Editor Responsável: Prof. Dr. Hermes Moretti Ribeiro da Silva

Recebido: 23/11/2020
Aprovado: 03/03/2021

Occupational health and safety and data mining: a bibliometric analysis

Camila Rafael[1], State University of Maringá, Paraná, Brazil
Mateus Vicente Peternella[2], State University of Maringá, Paraná, Brazil
Beatriz Lavezo dos Reis[3], State University of Maringá, Paraná, Brazil
Gislaine Camila Lapasini Leal[4], State University of Maringá, Paraná, Brazil
Rodrigo Clemente Thom de Souza[5], Federal University of Paraná, Maringá, Paraná, Brazil
Edwin Vladimir Cardoza Galdamez[6], State University of Maringá, Paraná, Brazil

RESUMO

Objetivo – Este artigo tem o objetivo de realizar uma análise bibliométrica sobre o tema de mineração de dados e saúde e segurança do trabalho, contemplando o período compreendido entre os anos de 2008 e 2020, sete bases de dados científicas e 68 registros selecionados.
Referencial teórico – Esta pesquisa se fundamentou teoricamente em conceitos que envolvem a mineração de dados, aprendizado de máquinas e saúde e segurança do trabalho.
Metodologia – Os artigos escolhidos foram submetidos a uma análise estatística, juntamente com a avaliação de uma das leis da bibliometria (Lei de Bradford), sobre a quantidade de citações, periódicos, autores, países de origem, categorias de publicação e avaliação da produtividade ao longo dos anos.
Resultados – Como resultado, constatou-se que o periódico mais influente é a *Safety Science*, e Taiwan é o país líder na origem dos artigos, com uma média de 115 citações por artigo. As revistas melhor ranqueadas são associadas aos temas *Engineering* e *Health*, ambas contendo 30% dos artigos e periódicos selecionados.
Contribuições – Com a pesquisa foi possível identificar insights sobre o crescimento da área de mineração de dados aliada a saúde e segurança do trabalho.
Palavras-chave – Análise de bibliometria. Saúde e segurança no trabalho. Mineração de dados.

ABSTRACT

Purpose *- This article aims to carry out a bibliometric analysis on data mining and occupational health and safety, covering the period between 2008 and 2020, for seven scientific databases and 68 articles.*
Theoretical framework *- This study was theoretically based on concepts that involve data mining, machine learning and occupational health and safety.*
Design/methodology/approach *- The selected articles were submitted to a statistical analysis, together with the evaluation of one of the bibliometric laws (Bradford's Law), comprising a number of citations, journals, authors, countries of origin, publication categories and an evaluation of production over the years.*
Findings *- As a result, it was found that the most influential journal was Safety Science, and Taiwan was the leading country in terms of articles produced, with an average of 115 citations per article. The best-ranked journals related to Engineering and Health, both corresponding to 30% of the selected articles and journals.*
Originality/value *- This study provides some insights into the growth of the data mining area together with occupational health and safety.*
Keywords *- Bibliometrics analysis. Occupational health and safety. Data mining.*

1. Av. Colombo, 5790 - Vila Esperança, Maringá - PR, 87020-270, Bloco 19/20, camilarafaelmga@gmail.com, https://orcid.org/0000-0002-4428-2914; 2. mateuspeternella@hotmail.com, https://orcid.org/0000-0001-5860-0664; 3. bia.lavezo@gmail.com, https://orcid.org/0000-0002-5916-3184; 4. gclleal@uem.br, https://orcid.org/0000-0001-8599-0776; 5thom@ufpr.br, https://orcid.org/0000-0003-2435-8528; 6. evcgaldamez@uem.br, https://orcid.org/0000-0002-1763-9332.
RAFAEL, C.; PETERNELLA, M.V.; REIS, B.L.; LEAL, G.C.L.; SOUZA, R.C.T.; GALDAMEZ, E.V.C. Occupational health and safety and data mining: a bibliometric analysis. **GEPROS. Gestão da Produção, Operações e Sistemas**, v.16, nº 2, p. 168 – 194, 2021.
DOI: http://dx.doi.org/10.15675/gepros.v16i2.2784

1. INTRODUCTION

Industrial revolutions have brought about innumerable changes within organizations in terms of working conditions, work environments and labor laws. These changes are considered positive in some aspects. For instance, services automation and technological advances have made life more efficient and faster (BADRI; BOUDREAU-TRUDEL; SOUISSI, 2018). However, they have also increased the risks of occupational accidents and diseases by the increasing exposure to machinery and computers. Thus, the work environment has made employees more susceptible to cardiovascular and stress-related diseases (RUSO; STOJANOVIĆ, 2012).

Regarding labor conditions and legislation, at the beginning of industrialization, commitment to health and safety of workers was not common. Yet, over time, it has become increasingly mandatory, besides being a competitive strategy for organizations, since the welfare of employees results in greater productivity (CIARAPICA; GIACCHIETA, 2009).

The International Labor Organization (ILO) states that Occupational Health and Safety (OHS) protects workers from general and occupational diseases and accidents at work (COMBERTI; DEMICHELA; BALDISSONE, 2018). In this context, OSH has become a recurring issue al around the world, as it represents a competitive strategy for organizations, and also because of the alarming rates of occupational occurrences. According to estimates of the ILO, approximately one employee dies due to occupational accidents or diseases every 15 seconds, and approximately 2.34 million employees die from work-related diseases every year (WANG et al., 2020).

Furthermore, financial loss is not limited to compensations claimed by employees due to accidents and diseases. There is also an economic burden on society, which is imposed, for instance, by Social Security expenditures (YILMAZA; ÇELEBIB, 2015). Thus, the costs that must be faced by companies and society in general demand occupational safety measures. In

addition to financial losses, accidents, diseases and deaths contribute to excluding thousands of people from the labor market (COLNAGO; SIVOLELLA, 2019).

One of the ways to find solutions for occupational occurrences is to study cases that have already happened by evaluating specific cases or databases. The analysis of the history of occupational occurrences is important because it provides a complete picture of activities to prevent accidents and diseases. Thus, it is a powerful tool to minimize risks (ZHANG; JIANG, 2012).

Data mining (DM) is an alternative to deal with the huge amount of occupational data. It is described as a set of techniques that allow the processing of large amounts of data (Big Data), extracting useful information for a specific purpose. Often, DM is linked to machine learning (ML), since both have automated pattern extraction techniques, which represent knowledge implicitly stored within databases (HAN; KAMBER; PEI, 2012).

