Review

A Review of Data Mining Strategies by Data Type, with a Focus on Construction Processes and Health and Safety Management

Antonella Pireddu *, Angelico Bedini, Mara Lombardi, Angelo L.C. Ciribini, Davide Berardi

Posted Date: 7 May 2024

doi: 10.20944/preprints202405.0322.v1

Keywords: Clustering; Principal Component Analysis (PCA); Meta-Analysis; Construction Industry; Data Mining; Machine Learning; Prediction Models; Workplaces Safety; Smart Technology (ST); State-of-the-art

Review

A Review of Data Mining Strategies by Data Type, with a Focus on Construction Processes and Health and Safety Management

Antonella Pireddu [1,*], Angelico Bedini [1], Mara Lombardi [2], Angelo L. C. Ciribini [3] and Davide Berardi [2]

[1] Department of Technological Innovations and Safety of Plants, Products and Anthropic Settlements (DIT), Italian National Institute for Insurance against Accidents at Work, Inail, Rome 00144, Italy
[2] Department of Chemical Engineering Materials Environment (DICMA), Sapienza-University of Rome, Rome 00184, Italy
[3] Department of Civil Engineering, Architecture, Land, Environment and Mathematics (DICATAM), Brescia University, Brescia 25121, Italy
* Correspondence: an.pireddu@inail.it

Abstract: Increasingly, information technology facilitates the storage and management of data useful for risk analysis and event prediction. Studies on data extraction related to occupational health and safety are increasingly available; however, due to its variability, the construction sector warrants special attention. This review is conducted under the research programmes of the National Institute for Occupational Accident Insurance (Inail). **Objectives:** The research question focuses on identifying which data mining (DM) methods, among supervised, unsupervised, and others, are most appropriate to be applied to certain investigation objectives, types, and sources of data, as defined by the authors. **Methods:** Scopus and ProQuest were the main sources from which we extracted studies in the field of construction, published between 2014 and 2023. The eligibility criteria applied in the selection of studies, were based on the Preferred Reporting Items for Systematic Review and meta-analysis (PRISMA). For exploratory purposes, we applied hierarchical clustering, while for in-depth analysis, we use principal component analysis (PCA) and meta-analysis. **Results:** The search strategy based on the PRISMA eligibility criteria, provided us with 61 out of 2,234 potential articles, 202 observation, 91 methodologies, 4 survey purposes, 3 data sources, 7 data types, and 3 resource type**.** Cluster analysis and PCA organized the information included in the paper dataset into two dimensions and labels: "supervised methods, institutional dataset, and predictive and classificatory purposes" (correlation 0.97÷8.18E-01; p-value 7.67E-55÷1.28E-22) and the second, Dim2 "not-supervised methods; project, simulation, literature, text data; monitoring, decision-making processes; machinery and environment" (corr. 0.84÷0.47; p-value 5.79E-25÷3.59E-06). We answered the research question regarding which method, among supervised, unsupervised, or other, is most suitable for application to data in the construction industry. **Conclusions**: The meta-analysis provided an overall estimate of the better effectiveness of supervised methods (Odds Ratio = 0.71, Confidence Interval 0.53÷0.96) compared to not-supervised methods.

Keywords: clustering; principal component analysis (PCA); meta-analysis; construction industry; data mining; machine learning; prediction models; workplaces safety; smart technology (ST); state-of-the-art

1. Introduction

The activities attributable to the construction sector according to the International Labour Organisation (ILO) classification are as follows: i) building, including excavation and the construction, structural alteration, renovation, repair, maintenance (including cleaning and painting) and demolition of all types of buildings or structures; ii) civil engineering, including excavation and the construction, structural alteration, repair, maintenance and demolition of structures such as airports, docks, harbours, inland waterways, dams, river and avalanche and sea defence works, roads

 doi:10.20944/preprints202405.0322.v1

and highways, railways, bridges, tunnels, viaducts and works related to the provision of services such as communications, drainage, sewerage, water and energy supplies; and iii) the erection and dismantling of prefabricated buildings and structures, as well as the manufacturing of prefabricated elements on the construction site [30].

Construction safety research is numerous and motivated by the alarming rates of accident and fatalities, focusing on two perspectives: management and technology [77]. In general, workplace safety management is based on organisational and technological strategies. Construction safety standards and accident reduction are achieved through information and worker training, aiming to enhance the level of risks perception associated with the production process. However, the impact of traditional accident prevention strategies has been limited due to their reactive and regulatory nature [45,77]. A relevant aspect is the increased risks associated with the organisation and production goals of construction companies.

According to Razi et al. [51], Artificial Intelligence (AI), is a broad field of computer science concerned with the developing intelligent robots capable of performing tasks that traditionally require human intellect. AI plays a crucial role in assisting construction supervisors in minimizing accidents, supporting project efficiency, and significantly improving operational safety. Alongside the advancement of information and communication technology, various innovative technologies have been investigated to aid and improve on existing management-driven safety management practices. Besides the aid of technologies, new injury prevention strategies have been developed for the construction industry. The risk analysis method is one of them which are used in safety programs to improve safety performance. A relevant factor is the relationship between the type of construction project and the type of accident.

Data mining methods are applicable in various fields, dealing with different types of data and objectives. Studies focusing on DM techniques applied to construction safety date back to no later than 2014 [52]. The study has been developed as part of the 2022-2024 Inail's research program and the objective “Study of the effectiveness and efficiency indices related to innovative technologies aimed at preventing the risk of injury in highly variable work environments”. The protocol of review is led by the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA) [50]. Section 1 of the article introduces the background and objectives of the investigation. Section 2 describes the materials and methods, while Section 3 presents the results obtained from the applying cluster analysis, PCA, and meta-analysis. Section 4 offers an extensive discussion of the results, considering the current state of the art and our future goals. Finally, Section 5 summarises the salient results achieved in this review.

2. Materials and Methods

The set of articles published from 2014 to September 2023, which were useful for the purposes of this review, was extracted from Scopus [9,20] and ProQuest [15]. Authoritative sites on conferences in the field of computer science and DM and Management in Construction field were queried; however, only Web of Conference provided an eligible contribution for the purposes for our review.

