optimum Path Forest; SVM	Agriculture
S50	Sarkar et al. (2018)		Decision Trees; Adaboost;	Steel industry
S51	Pekel et al. (2018)		Bayesian network	Civil construction
S52	Del Pozo-Antúnez et al. (2018)	44589	ANN	Administrative
S53	Lee & Kim (2018)	89921	Decision Trees; Random Forest; SVM	
S54	Sanmiquel et al. (2018)	56034	Decision Trees; Association Rules;	Mining
S55	Shirali et al. (2018)	2127	Decision Trees	Steel industry
S56	Meyers et al. (2018)	9600	Naïve Bayes; Principal Component Analysis	
S57	Sarkar et al. (2019c)	3308	ANN; Decision Trees; Random Forest; SVM	Steel industry
S58	Kang & Ryu (2019)	9796	Random Forest	Civil construction
S59	Kakhki et al. (2019)	33458	Naïve Bayes; SVM; Boosted Trees	Agroindustry
S60	Kao et al. (2018)	200	Non-negative matrix factorization	Healthcare
S61	Heo et al. (2019)		CNN	
S62	Tomiazzi et al. (2019)	127	ANN; KNN; SVM	Agriculture
S63	Di Noia et al. (2019)		K-means; KNN; SVM	Healthcare
S64	Xie & Chang (2018)		SVM	Energy
S65	Sarkar et al. (2019b)	1964	Random Forest	Steel industry
S66	Zhao et al. (2019)		Adaboost; Random Forest; SVM	Manufacturing
S67	Sarkar et al. (2019a)	3488	DNN; K-means; Random Forest; SVM	Steel industry
S68	Khosrowabadi & Ghousi (2019)	5600	Association Rules; K-means; GRI-3	Automotive