The best-known techniques used in DM are Association Rules (AR), Artificial Neural Networks (Neuro-Fuzzy) and Regression Trees (RT). All these techniques have a common path: data preparation stages, model execution and analysis of the results (BUCZAK; GUVEN, 2016; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). DM can help with OSH by extracting patterns from databases where records are stored, so that, studies can be conducted and preventive measures can be taken. As an example, Sanmiquel, Rossell and Vintró (2015) carried out a case study in the Spanish mining sector in which decision tree techniques and Bayesian classifiers were applied to explore occupational accidents in mines. The study contributed to the development of safety policies in the mine where the research took place.

DM and OSH are interconnected and represent a recurring issue. That leads to an increase in the scientific production on the subject (RUSO; STOJANOVIĆ, 2012). In this context, conducting a bibliometric analysis of the studies available on the topic is important to identify trends and research growth. Moreover, it is a way to identify the main journals related

to the topic and predict productivity of individual authors and countries, as well as analyze the emergence of new study areas (PIMENTA et al., 2017; QUEVEDO-SILVA *et al.*, 2016).

In light of the foregoing, this study aims to present a bibliometric analysis to identify the most impactful articles, authors, journals, countries and publication categories associating OSH to DM. The selected articles were identified through a systematic mapping conducted by a research group whose goal was to answer the following question: "How does DM support decision-making in OSH?" The mapping was performed to identify what bases and types of OSH data, in addition to techniques and tools, are used by DM.

This bibliometric study has an investigative and quantitative approach, a descriptive objective and bibliographic research procedures. It is divided into five sections. Section 1 introduces the topic for research contextualization. Section 2 briefly presents a theoretical framework on OSH and DM. Section 3 details the methodology. Section 4 provides research results through statistical analysis. Finally, section 5 presents final considerations with some insights for future research on the topic.

2. THEORETICAL FOUNDATION

2.1 Occupational Safety and Health (OSH)

The occupational safety and health concept has risen in the academic and business environments in recent decades, mainly influenced by an increase in cases of work-related incidents all over the world (SÁNCHEZ-HERRERA; DONATE, 2019). Attention to OSH has become fundamental for the survival of organizations. It aims to identify, manage and mitigate the risks involving workers' health and safety (MUTLU; ALTUNTAS, 2019).

As employees are the pillars of organizations development, the concern with health and safety has led to the creation of new policies, laws and regulations, both by private and public organizations (AZIZ; OSMAN, 2019; CHEN *et al.*, 2020). Those measures have been taken not only due to the damage suffered by workers and their families, but also due to economic losses. The ILO estimates that losses caused by occupational accidents and diseases

correspond to US $ 3.3 trillion, which represents about 4% of the global gross domestic product (WANG *et al*., 2020).

To reduce human and financial losses, it is necessary to understand the concepts involved in OSH. A risk means the possibility of an occurrence whose outcome is detrimental to an employee. That can be an injury or loss caused by a dangerous situation (MUTLU; ALTUNTAS, 2019). Regarding diseases, they are divided into two classes. Profession-related diseases are those associated with the profession of the victim, whereas occupational diseases are the ones caused by special conditions a worker is subject to (AZZOLIN *et al*., 2012).

As for accidents, there are five categories according to the NBR 14280: 2001: working, commuting, without injury, impersonal and personal. However, the most serious outcome is death. A work death is that related to one's profession, regardless of the interval between accident and death confirmation (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2001).

All occurrences must be registered for companies' control, but also for the government and regulatory organizations monitoring. In Brazil, all diseases, accidents and deaths related to work must be registered through a Work Accident Communication (*Comunicação de Acidentes de Trabalho* - CAT). Then, the information is formalized, stored and made publicly available (BATISTA; SANTANA, FERRITE, 2019).

Other countries also register and disclose their data publicly, such as Taiwan, with the Council of Labor Affairs (Executive Yuan) (CHENG; YAO; WU, 2013; LIAO; PERNG, 2008), Italy with the *Istituto nazionale per l'assicurazione contro gli infortuni sul lavoro* (INAIL) (COMBERTI; DEMICHELA; BALDISSONE, 2018; PALAMARA; PIGLIONE; PICCININI, 2011) and the United States, with the Occupational Safety and Health Administration (OSHA) (SHIN *et al*., 2018; TIXIER *et al*., 2017).

Creation and dissemination of these occupational data sets are useful in investigating the OSH scenario, developing actions, supporting new legislation and assisting organizations in decision-making (DEL POZO-ANTÚNEZ *et al*., 2018; YANAR; LAY; SMITH, 2019). In

order to carry out these analyses, some tools should be used, such as the exploratory analysis of data and statistical methods (CHENG *et al*., 2010) and computational resources, such as data mining (CHOI *et al*., 2020).

2.2 Data Mining (DM)

According to Witten and Frank (2016), DM covers the use of ML techniques to extract knowledge from large amounts of data, with applications in several areas, such as medicine, finances, retail, manufacturing, engineering and others. Medical diagnostics, credit analysis and fraud detection are examples of applications in some of these areas. The term "mining" was chosen because this process of extracting knowledge from data leads us, metaphorically, to perform mineral processing activities that extract economic interest products from large deposits.

By making use of ML algorithms, DM is often linked to Artificial Intelligence (AI). This relationship is based on the fact that DM models can have learning and adaptation capacities. These models can provide solutions for situations which they were not explicitly programmed for. They make use of mathematical and statistical theories and are programmed (as well as parameterized) to optimize a performance criterion. This optimization uses data as examples, based on past experiences and with the ultimate goal of making predictions (WITTEN; FRANK, 2016).

DM can also be associated with the KDD (Knowledge Discovery in Databases) process, as it is one of the stages involved in the entire process. However, there are nine steps that compose the KDD procedure, as presented by Fayyad, Piatetsky-Shapiro and Smyth (1996): understanding the customer's demand; selecting the data set; performing cleaning and pre-processing; reducing and projecting; aligning methods to the demand; defining the model and its elements; performing data mining; interpreting results; and using the knowledge discovered.