2.1. Selection and Inclusion Criteria

All searches were conducted using a combination of subject headings and free-text terms. We focused exclusively on peer-reviewed articles, conference papers and book chapters. The topics included were: “Machine Learning,” “Work,” “Construction,” and “Risk,” accross the following subject areas: (i) Engineering, (ii) Social and Environmental Sciences, (iii) Computational Sciences. The criteria applied in the search strategy are defined in Table 1. The final search strategy was developed through several preliminary searches including (i) articles, (ii) conference papers, (iii) book chapters.

 doi:10.20944/preprints202405.0322.v1

Table 1. Query input for document search inclusion criteria Source: Scopus and ProQuest data.

Stream	Query
TIT-ABS-KEY	Machine Learning AND Work AND Construction AND Risk
SUBJECT AREA	Engineering AND Social Science AND Environmental Science AND Computational Science
PUBLICATION YEAR	From 2014 to September 2023
DOCUMENT	Article, Conference Paper, Book chapter (Peer reviewed)
LANGUAGE	Not restriction

Figure 1 summarise the result of the PRISMA document selection process. The collected paper dataset includes the information on Authors, Title, Year of publication, Source of title, Volume, Issue, Number of Pages, Citation number, DOI, Affiliations Author information, Abstract, Keywords, Type of Publication, and further information. Three authors (AP, AB, and DB) independently reviewed the titles and abstracts to assess the eligibility of all studies. We applied PRISMA procedures and checklists [50] in identifying topic, datasets and keywords and filtering content according to abstract, assessing the eligibility of publications in the research scope. Further insights were made into the selected articles by conductingfull-text review and analysing the content for search purposes (see Figure 1) [23]. Disagreements were resolved by a fourth evaluator (ML) until consensus was reached between the authors. Only studies that met the eligibility criteria were included.

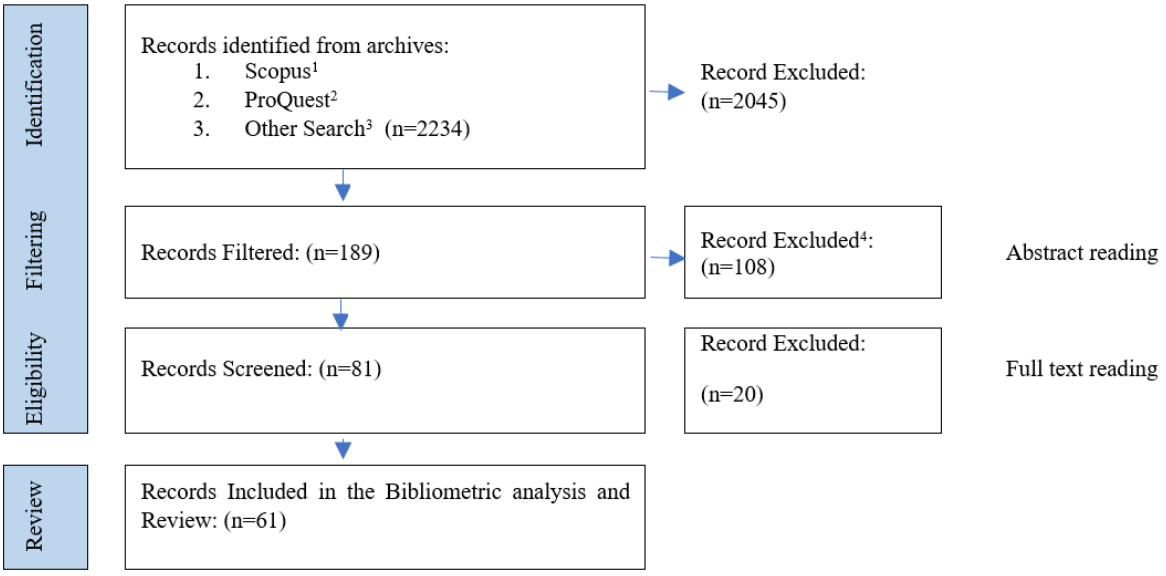

Figure 1. PRISMA criteria for the selection of documents and eligibility flow-chart. Source: Scopus and ProQuest data. 2014-2023.

2.2. Risk of Bias for Selected Studies

The risk in non-randomised studies was assessed on the basis of the following biases: (1) due to confounding, (2) in the selection of the types of data in the study, (3) in the classification of the study objective, (4) due to missing data, (5) in the measurement of outcomes, (6) in the evaluation metrics and (7) in the selection of the reported outcome. Each individual study included was assessed as having a low, moderate, severe, and critical risk of bias. If critical information was missing for the assessment of risk of bias, these studies were considered devoid of information.

 doi:10.20944/preprints202405.0322.v1

2.3. Data Quality and Items

The titles and abstracts of the identified studies were independently checked at two different points in time. Eligibility and inclusion criteria were initially assessed on a subset of 30 studies before searching all databases. Decisions were made by examining both the abstracts and the full texts. Only studies that were complete and met all inclusion criteria were included in the qualitative and quantitative synthesis. The information and data included in the papers obtained through the PRISMA method, were then included in the review.

2.3. Study Design

The scientific articles falling under the eligibility criteria of PRISMA were pre-processed to extract information suitable for the purpose of review. The 61 papers included in the review were categorized by 33 source titles and by publication year (Table 2). The bibliometric analysis involved a review of the global literature and geographic mapping worldwide (Figure 2). The cluster analysis (HC) was used to find the best aggregations between groups. By the Silhouette index, it was found the best degree of aggregation to be represented in a cluster plot based on correlations and variances (Figures 3 and 4). Principal component analysis (PCA) was useful to find the correlation classes between the various parameters of the dataset in a simplified reading of the results. Through PCA, we reduced the items and obtained the extent of correlation between variables, methods, and components. The meta-analysis of this classes was useful to estimate the reliability of HC and PCA results and the odds ratios OR and confidence intervals CI of groups of items. Spatial data collection, analysis, classification, and bibliometric analysis were performed with VOS viewer [65], R and GIS software.

Table 2. Papers included in the review by source. Years: 2014-2023 (September). Source: Scopus, ProQuest data.