A way to classify data mining methods refers to model learning, which can be supervised or unsupervised. The first method is guided by the developer, while the second is not (HAJAKBARI; MINAEI-BIDGOLI, 2014; ZHAO *et al*., 2019). In addition, DM methods have two main objectives, namely, prediction of future events, based on the mined data, and description, in which there is a search for patterns that can be understood by the decision-makers (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). The application of data mining to OSH is often associated with forecasting and extracting patterns. In those cases, the techniques may be related to injury risk prediction, accidents or even to predict the seriousness of those events (SARKAR; MAITI, 2020).

3. METHODOLOGICAL PROCEDURES

The 68 articles selected for bibliometric analyses were identified through a systematic mapping of the literature, conducted by a research group. The selected articles (also used in this research) addressed topics related to DM and ML, with applications in health and safety at work. These records were selected by using IEEE Xplore, Ingenta, Science Direct, Scopus, SpringerLink, PubMed and ProQuest databases.

The inclusion criteria met the following restrictions: published since 2008; English-language publications; and showing an application of DM in OSH. The exclusion criteria were: not peer reviewed; not showing an application of DM in OSH; and not responding satisfactorily to our research questions. As for duplicate articles, only the most complete one was selected.

The initial search on the seven databases, based on the aforementioned inclusion and exclusion criteria, resulted in 1424 articles from which 474 duplicate studies were excluded. After partial and total reading of the remaining articles, the final set reached 68 articles. By exploring them, we sought to answer five questions:

(i) What kind of occupational safety and health data are explored?

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

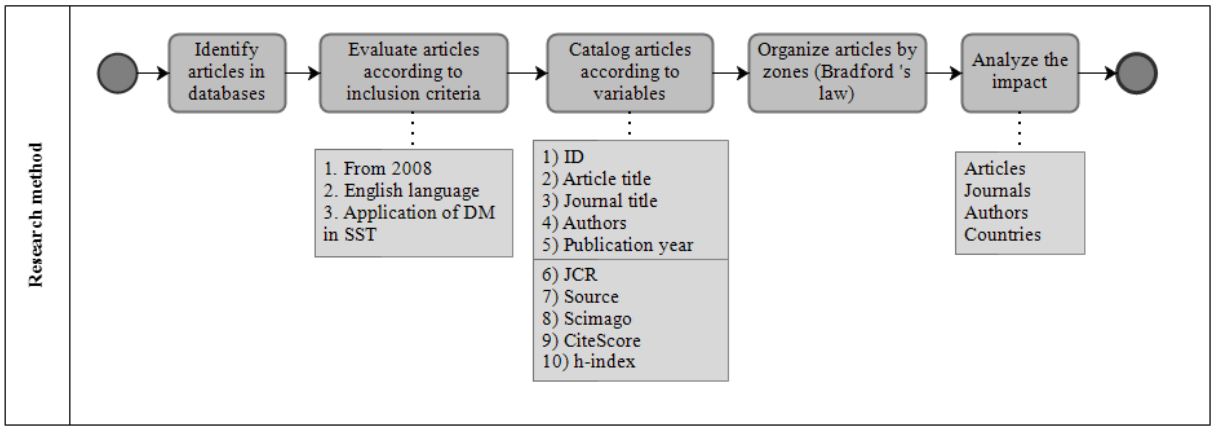
(iii) What industrial activity sector is explored in the research?

(iv) Which occupational safety and health database was used?

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

This research was carried out by using the same 68 articles. Figure 1 shows all the steps of the research methodology, from the identification of the articles to the final analysis

Figure 1 – Research method used



Source: Authors (2021).

Based on the selected articles, it was possible to start the analysis of the records, in which the publications were cataloged considering ten variables, namely, Code; Title of the article; Journal title; Authors; Year of publication; JCR; Source; Scimago; CiteScore; and Index H. After a synthesis of the information, Bradford's Law was applied to verify the degree of attraction of the journals by adopting their reputation as a criterion to identify the most relevant ones, as well as the ones that give more attention to a specific topic.

The titles and the number of articles published by each journal are organized in a table in descending order of productivity, divided into three zones, each of them containing one third of the total of journals. Thus, each zone averaged 22 articles. The analysis also considered the impacts of the productions, journals, authors and countries.

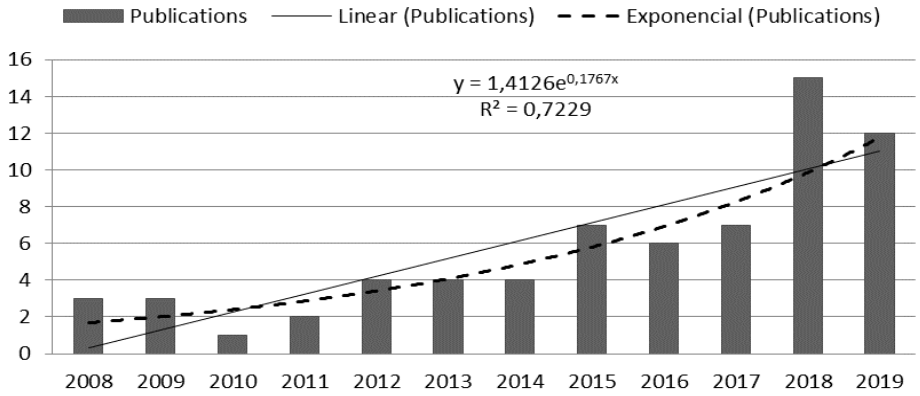
4. RESULTS AND DISCUSSION

4.1. Productivity evaluation over the years

Statistical analysis showed that, within the period of publication of the articles (2008 - 2020), the amount of research had an increase. The increase in the number of publications on OSH and DM in recent years can be explained by the following factors: a) the number of researchers has grown exponentially and, thus, the number of submissions to journals has also increased; b) the use of technologies, such as computers connected to the internet, facilitates the access to updated information sources; c) nowadays, workers' health has been a very discussed topic and has gained emphasis in large companies that try to promote health, safety, comfort and satisfaction of their employees. There are some other factors that have influence on efficiency and productivity, such as reducing compensation costs due to work accidents.

All of these factors led to an increase in research and publications in the area, with an average growth of 17% in the last 4 years, which is shown in Figure 2.

Figure 2 – Number of publications per year



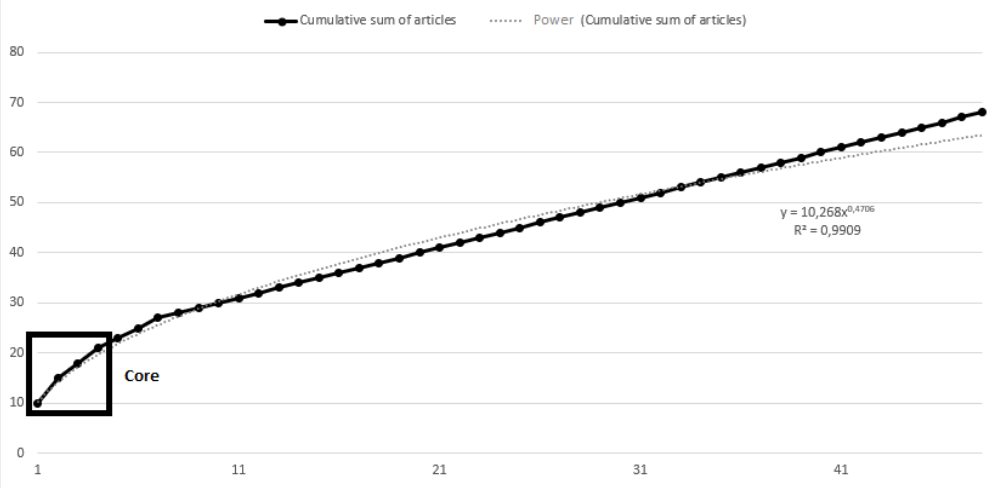
Source: Authors (2021).

A projection of the total production in 2020 was also made based on the exponential growth of the previous years, and the average number was 14 publications.

4.2. Journals evaluation

Regarding the journals themselves, Bradford's Law was used to assist us in their evaluation. By arranging the 48 journals in descending order of productivity and separating them into three zones, with an average of 22 articles each zone, it was possible to identify the most relevant and productive journals on the list. Figure 3 shows the distribution of the accumulated sum of articles indicating that, as the number of journals increases, production decreases.

Figure 3 – Number of publications per year



Source: Authors (2021).

Thus, it was possible to identify that the first zone (core) gathered the most productive journals, totaling four with 21 articles, followed by zone 2, with 20 journals and 23 articles. The third zone, which was the least productive, reached only one publication per journal, with 24 journals and 24 articles. As provided by Bradford's law, "few produce much and many produce little" (ARAÚJO, 2006).

Calculation of the Minimum Bradford Zone (MBZ) also justifies the number of articles in each zone:

$$MBZ = \frac{NR1a}{2}; MBZ = \frac{41}{2}; MBZ = 20{,}5$$

where NR1a is the total number of journals with a single article.

Table 1 shows the most productive and relevant journals on the list, found in the core zone, as well as their TP (total publications), hi% (relative frequency) and Hi% (cumulative frequency) indexes.

Table 1 – List of journals belong to the core

Ranking	**Journals**	**TP**	**∑ Cumu.**	**hi%**	**Hi%**	**Citation < 2020**
1º	Safety Science	10	10	14,70	14,70	493
2º	*Accident Analysis & Prevention*	5	15	7,35	22,05	242
3º	*Environmental Science and Pollution Research*	3	18	4,41	26,46	10
4º	*Int. J. of Environmental Research and Public Health*	3	21	4,41	30,87	27

Source: Authors (2021).

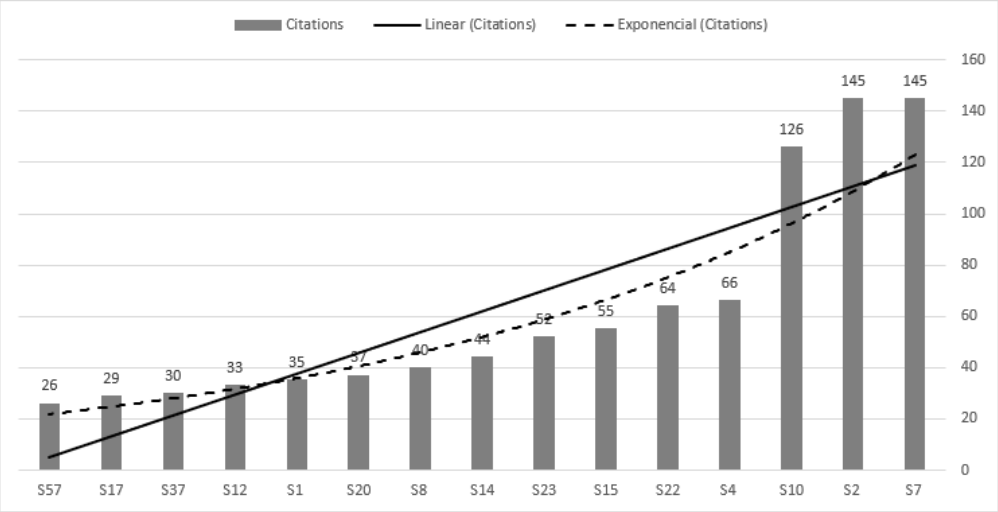
Although the core zone contains only four journals, it corresponds to about 30% of the total number of articles. Safety Science leads the ranking of publications with 10 of the 68 articles, showing a strong influence on OSH and DM. The journal addresses multidisciplinary research on employees' health and safety, involving aspects related to social studies, technologies, legislation, control, safety techniques and others.

The second most productive journal is Accident Analysis & Prevention, which deals with specific issues involving occupational accidents and diseases. It covers medical research, studies of human and environmental factors that can cause diseases, injuries and fatalities. The third and fourth journals, in terms of relevance, are also multidisciplinary. They discuss issues related to occupational hygiene, public health research and engineering sciences, such as programming.

4.3. Evaluation of citations between articles

Regarding evaluation of the articles, the number of citations each publication had was analyzed, with a total of 1291. Figure 4 shows the 15 articles that were most referenced by other publications. Three of them stood out (S2, S7 and S10), with 32% of the total of citations, but only 4% of the total number of articles.

Figure 4 - Ranking of the 15 most cited articles



Source: Authors (2021).

It is noteworthy that articles S2 and S7 had the maximum number of citations. They were referenced by 145 publications. In addition, the two articles were published by Safety Science, which was the most influential journal among those selected for the research, as previously reported. Details of the characteristics of the 15 most cited articles are presented in Table 2, as well as the data mining techniques and tools used, the sector of application, country of origin and type of result presented.