Source Title	Author	Papers
Accident Analysis and Prevention	Goh et al., 2017	1
Advances in Civil Engineering	Lim et al. 2022	1
Applied Sciences	Hoła et Szóstak, 2019	1
Applied Sciences (Switzerland)	Bai et al., 2022, Lee et al., 2020, Liu et al., 2022, Zhang J. et al., 2020	4
Applied Soft Computing	Lin et al., 2023	1
Automation in Construction	Antwi-Afari et al., 2022, Choo et al., 2023, Tixier et al., 2016, Tixier et al., 2017, Yu et al., 2019, Zhang F. et al., 2019	6
Buildings	Al-Kasasbeh et al. 2022, Dutta et al., 2023, Gao et al. 2022, Liu et al. 2023, Ankit et al. 2023, Maqsoom et al. 2023, Numan et al., 2023, Shuang et al., 2023, Toğan et al. 2022, Wang J. et al., 2022, Yin et al., 2023	11
Chinese Journal of Mechanical Engineering (English Ed.)	Li L.et al., 2017	1
Civil and Environmental Engineering	Erzaij et al., 2021	1
Computer-Aided Civil and Infrastructure Engineering	Li X et al., 2023	1
E3S Web of Conferences	Passmore et al., 2019	1
Engineering, Construction and Architectural Management	Duan et al., 2023	1

 doi:10.20944/preprints202405.0322.v1

IEEE Access	Leng et al., 2021, Lin et al., 2014	2
IEEE Robotics and Automation Letters	Osa et al., 2023	1
International Journal of Computational Methods and Experimental Measurements	Fernández et al., 2023	1
International Journal of Environmental Research and Public Health	Aminu Darda'u et al., 2023, Khairuddin et al., 2022 Sadeghi et al., 2020, Yedla et al., 2020	4
IOP Conference Series. Earth and Environmental Science	Yao et al., 2022, Razi et al., 2023	2
Journal of Civil Engineering and Management	Wei, 2021	1
Journal of Safety Research	Goldberg, 2022	1
Lecture Notes in Civil Engineering	Jha et al., 2023, Sapronova et al., 2023	2
Mathematical Problems in Engineering Volume	Zhang X. et al., 2020	1
PLoS One	Ensslin et al., 2022	1
Rock Mechanics and Rock Engineering	Hasanpour et al., 2015	1
Safety Science	Alkaissy et al., 2023, Wang F. et al., 2016, Zermane et al. 2023	3
Scientific Programming	Zhao et al., 2022	1
Sensors (Switzerland)	Dong et al., 2021	1
Sustainability	Alateeq et al. 2023, Alhelo et al., 2023, Muhammad et al., 2023, Topal & Atasoylu, 2022	4
Sustainability (Switzerland)	Mostofi et al., 2022, Yan et al., 2022, Zhu&Liu, 2023	3
Visualization in Engineering	Schindler et al., 2016	1
Wireless Communications and Mobile Computing	Kumari et al., 2022	1
Total		61

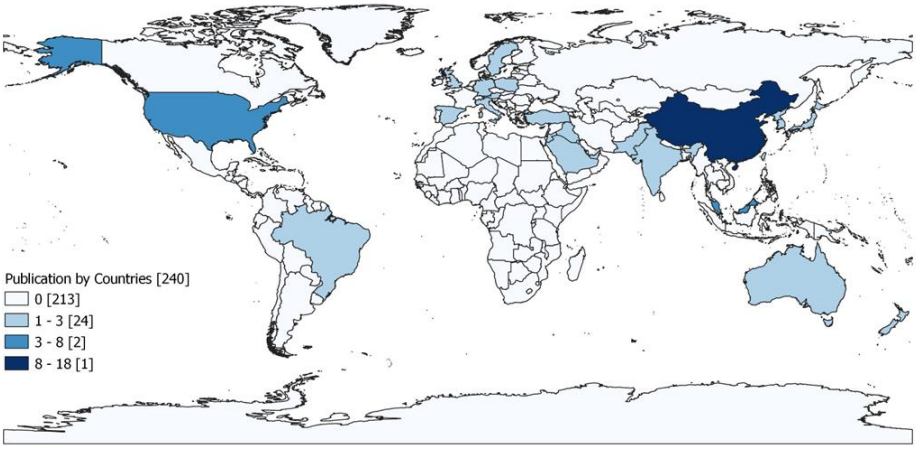

Figure 2. Map of papers included in the review by country of origin. Years 2014-2023 (September). Source: Author's processing from Scopus and ProQuest data. GIS and VOS viewer.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2024 doi:10.20944/preprints202405.0322.v1

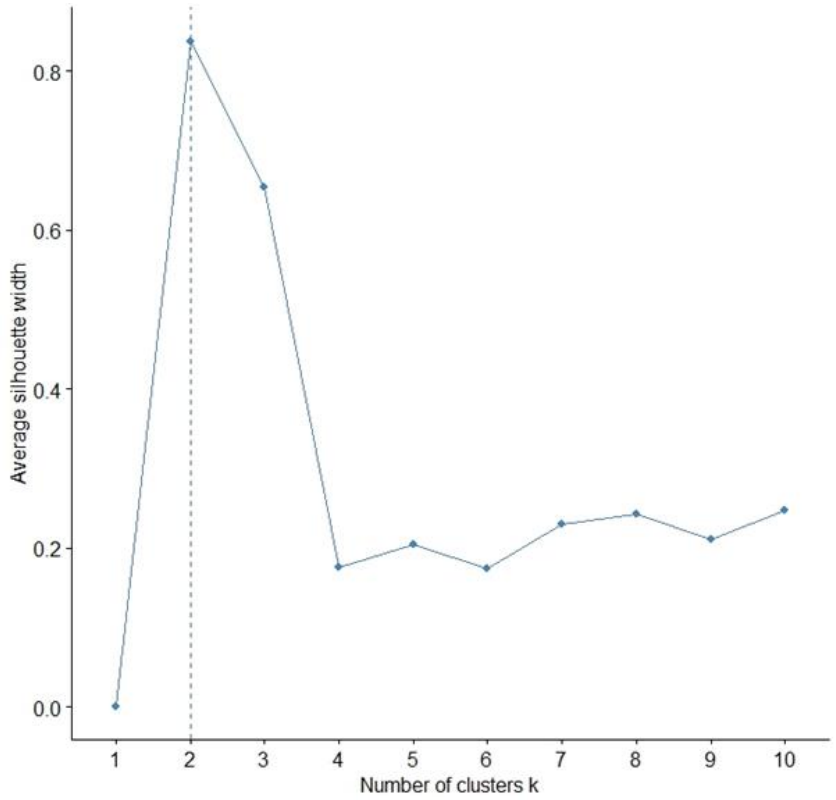

Figure 3. Silhouette test. Representation of optimal number of clusters. Years: 2014-September 2023. Source: Author's processing from Scopus and ProQuest data.