Table 2 - Characteristics of the 15 most referenced articles (to be continued)

ID	Techniques	Tools	Application sector	Country	Results
S2	*Association Rules*	Not Specified	Civil Construction	Taiwan	Data visualization and analysis only

S7	*Association Rules*	Statistica Data Miner®	Civil Construction	Taiwan	Data visualization and analysis only
S10	*Decision tree*	Statistica Data Miner®	Civil Construction	Taiwan	Visualization and analysis with suggestions of actions and/or tools
S4	*Neuro-fuzzy*	MATLAB®; Microsoft Excel® Solver	Not Specified	Italy	Data visualization and analysis only
S22	*Decision tree, bayes*	WEKA®	Mining	Spain	Data visualization and analysis only
S15	*Decision tree, Association Rules*	SAS®	Not Specified	Finland	Data visualization and analysis only
S23	*Decision tree*	Clementine®	Civil Construction	Turkey	Data visualization and analysis only
S14	*Regression tree*	Statistica Data Miner®	Petrochemical	Taiwan	Visualization and analysis with suggestions of actions and/or tools
S8	*k-means, Hierarquical clustering*	MATLAB®	Timber sector	Italy	Data visualization and analysis only
S20	*Neuro-fuzzy, Naive*	TextMiner®	Not Specified	United States	Visualization and analysis with suggestions of actions and/or tools
S1	*Regression tree*	Not Specified	Petrochemical	Italy	Visualization and analysis with suggestions of actions and/or tools

Table 2 - Characteristics of the 15 most referenced articles (continued)

ID	Techniques	Tools	Application sector	Country	Results
S12	*Naive-Bayes*	Not Specified	Not Specified	United States	Data visualization and analysis only

S37	*Hierarquical clustering*	Not Specified	Civil Construction	France	Visualization and analysis with suggestions of actions and/or tools
S17	*Naive-Bayes, Decision tree*	SMOTE®	Not Specified	Canada	Visualization and analysis with suggestions of actions and/or tools
S57	*Decision tree (C 5.0)*	Not Specified	Steel sector	India	Visualization and analysis with suggestions of actions and/or tools

Source: The Authors (2021).

[1] Clustering, also known as grouping, is a data mining task used when there are no predefined classes for dividing the data. In these cases, the algorithm itself decides which groups (clusters) will be used to classify the data (WITTEN; FRANK, 2016).

Liao and Perng (2008), authors who lead the ranking (S2), address in their research some DM techniques used to identify characteristics of work accidents in the construction industry. Their research revealed that many accidents are related to weather conditions. The second article (S7), by Cheng et al. (2010), who also lead the ranking, discusses the possible patterns of accidents in civil construction to be detected by using DM. Their study listed some risky combinations responsible for many accidents, such as working in heights without the necessary safety equipment.

In the third most referenced article (S10), which was cited 126 times, Cheng et al. (2012) address the most serious occupational accidents, such as injuries and fatalities, which have a high rate of occurrence in the construction sector worldwide. The authors also address prevention measures.

Another classification applied to the 15 articles refers to the tasks used in the DM process, divided into four types: three of them related to supervised learning (association, classification and regression) and one linked to unsupervised learning (clustering[1]). Data processing during ML, with previously labeled inputs and outputs, can be understood as supervised learning. It means that the types of data and their meaning are already known.

Unsupervised learning, however, refers to ML whose data are unknown, presenting several results.

Most articles presented supervised learning tasks. Classification task was the most used. It was found in 78% of the records, in some cases as a single task and, in others, in comparisons or in association with other types. Clustering was used in 15 studies (22%), and association was found in eight studies (12%). Regression was the least applied task, corresponding to only 7%.

4.4. Evaluation of the impact of the articles considering the number of authors

In order to determine whether the number of authors in each article directly affects the impact of the production, some parameters were collected (Table 3). In this study, articles written by one to 12 authors and their respective indexes TP, TC (Total citations), hi% and Hi% were identified.

Table 3 - Indexes for assessing the impact of the articles

Number of authors	TP	hi%	Hi%	Citations < 2020	TC/TP
1	3	4.41	4.41	55	18,33
2	11	16.18	20.59	279	25,36
3	18	26.47	47.06	454	25,22
4	14	20.59	67.65	208	14,86
5	10	14.71	82.35	163	16,30
>5	12	17.65	100	129	10,75
∑	68	100	-	1288	-

Source: Authors (2021).

Productions with up to three authors represented almost 50% of the total number of publications (Hi%). When comparing the total number of citations that the articles had in other publications, those written by up to three authors were the most referenced ones in the academic environment. In contrast, articles produced by five or more authors had an average of citations per published article (TC / TP) of 10.75. This corresponds to the lowest average in

the table, which shows that the smaller the number of authors in the articles, the greater the impact and the number of publications on the topic addressed.

4.5. Evaluation of the publication category of the journals

According to the classification category for journals, by Scimago Journal & Country Rank (Table 4), there were two main areas: Engineering (25 journals) and Health (19 journals). When analyzing the h-index (Hirsch index) of the categories, Engineering still remains the most productive area, with a score of 17. It means that there were, at least, 17 articles with, at least, 17 citations. Engineering was followed by Health, with a score of 15. Social Sciences, which previously had the lowest score in number of publications, ranked third in the h-index, with 13.

Table 4 - Categories of journals and articles

Category	Journals		Articles		Citations	TC/TP	*h-index*
	TP	hi%	TP	hi%	< 2020		
Engineering	25	32,05%	41	31,78%	1105	26,95	17
Health	19	24,36%	36	27,91%	916	25,44	15
Computer Science	19	24,36%	19	14,73%	191	10,05	7
Social Sciences	7	8,97%	20	15,50%	822	41,10	13
Others	8	10,26%	13	10,08%	140	10,77	6

Source: Authors (2021).

4.6. Scientific production by country

In total, about 20 countries produced the 68 articles selected for this study. As a way of evaluating the productivity of each country that contributed with an article on OSH and DM, some parameters were identified, as shown in Table 5.

Table 5 - Classification of countries

Citations	Articles			Citations	TC/TP	*h-index*
	TP	hi%	Hi%	< 2020		
United States	14	20,59%	20,59%	157	11,21	6
India	8	11,76%	32,35%	85	10,63	5
Italy	6	8,82%	48,53%	170	28,33	6
Spain	5	7,35%	39,71%	99	19,80	4
Brazil	4	5,88%	54,41%	17	4,25	3
China	4	5,88%	60,29%	6	1,50	2
South Korea	4	5,88%	66,18%	11	2,75	2
Taiwan	4	5,88%	72,06%	460	115,00	4
Canada	3	4,41%	76,47%	33	11,00	2
Iran	3	4,41%	80,88%	22	7,33	1
Turkey	3	4,41%	85,29%	59	19,67	2
France	2	2,94%	88,24%	48	24,00	2
Saudi Arabia	1	1,47%	89,71%	14	14,00	1
Slovenia	1	1,47%	91,18%	17	17,00	1
Finland	1	1,47%	92,65%	55	55,00	1
Hong Kong	1	1,47%	94,12%	5	5,00	1
Japan	1	1,47%	95,59%	3	3,00	1
Serbia	1	1,47%	97,06%	2	2,00	1
Singapore	1	1,47%	98,53%	22	22,00	1
Thailand	1	1,47%	100,00%	6	6,00	1

Source: Authors (2021).

The first parameter to be assessed was the total number of publications by country (TP). The United States leads the ranking, with 20% of the articles. In addition, more than 50% of the publications were produced by only five countries, namely, the United States, India, Italy, Spain and Brazil, an index called “cumulative frequency of publications” (Hi%).

Based on the h-index, it was possible to analyze the production capacity of the countries and its scientific-academic impact. As a result, it was noticed that Italy and the United States were the leading countries with the same score, reaching the h-index of six. It means that the two countries have, at least, six articles with, at least, six citations in each publication. The main works are, respectively, from the United States (S20, S12, S32, S35, S29 and S3) and Italy (S4, S8, S1, S42, S63 and S28).

India comes in second place, with a score of five (S57, S33, S50, S26 and S38). Then come Spain (S22, S54, S40 and S52) and Taiwan (S2, S7, S10 and S14), with the same score, reaching an h-index of four. For the analysis, the average of citations is also evaluated by the number of publications per country (TC / TP). In first place, comes Taiwan, with an average of 115 citations for each publication (Table 5), followed by Finland and Italy, which had 55 and 28 citations per article published, respectively. Among the 15 most cited articles, Taiwan has four, Italy three and the United States two, while Spain, Finland, Turkey, France, Canada and India have only one.

6. CONCLUSION

This research shows the importance of bibliometric literature reviews, not only as an instrument to identify and classify a wide variety of studies within areas related to OSH and DM, but also to analyze information and search for trends. Based on the results, it was possible to raise some insights and possibilities for future research.

The results of this study show an increasing trend in the number of annual publications, despite the reduction in the number of articles published in the last two years. For that reason, this area presents itself as an area of great interest among academics. Another indicator of interest is the progressive increase in the number of citations that the articles had, with three articles standing out from the rest: *S2 - Data mining for occupational injuries in the Taiwan construction industry* (LIAO; PERNG, 2008), *S7 - Use of association rules to explore*

cause–effect relationships in occupational accidents in the Taiwan construction industry (CHENG *et al*., 2010), *S10 - Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry* (CHENG *et al*., 2012).

Classification of the number of authors per article can be useful for researchers who need to carry out an initial review of the literature. In that sense, it is evident that the smaller the number of authors, the greater the relevance of the article, since the average number of citations per publication was high. In contrast, articles with five or more authors had the lowest average and, thus, were considered less relevant.

Regarding the analysis of the journals based on Bradford's law, only four journals belonged to zone 1, which was the most influential. Safety Science was the most productive among the 48 selected journals, once it had 10 of the 68 articles. Thus, we can infer that it has a strong influence on OSH and MD studies.

Regarding the number of articles per country, the United States stood out with the highest number of publications (14). However, it was not the country with the highest average number of citations per article. Taiwan leads this ranking with an average of 115 citations for each of its four articles. The countries' h-index shows that both the USA and Italy had a minimum of six articles with six citations.

Finally, this study revealed that there are two main areas among the publication categories of the journals: Engineering and Health. Engineering has the best performance, since it had, on average, 30% of the articles and journals selected. In addition, it has the highest h-index, reaching a score of 17 articles with, at least, 17 citations. The second best classified area was Health, with 25 citations on average for each article. Both areas have a strong influence on issues related to technology, data analysis, accidents and occupational diseases.

In light of the above, this research enabled the understanding of publications on data mining associated with health and safety, as well as insights and future research possibilities.

It was also possible to identify the increasing use of DM techniques to understand and predict behaviors of occupational incidents. That results in an improvement in the OSH scenario.

Regarding future research, this study suggests deepening into the selected articles, in order to carry out a comparative analysis of their results. Due to the increasing interest of the academia and the industry in OSH and MD, research with applications is also recommended to understand the behavior of real data and specific sectors of the industry, or even predictions of future accidents. That can be useful, for it would support new norms and policies, both for public and private organizations.

Acknowledgments

This study was partially funded by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), the National Council for Scientific and Technological Development - Brazil (CNPq) and Araucária Foundation.

References

ARAÚJO, C. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, v. 12, n. 1, p. 11-32, 2006.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14280**: Cadastro de acidente do trabalho - Procedimento e classificação. Rio de Janeiro, p. 1-94, 2001.

AZIZ, S. F. A.; OSMAN, F. Does compulsory training improve occupational safety and health implementation? The case of Malaysian. **Safety Science**, v. 111, p. 205–212, 2019.

AZZOLIN, K.; SOUZA, E. N.; RUSCHEL, K. B.; MUSSI, C. M.; LUCENA, A. F.; RABELO, E. R. Consenso de diagnósticos, resultados e intervenções de enfermagem para pacientes com insuficiência cardíaca em domicílio. **Revista Gaúcha de Enfermagem**, v. 33, n. 4, p. 56–63, 2012.

BADRI, A.; BOUDREAU-TRUDEL, B.; SOUISSI, A. S. Occupational health and safety in the industry 4.0 era: A cause for major concern?. **Safety Science**, v. 109, p. 403-411, 2018.

BATISTA, A. G.; SANTANA, V. S.; FERRITE, S. Registro de dados sobre acidentes de trabalho fatais em sistemas de informação no Brasil. **Revista Ciência e Saúde Coletiva**, v.24, p. 693-704, 2019.