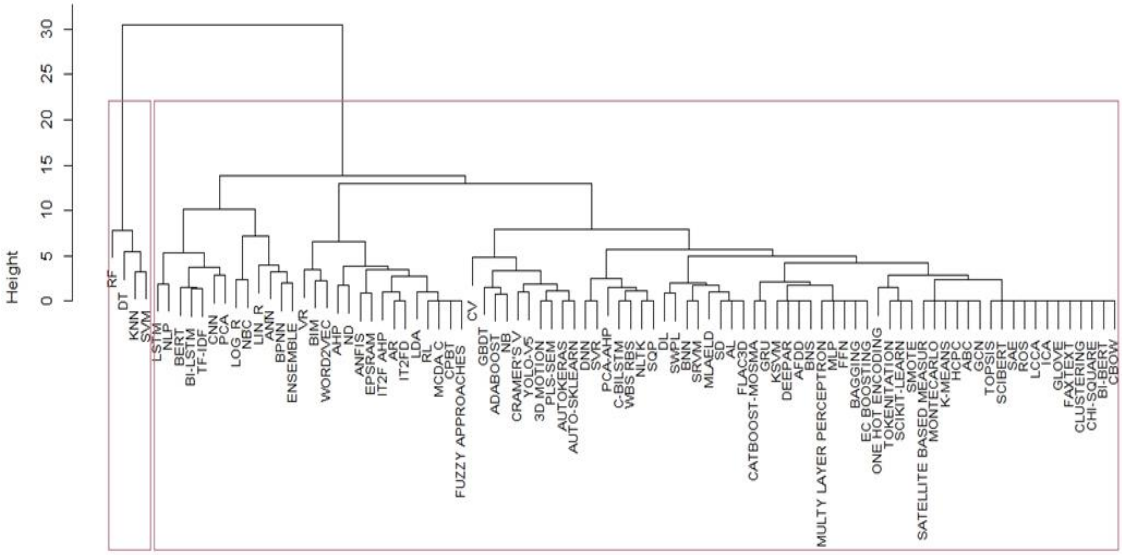

Figure 4. Hierarchical cluster dendrogram. Euclidean distance VS height. 91DM methods by cluster [hclust (*, "Ward.D2")]. Years: 2014-September 2023. Source: Author's processing from Scopus and ProQuest data.

3. Results

3.1. Study Selection and Bibliometric Analysis

The search strategy, based on the PRISMA eligibility criteria, yielded 61 papers that were included in the review and categorized by 33 source titles, (Table 2) and publication year. Regarding

 doi:10.20944/preprints202405.0322.v1

the latter, there was an increasing trend in publication from 2014 to September 2023, where the articles recorded the following trend: 1 paper each in 2014 and 2015, 3 papers in 2016 and 2017, 4 papers in 2019, 6 in 2020, 3 papers in 2021, 17 papers in 2022 and 23 papers in 2023.

Figure 2 shows a map of the 61 scientific articles included in the review from 2014 to September 2023, by country of origin, worldwide.

3.2. Classes of Data and DM Methods

In a separate dataset, study objective, type of data under investigation, applied DM methods, applied DM type (supervised, unsupervised, other), validation metrics (if available), the DM method found to be most effective, number of rows and columns in the dataset used by the authors (if available). The 61 selected articles provided 202 observations, 91 DM methods (50 of which were considered the best method), 4 survey purposes, 3 field, 7 data types, 3 resource type (as detailed in Table 3 and Appendices A and B). Among study objectives, we obtained X1 classifying (18%), X2 decision-making (15%), X3 monitoring (16%), X4 predicting (51%). As a field we obtained: X5 construction process (38%), X6 occupational accident (34%), X7 risk management process (28%). The types of data investigated were: X8 construction project (5%), X9 institutional dataset (70%), X10 interview report (2%), X11 literature data (3%), X12 narrative text (6%), X13 signal (10%), X14 simulation (4%). Regarding the types of DM method, we found: X15 supervised method (58%), X16 unsupervised method (24%), X17 other method (18%). As a resource type we found: X18 process (63%), X19 environment resource (15%), X20 plant and machinery resource (22%).

Table 3. Classification of content included in the 61 articles selected by the PRISMA method. Years: 2014-2023. Author's processing from Scopus and ProQuest data.

CLASS	N	DESCRIPTION	INDEX
DM METHODS	91	Appendix A	
STUDY OBJECTIVE	4	classifying, decision making, monitoring, predicting	X1-X4
FIELD	3	construction process, occupational accident, H&S management	X5-X7
DATA TYPE	7	project, institutional data, interview, literature, text, signal & video, simulation	X8-X14
DM TYPE	3	supervised, unsupervised, other	X15-X17
RESOURCE TYPE	3	PROCESS, ENVIRONMENT, MACHINERY	X18-X20

3.3. Cluster Analysis

Clustering is a significant approach in DM that aims to identify groups within data sets. In real-world applications, both numeric and categorical features are often used to define the data. Clustering analysis is one of the most important approaches in DM, and it seeks to find the nature of groupings or clusters of data objects within an attribute space [8,11,16]. For an exploratory approach, we applied clustering analysis to the dataset in Appendix B. With this unsupervised ML approach, the algorithm processes input data and generates a sequence of cluster based on relational similarities with surrounding data points. The questions to answer in this DM method are: "when do we stop combining clusters?", "How do we represent clusters?". By applying Hierarchical clustering (HC) and appropriate indexes, we identified the optimal number of clusters of our data.

The "silhouette" index provided the best determination of cluster number; the highest average silhouette width indicates the optimal number of clusters [13]. The concept of silhouette width represented involves the difference between the within-cluster tightness and separation from the rest. Specifically, the silhouette width s_i for entity $i \in I$ is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (1)$$

 doi:10.20944/preprints202405.0322.v1

where "a_i" is the average distance between "i" and all other entities in the cluster to which "i" belongs and "b_i" is the minimum of the average distances between "i" and all entities in every other cluster. Silhouette width values are between -1 and 1. If the silhouette width value for an entity is approximately zero, it means that the entity could also be assigned to another cluster. If the silhouette width value is close to -1, it means that the entity has been incorrectly classified. If all silhouette width values are close to 1, it means that the set "i" is well clustered. The silhouette index suggests that the best aggregation of the Appendix B dataset is two clusters having obtained an index greater than 0.7 (Figure 3).