BUCZAK, A. L., GUVEN, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. **IEEE Commuications Surveys & Tutorials** v. 18, n. 2, p. 1153–1176, 2016.

CHEN, H.; HOU, C.; ZHANG, L.; LI, S. Comparative study on the strands of research on the governance model of international occupational safety and health issues. **Safety Science**, v. 122, p. 104513, 2020.

CHENG, C. W.; LIN, C. C.; LEU, S. SEN. Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. **Safety Science**, v. 48, n. 4, p. 436–444, 2010.

CHENG, C. W.; YAO, H. Q.; WU, T. C. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. **Journal of Loss Prevention in the Process Industries**, v. 26, n. 6, p. 1269–1278, 2013.

CHENG, C.-W., LEU, S. S., CHENG, Y. M., WU, T. C., LIN, C. C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. **Accident Analysis & Prevention**, v. 48, p. 214–222, 2012.

CHOI, J., GU, B., CHIN, S., LEE, J. S. Machine learning predictive model based on national data for fatal accidents of construction workers. **Automation in Construction**, v. 110, p. 102974, 2020.

CIARAPICA, F. E.; GIACCHETTA, G. Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. **Safety Science**, v. 47, n. 1, p. 36–49, 2009.

COLNAGO, L; SIVOLELLA, R. Convenção 187 da OIT: promoção da saúde e segurança do trabalho no Brasil e a viabilidade de sua ratificação. **Revista eletrônica do Tribunal Regional do Trabalho da 9ª Região**, v. 8, p. 144-156, 2019.

COMBERTI, L.; DEMICHELA, M.; BALDISSONE, G. A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making. **Safety Science**, v. 106, p. 191–202, 2018.

DEL POZO-ANTÚNEZ, J. J., ARIZA-MONTES, A. FERNÁNDEZ-NAVARRO, F. MOLINA-SÁNCHEZ, H. Effect of a job demand-control-social support model on accounting

professionals' health perception. **International Journal of Environmental Research and Public Health**, v. 15, n. 11, 2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in databases**. AI Magazine. v. 17, n. 3, p. 37–54, 1996.

HAJAKBARI, M. S.; MINAEI-BIDGOLI, B. 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**, v. 32, p. 443–453, 2014.

HAN, J; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**. 3 ed. Waltham: Morgan Kaufmann Publishers, 2012.

LIAO, C.-W.; PERNG, Y.-H. Data mining for occupational injuries in the Taiwan construction industry. **Safety Science**, v. 46, n. 7, p. 1091–1102, 2008.

MUTLU, N. G.; ALTUNTAS, S. Risk analysis for occupational safety and health in the textile industry: Integration of FMEA, FTA, and BIFPET methods. **International Journal of Industrial Ergonomics**, v. 72, p. 222-240, 2019.

PALAMARA, F.; PIGLIONE, F.; PICCININI, N. Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. **Safety Science**, v. 49, n. 8–9, p. 1215–1230, 2011.

PIMENTA, A. A., PORTELA, A. R. M. R., OLIVEIRA, C. B. O. RIBEIRO, R. M. A Bibliometria as Pesquisas Acadêmicas. **Revista de Ensino, Pesquisa e Extensão**, v. 4, n. 7, 2017.

QUEVEDO-SILVA, F., SANTOS, E. B. A., BRANDÃO, M. M., VILS, L. Estudo Bibliométrico: Orientações Sobre Sua Aplicação. **Revista Brasileira De Marketing**, v. 15, n. 2, p. 246-262, 2016.

RUSO, J.; STOJANOVIĆ, V. Occupational health and safety using data mining. **International Journal for Quality Research**, v. 6, n. 4, 2012.

SÁNCHEZ-HERRERA, I. S.; DONATE, M. J. Occupational safety and health (OSH) and business strategy: The role of the OSH professional in Spain. **Safety Science**, v. 120, p. 206–225, 2019.

SANMIQUEL, L.; ROSSELL, J. M.; VINTRÓ, C. Study of Spanish mining accidents using data mining techniques. **Safety Science**, v. 75, p. 49–55, 2015.

SARKAR, S.; MAITI, J. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. **Safety Science**, v. 131, p. 104900, 2020.

SHIN, D.-P., YOUNG-JUN, P., SEO, J., DONG-EUN, L. Association Rules Mined from Construction Accident Data. **KSCE Journal of Civil Engineering**, v. 22, n. 4, p. 1027-1039, 2018.

TIXIER, A. J.-P., HALLOWELL, M. R., RAJAGOPALAN, B. BOWMAN, D. Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining. **Automation in Construction**, v. 74, p. 39, 2017.

WANG, Y.; CHEN, H; LIU, B; YANG, M; LONG, Q. A Systematic Review on the Research Progress and Evolving Trends of Occupational Health and Safety Management: A Bibliometric Analysis of Mapping Knowledge Domains. **Frontiers in Public Health**, v. 8, 2020.

WITTEN, I., FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 4 ed. São Francisco: Morgan Kaufmann, 2016.

YANAR, B.; LAY, M.; SMITH, P. M. The Interplay Between Supervisor Safety Support and Occupational Health and Safety Vulnerability on Work Injury. **Safety and Health at Work**, v. 10, n. 2, p. 172–179, 2019.

YILMAZA, F., ÇELEBIB, U. B. The Importance of Safety in Construction Sector: Costs of Occupational Accidents in Construction Sites. **Business and Economics Research Journal**, v. 6, n. 2, p. 25-37. 2015.

ZHANG, D; JIANG, K. Application of Data Mining Techniques in the Analysis of Fire Incidents. **Procedia Engineering**, v.43, p. 250–256, 2012.

ZHAO, Y., ZHANG, C., ZHANG, Y., WANG, Z., LI, J. Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise. **Ear and Hearing**, v. 40, n. 3, p. 690–699, 2019.

Appendix

Appendix A - Table with the articles selected in the systematic mapping, their respective identification and citation codes.