We created the item groupings through an iterative hierarchical process of aggregating pairs of "most similar" groups of methods by calculating the dissimilarity ("distance," for triangular inequality). Thus, we obtained the dendrogram in which are represented the Euclidean distance between the elements, the similarity, and the shape of the clusters.

3.4. Principal Component Analysis (PCA)

The objective of PCA is to identify suitable Y linear transformations of the observed variables that are easily interpretable and capable of highlight and synthesise the information inherent in the initial matrix X [11,16]. This tool is particularly useful when dealing with a considerable number of variables from which one wants to extract information as possible while working with a smaller set of variables. The source data are organised in a matrix, denoted X, where the columns stand for the p observations made and the rows are the p variables considered for the phenomenon under analysis.

$$X = (X_1, X_2 \dots X_p)^T \quad (2)$$

The source data matrix is synthetically represented by a multivariate random vector. Given a matrix X which holds p interrelated variables, we obtained a matrix of new data Y, consisting of p interrelated variables to each other, which turn out to be linear combinations of the former. Each principal component can be expressed as follows:

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & l_{1p} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pp} \end{bmatrix} \cdot \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{p1} & \cdots & X_{pp} \end{bmatrix} \quad (3)$$

$$\vec{Y} = l_{ij}X_1 + l_{ij}X_2 + \cdots l_{ip}X_p \text{ where } \quad i=1, 2\dots p \quad (4)$$

The generic coefficient l_{ij} is the weight that the variable Xj has in finding the principal component Yi (with $i = 1, 2, k, p$). The more the larger l_{ij} is (in absolute value), the greater the weight that the values Xj ($j = 1, 2, k, p$) have in deciding a given principal component. The principal component Yi will be most strongly characterised by the variables Xj to which correspond the largest absolute coefficients l_{ij}.

Once the significance of the correlation between the different variables had been found from Formula 5, the PCA principal component analysis was applied, leaving out the outliers between the DM methods, the most important of which were analysed separately.

$$r_{xy} = Corr(y_i, x_j) = e_{ij}\frac{\sqrt{\lambda i}}{\sigma j} \quad (5)$$

The outcome of the PCA resulted in 11 components of which only the two principals of the total explained the total variance.

3.4.1. Inertia Distribution

The dataset contains 91 individuals corresponding to DM methods and 20 variables. Analysis of the graphs reveals no outlier. The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied. The first two

 doi:10.20944/preprints202405.0322.v1

dimensions of analyse express 69.71% of the total dataset inertia; that means that 69.71% of the individuals (or variables) cloud total variability is explained by the plane. This percentage indicates that the first plane effectively represents the data's variability.

The first factor is the main one: it expresses 57.36% of the variability of the data (Figure 5). In this case, the variability relating to the other components may be less significant, despite the high percentage. The first axis has a higher amount of inertia than the 0.95 quadrant of the random distributions. This observation suggests that only two axes carry factual information. Consequently, the description will stick to these axes.

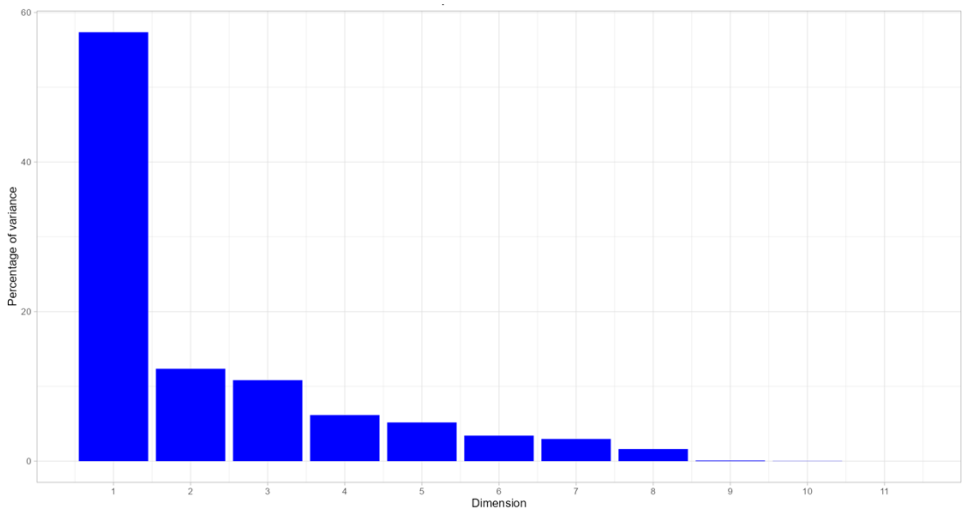

Figure 5. Decomposition of the total inertia by axes. Dimension VS Percentage of variance.

The criteria for selecting dimensions in the final model are threefold: the Kaiser rule, where eigenvalues are greater than 1 (Table 4); the proportion of variance explained by the components at least equal to 60%-80% of the overall variability (Table 4); the Cattell rule, according with it the right number of components corresponds at the elbow or change of slope in the component-eigenvalue graph (Figure 5). From these observations, it should be better to also interpret the dimensions greater or equal to the second one. The above criteria allowed us to assign a "label" to each component.

Table 4. Eigenvalues, percentage of variance and cumulative percentage of variance (in bolt the significant values).

Dim	**eigenvalue**	**% of variance**	**cumulative % of variance**
Dim1	**6.31**	**57.36**	**57.36**
Dim2	**1.36**	**12.35**	**69.71**
Dim3	1.19	10.83	80.54
Dim4	0.68	6.16	86.71
Dim5	0.57	5.18	91.88
Dim6	0.38	3.42	95.30
Dim7	0.33	2.97	98.27
Dim8	0.18	1.62	99.89
Dim9	0.01	0.09	99.97
Dim10	0.00	0.03	100.00
Dim11	0.00	0.00	100.00

3.4.2. Axes Description

Dimension 1 opposes individuals such as dt (32), knn (49), svm (81) and rf (69), to the right of the graph characterised by a strongly positive coordinate on the axis, to individuals such as MCDA C (58), characterised by a strongly negative coordinate on the axis (to the left of the graph).

 doi:10.20944/preprints202405.0322.v1

Dimension 2 opposes individuals such as lstm (54), word2vec (88), nlp (63) and BIM (16), who at the top of the graph, and characterised by a low positive co-ordinate on the axis, with individuals such as ann (8), adaboost (3), who have low negative coordinate on the axis and are located at the bottom of the graph.

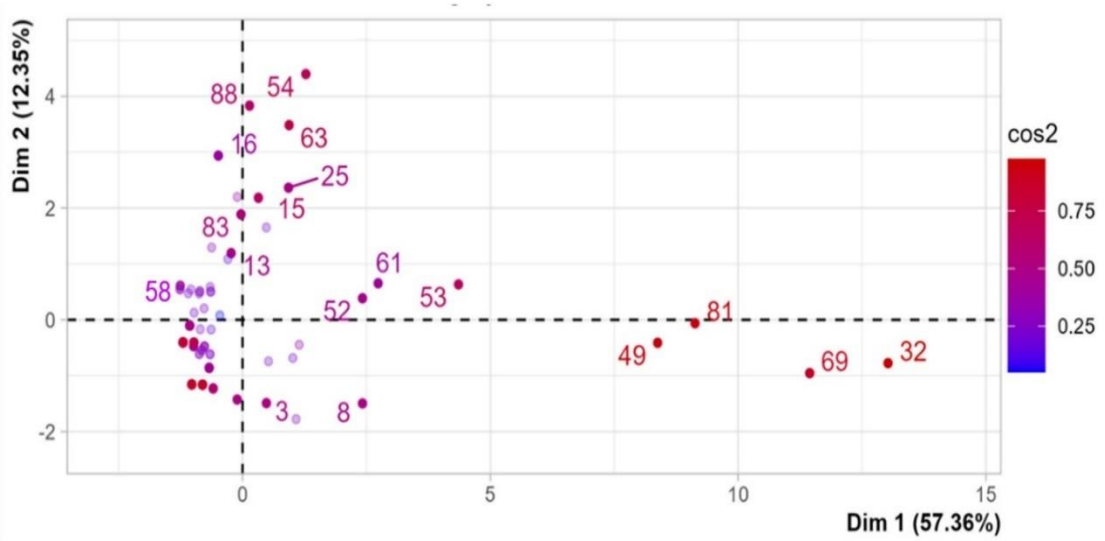

Figure 6. PCA. The graph of individual (DM Methods). Dim1 vs Dim2 (correlation or cos^2>0.4).

The Dim1, group 1 (dt , knn, svm and rf) is sharing high values for the variables "predicting", "supervised", "monitoring", "frequency", "institutional data", "data project-simulation-signal", "classifying", "best method and "interview-literature-text" (variables are sorted from the strongest). The group 2 characterised by a negative coordinate on the axis, the individual MCDA C (58) is sharing low values for the variables "interview-literature-text", "classifying", "frequency", "institutional data", "monitoring", "predicting", "supervised", "project-simulation-signal", "best method" and "other methods (variables are sorted from the weakest). The variables "supervised" and "freq." are highly correlated with this dimension (respective correlation of 0.94, 0.98). These variables could therefore summarize themselves the dimension 1. The Dim2, group 1 shares high values for the variables "not-supervised" and "decision-making", while the group 2 with "monitoring", "machinery" and "environment" (Tables 5 and 6).

Table 5. PCA. Axes description and correlation between axes, methods, and variables (cos^2>0.4). Years: 2014-2023. Source: Author's processing from Scopus and ProQuest data. R.

Axes	(+)	(-)	DM Class	Study Objective	Data type	Resource type
Dim1	dt (32), knn (49), svm (81) and rf (69)	MCDA C (58)	supervised	classifying predicting	institutional data, interview-literature-text	-
Dim2	lstm (54), word2vec (88), nlp (63), BIM (16)	ann (8), adaboost (3)	other-supervised (not-supervised)	decision making monitoring	project-simulation-signal; interview-literature-text	machinery environment

 doi:10.20944/preprints202405.0322.v1

Table 6. PCA. Axes description, correlation between methods and axes (in bolt the correlation or cos^2>0.4). Source: Author's processing from Scopus and ProQuest data. R.

Dim1	Correlation (cos^2)	p.value	Dim2	Correlation (cos^2)	p-value
frequency	**9.874E-01**	1.603E-71	other type	**8.413E-01**	5.790E-25
supervised	**9.694E-01**	7.675E-55	decision making	**5.077E-01**	3.801E-07
institutional data	**9.412E-01**	8.809E-43	interview-literature-text	**4.688E-01**	3.593E-06
predicting	**9.361E-01**	2.984E-41	classifying	3.060E-01	3.547E-03
classifying	**8.181E-01**	1.286E-22			

According to the correlation method-variable and axes, the x-axis (Dim1) can then be renamed "Supervised methods (dt, knn, svm and rf), applied to institutional data in classifying and making inference (predicting)". The y-axis (Dim2) can instead be renamed "Not-supervised methods (lstm, word2vec, nlp, BIM) applied to project, simulation signal, interviews, literature, or textual data in making decisions and classifying.

3.5. Meta-Analysis

The data from the complete collection of studies selected according to the PRISMA methodology, aggregated according to the classes defined in Table 1, allowed us to derive a single conclusive result that answered our research question. Through meta-analysis, we assessed whether supervised methods were more effective than not-supervised ones, for the various classes. The forest plot summarises the results of the meta-analysis, which include the OR with its CIs, the sample size weight, the heterogeneity of the data and a quantitative, whole-data assessment of the effectiveness of the treatment with supervised methods (Figure 7).