Study ID	Title of Articles	Citation
S1	Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach	Bevilacqua *et al*., 2008
S2	Data mining for occupational injuries in the Taiwan construction industry	Liao e Perng, 2008
S3	Prioritizing Health Promotion Plans with k-Bayesian Network Classifier	Ueno *et al*., 2008
S4	Classification and prediction of occupational injury risk using soft computing techniques: An Italian study	Ciarapica e Giacchetta, 2009
S5	Application of data mining in classification analysis of safety accidents based on alternate covering neural network	Qu, 2009
S6	Signal processing and machine learning for real-time classification of ergonomic posture with unobtrusive on-body sensors; application in dental practice	Olsen *et al*., 2009
S7	Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry	Cheng *et al*., 2010
S8	Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases	Palamara *et al*., 2011
S9	Spatial clustering applied to health area	Valêncio *et al*., 2011
S10	Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry	Cheng *et al*., 2012
S11	Application of Pharmacovigilance Methods in Occupational Health Surveillance: Comparison of Seven Disproportionality Metrics	Bonneterre *et al*., 2012
S12	Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims	Bertke *et al*., 2012
S13	Occupational Health and Safety using Data Mining	Ruso *et al*., 2012
S14	Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry	Cheng *et al*., 2013
S15	Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database	Nenonen, 2013
S16	Analytical study using data mining for periodical medical examination of employees	Waghmare and Pai, 2013

S17	Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers	Gross *et al*., 2013
S18	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	Hajakbari e Minaei-Bidgoli, 2014
S19	Office workers syndrome monitoring using kinect	Paliyawan *et al*., 2014
S20	Near-miss narratives from the fire service: A Bayesian analysis	Taylor *et al*., 2014
S21	An information fusion framework for context-based accidents prevention	Sanchez-Pi *et al*., 2014
Study ID	**Title of Articles**	**Citation**
S22	Study of Spanish mining accidents using data mining techniques	Sanmiquel *et al*., 2015
S23	Decision tree analysis of construction fall accidents involving roofers	Mistikoglu *et al*., 2015
S24	A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining	Saâdaoui *et al*., 2015
S25	Assessing ergonomic and postural data for pain and fatigue markers using machine learning techniques	Shein *et al*., 2015
S26	Assessment of Risk of Musculoskeletal Disorders among Crane Operators in a Steel Plant: A Data Mining-Based Analysis	Krishna *et al*., 2015
S27	Data-mining and expert models for predicting injury risk in ski resorts	Bohanec e Delibasic, 2015
S28	Workplace accidents analysis with a coupled clustering methods: S.O.M. and K-means algorithms	Comberti *et al*., 2015
S29	Bayesian decision support for coding occupational injury data	Nanda *et al*., 2016
S30	Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning	Chokor *et al*., 2016
S31	A novel hidden danger prediction method in cloud-based intelligent industrial production management using timeliness managing extreme learning machine	Luo *et al*., 2016
S32	Automatic Detection of Helmet Uses for Construction Safety	Rubaiyat *et al*., 2016
S33	Text mining based safety risk assessment and prediction of occupational accidents in a steel plant	Sarkar *et al*., 2016
S34	Classifying construction site photos for roof detection	Siddula *et al*., 2016
S35	Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review	Marucci-Wellman *et al*., 2017

S36	Coupling risk attitude and motion data mining in a preemtive construction safety framework	Rashid *et al*., 2017
S37	Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining	Tixier *et al*., 2017
S38	Predictive model for incident occurrences in steel plant in India	Sarkar *et al*., 2017
S39	Construction accident narrative classification: An evaluation of text mining techniques	Goh e Ubeynarayana, 2017
S40	Bayesian Decision Tool for the Analysis of Occupational Accidents in the Construction of Embankments	Gerassis *et al*., 2017
S41	Safety in ready mixed concrete industry: Descriptive analysis of injuries and development of preventive measures	Akboğa e Baradan, 2017
S42	A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making	Comberti *et al*., 2018
S43	Wearable insole pressure system for automated detection and classification of awkward working postures in construction workers	Antwi-Afari *et al*., 2018
Study ID	**Title of Articles**	**Citation**
S44	Application of Inertial Measurement Units for Advanced Safety Surveillance System Using Individualized Sensor Technology (ASSIST): A Data Fusion and Machine Learning Approach	Baghdadi, 2018
S45	Does background really matter? Worker activity recognition in unconstrained construction environment	Jiang *et al*., 2018
S46	Estimation of probability of harm in safety of machinery using an investigation systemic approach and Logical Analysis of Data	Jocelyn *et al*., 2018
S47	Association Rules Mined from Construction Accident Data	Shin *et al*., 2018
S48	A Bayesian assessment of occupational health surveillance in workers exposed to silica in the energy and construction industry	Abad *et al*., 2018
S49	Evaluation of genotoxic effects in Brazilian agricultural workers exposed to pesticides and cigarette smoke using machine-learning algorithms	Tomiazzi *et al*., 2018
S50	Prediction of Occupational Incidents Using Proactive and Reactive Data: A Data Mining Approach	Sarkar *et al*., 2018
S51	A Bayesian Network Application in Occupational Health and Safety	Pekel *et al*., 2018
S52	Effect of a job demand-control-social support model on accounting professionals' health perception	Del Pozo-Antúnez *et al*., 2018

S53	Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques	Lee e Kim, 2018
S54	Analysis of occupational accidents in underground and surface mining in Spain using data-mining techniques	Sanmiquel *et al*., 2018
S55	Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran	Shirali *et al*., 2018
S56	Applying machine learning to workers' compensation data to identify industry-specific ergonomic and safety prevention	Meyers *et al*., 2018
S57	Application of optimized machine learning techniques for prediction of occupational accidents	Sarkar *et al*., 2019c
S58	Predicting types of occupational accidents at construction sites in Korea using random forest model	Kang e Ryu, 2019
S59	Evaluating machine learning performance in predicting injury severity in agribusiness industries	Kakhki *et al*., 2019
S60	Discovering Latent Psychological Structures from Self-Report Assessments of Hospital Workers	Kao *et al*., 2019
S61	Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data	Heo *et al*., 2019
S62	Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke	Tomiazzi *et al*., 2019
S63	Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction	Di Noia *et al*., 2019
S64	Intelligent Wearable Occupational Health Safety Assurance System of Power Operation	Xie e Chang, 2019
Study ID	**Title of Articles**	**Citation**
S65	An optimization-based decision tree approach for predicting slip-trip-fall accidents at work	Sarkar *et al*., 2019b
S66	Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study	Zhao *et al*., 2019
S67	Text-clustering based deep neural network for prediction of occupational accident risk: A case study	Sarkar *et al*., 2019a
S68	Decision support approach to occupational safety using data mining	Khosrowabadi and Ghousi, 2019