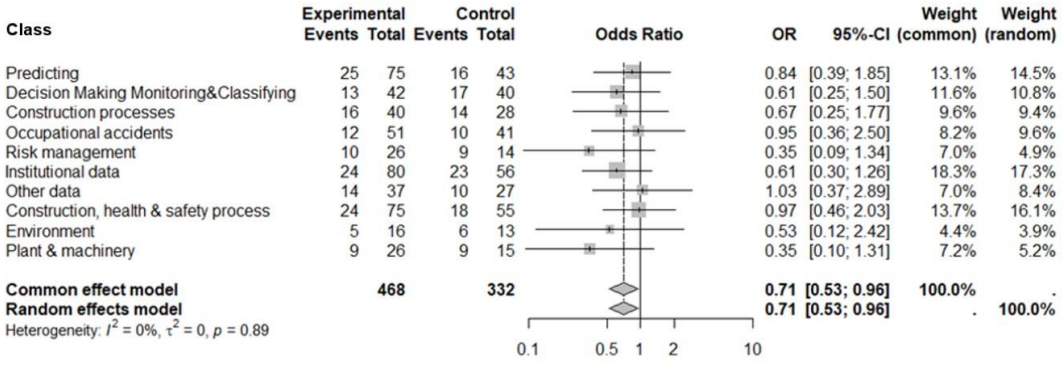

Figure 7. Funnel plot. Odds ratio (OR) and relative confidence interval (95% CI) for the total number of data mining analysed, by classes. The relative weight of each estimate in the analysis is marked with a box. The diamond represents the meta-analytical OR.

The heterogeneity is null (the sets under study are compatible). The analysis of the groups shows that the CIs intercept the "no effect" line and lose significance when taken individually, however, they are consistently overlapping and like each other. The Figure 7 indicate a general positive trend towards data treatment with supervised methods (on the left from "no-effect" line) which can be summarised by the OR = 0.71 and a CI (0.53-0.96).

4. Discussion and Future Directions

Studies focusing on DM techniques applied to the construction industry are relatively recent, dating back to 2014 at the latest and therefore, the review date from 2014 to September 2023. The number of articles on this sector has increased from 1 in 2014 to 23 in 2023. Similarly, the evolution

 doi:10.20944/preprints202405.0322.v1

of the total number of applied DM techniques increased from 5 between 2014 and 2016 to approximately 60 in 2023 (data not yet completed at the time of the survey).

In construction process field, 20 out of 61 observations were made regarding the construction of buildings, dams, roads, and tunnels. Within this field, the 60 out of 202 observations covered topics such as construction delays [22], crane, drilling and excavation tasks [14,18,24,39,48,70,74]; geological conditions [54], scaffolding collapse [68]; transport delays [31]; tunnelling [28,36,37,41,43,55,67]; workers and machinery location [34,40]. According to Erzaij et al. [22], project suspensions are among the most persistent tasks facing the construction sector, due to the difficulty of the industry and the essential interdependence between the bases of delay risk. The influence of delays can lead to increased time, costs, disputes, litigation, and overall rejection. The study aims to develop a data prediction tool to examine and learn the sources of delay based on previous data from construction projects, using decision trees and Bayesian naïve classification algorithms. Kumari et al. [34], investigated a machine learning architecture for excavator position detection by Global Positioning System (GPS), which can guarantee an excavator and driver position remarkably close to the real one. Wang J. et al. [67], used the principal component analysis (PCA) approach to select input factors for the prediction of tunnel boring machine (TBM) performance, particularly the travel speed. Liu et al [42], developed a model capable of predicting tunnel boring machine disc replacements based on a binary classification algorithm of Gaussian kernel support vector type cutting performance. After being trained over a period of historical data, the proposed model can predict whether cutter disc replacement is necessary, thus reducing the time required for periodic inspections. Lin et al. [40], investigated the feasibility of a real-time location service system using the Wi-Fi fingerprinting algorithm for safety risk assessment of tunnel workers. A location algorithm based on signal strength (RSS) and an artificial neural network (ann) was used for location analysis and risk assessment. Wei [68], developed wind speed prediction models based on various deep learning and machine learning techniques, in particular deep neural networks, neural networks with short-term memory, support vector regressions, random forest, and k-nearest neighbours. Subsequently, the author analysed the wind force on the scaffold and assessed the probability of the scaffold collapsing under the action of the wind.

In Occupational accident field, 16 out of 61 papers dealt with data on accidents and injuries at work from 2014 to 2023. In the class "occupational accidents," the 89 out of 202 submissions covered the following topics: reporting of accidents [3,26,27,29,32,35,43,45,49,59–62,75,76] and days away from work. On this topic, Yelda et al. analysed textual narratives to predict injury outcome and days off work in a mining operation. For this purpose, they used decision trees, random forest and ANN and the performance of these models was compared with that of logistic regression [71]. Lee et al. [35] proposed an optimised data preprocessing method to minimise variables and main elements in diverse and complex work accident data and built an ML prediction model to achieve this. Specifically, they analysed the correlations using a flood flow diagram and applied clustering and principal component analysis (PCA) to analyse the relationships between the main variables and to be able to draw broader conclusions. However, accidents are unevenly recorded in narrative form. To date, unfortunately, there is little research on text mining, natural language processing (NLP) and deep learning (DL) techniques for analysing construction accident narratives [7]. Construction accident reports hold a wealth of empirical knowledge that could be used to better understand, predict, and prevent the occurrence of accidents in the construction sector. Large construction companies and federal agencies, such as the Occupational Safety and Health Administration (OSHA), hold these reports in the form of huge digital databases [76]. Zhang J. et al. [76], utilised accident narrative data obtained from the official OSHA website, presenting a new unified architecture having a bi-directional short-term memory model (BiLSTM) and a convolutional layer for classification of construction accident causes. Tixier et al. [60], and Zhan F. et al. [75] proved how the study of safety attributes and outcomes can be automatically and accurately processed from unstructured accident reports using natural language processing (NLP).

In risk management field 25 out of 61 papers dealt with data on the "risk management process". Into this class the 54 observations out of 202 concerned the following topics: awkward working

 doi:10.20944/preprints202405.0322.v1

postures [7,19]; compliance with Health and Safety standards [2,4,47]; risk assessment [21,51,53,63,66,77,79]; safe climate [44]; slope instability [10,17]; teaching-training tasks [5,6,78]; unsafe behaviours [25,72]; worker fatigue-heat stress [38,69,73]; site image [1,46]. Antwi-Afari et al. [7] used deep learning networks to automatically extract relevant features with spatial-temporal dependence acquired by a wearable insole pressure system. The aim was to use deep learning-based networks and sensor data from wearable insoles to automatically recognise and classify types of awkward working postures for construction workers. So, they adopted recurrent neural networks (RNNs), deep learning models to train time series of plantar pressure data acquired from a wearable insole pressure sensor. Wang F. et al. [66] provided a strategic view of the relationships between different organisational objectives and technical risks that may arise during the construction of a tunnel. They created a systems-based model integrating Systems Dynamics (SD), Bayesian Belief Networks (BBN) and Smooth Relevance Vector Machines (sRVMs) called Organisational Risk Dynamics Observer (ORDO). The model was applied to an urban metro project built in Wuhan, China, and was used to provide guidance on effective accident prevention strategies. Mostofi et al. [45], explored the predictive ability of a multilayer GCN algorithm that learns the connection between construction accidents and project types, believing that richer information from existing safety and construction accident datasets by project type would provide better learning for the predictive model adopted. In addition, it would have supplied more information to predict the severity of accident consequences. The authors proved the effectiveness of the network representation of construction accidents in improving the learning capability of the ML model by using a feedforward reference network (FFN) algorithm with parameters like those used in the GCN algorithm to predict severity outcomes. The use of prefabricated is attracting increasing interest in the construction industry due to sustainability aspects, product quality, high production efficiency and cost-effectiveness. Dealing with this topic, Zhu & Liu [79] developed a prediction and risk assessment model related to the supply chain management of precast buildings. The BP neural network can be used to predict the risk of the prefabrication supply chain.

Soil instability and landslides are a major problem in the construction sector that can lead to safety risks for workers and the public, but also to considerable economic damage due to work stoppages. In this regard, Bay et al. [10] evaluated 102 cases of slopes with arc-shaped failure modes using eight machine learning regression methods. The slope safety factor prediction models were set up by performing cross-validation and hyper-parameter adjustment of the model. Furthermore, based on objective weighting and TOPSIS methods, was developed a model to evaluate the performance of the machine learning model and find the best FOS prediction model. Sadeghi et al. [53] developed an Ensemble Predictive Safety Risk Assessment Model (EPSRAM) to assess the health and safety risks of workers on construction sites based on the integration of neural networks and fuzzy inference systems. The model introduces an innovation in countries such as Malaysia, where there is continued growth in the construction industry but where there is a lack of studies on OHS assessment of workers involved in construction activities. Such circumstances may expose construction workers to the risk of developing fatigue. If workers continue to work under fatigued conditions, they are prone to the development of work-related musculoskeletal disorders (MSDs). Yu et al. [73] and Yan et al. [69], developed a combination of computer vision technology and biomechanical analysis for non-intrusive whole-body fatigue monitoring of construction workers using 3D model data from the motion capture algorithm and biomechanical analysis.

Zhao et al. [78], conducted a study on efficient and parallel DM and machine learning methods and algorithms distributed on a large scale and proposed an experiential teaching model focused on the cultivation of independent learning ability and the subjective initiative of individual learners. The article, which could have been excluded for review, was nevertheless kept as it combined the importance and technical challenges of the algorithms themselves and the context of the practical application needs of the field. It reported research on methods and algorithms for DM and machine learning, distributed on a large scale for training purposes. As an innovative teaching model, the experiential teaching model described in it, focuses among other things on cultivating individual learners' independent learning ability and subjective initiative, which was found to effectively

 doi:10.20944/preprints202405.0322.v1

activate the atmosphere of the working class/environment and improve the teaching effect. It has been included as one of the articles dealing in an innovative way with risk management process, including health and safety training in the workplace. Other studies, not included in the review, report an analysis based on the effectiveness of combinations of Smart Construction Safety Technologies (SCST), potentially able to generate information useful for DM and the measurement of the effectiveness of the same technologies, single and combined [33].

Regarding the type of data used in the DM, 39 out of 61 papers dealt with institutional datasets (2016-2023), 8 used signal and video data (2014-2023), 4 narrative texts (2016-2022), 3 construction projects (2016-2023), 3 literature data (2020-2023), 3 simulations (2015-2023) and 1 out of 61 used an interview report (2023). The data used may have different characterisations, as refer to specific aspects of an occupational injury such as, for example, the body parts affected and the expected probability. Other studies focus on the observation of environmental and meteorological precursors of accidents, e.g. associated with the collapse of scaffolding [41] and slope instability [10,17]. Liu et al. analysed data from sophisticated and technologically innovative machine monitoring, capable of returning and processing geological data, faults and predicting the progress of TBM and maintenance, avoiding downtime and inspections, were analysed [43]. According to Schindler et al. [55] and Leng et al. [36], the use of satellite data has proved to be a winning strategy compared to ground surveys. Data collected by sensors was used to assess the state of effort associated with the awkward working postures taken by the worker while performing his work on the construction site [7] or the data of physical fatigue and worker's heat stress [69]. Another interesting use of data involves the construction practitioners' interview through which it was integrated processes and occupational risk information [51]. By focusing on health and safety aspects, quality, in terms of the homogeneity and standardisation of the various sources of institutional accident data included in the review, it can be affected by the different methods of acquisition, from one institution to another and from one country to another. It can also assume that the data produced by technologies and machines used in the processes, have a higher degree of homogeneity and standardization than the former. Liu et al. underline the significance of employing innovative and efficient safety management technologies, along with new management approaches and automated methods based on artificial intelligence, to detect and eliminate risks promptly. According to the authors, these innovative technologies would mitigate any deficiencies in site management, significantly improve site safety management and eliminate risks at source [43]. An increasingly widespread orientation towards an automated management of the site or parts of it, would not only lead to an improvement in the health and safety of the processes but also a significant improvement in the quality of the data coming from the construction field. It can be assumed that soon, accident data collection techniques will not be able to do without innovative technologies capable of automatically acquiring information on near misses, accidents, and injuries in the construction sector.

Intelligent technologies can generate a range of data that pertains to both the individual (e.g. worker) and the interaction and connection between different technologies. The Internet of Things (IoT) is gradually spreading in the construction sector, thus making an important contribution to the production of new data. Robots and collaborative robots play a significant role in technological innovation and data extraction, as they can produce quality in terms of productivity, product quality, and standardization of production processes. Furthermore, these technologies have the potential to produce high-quality data, which could play a significant role in the pre-processing of data required for the use of DM techniques. The use of these technologies in construction sites is still limited due to unresolved difficulties, attributable to the high variability of environmental conditions, the need to protect the secrecy of processes and the privacy of workers. Moreover, to accompany change, workers and enterprises need vocational training and management training [30].

5. Conclusions

Cluster analysis and PCA was applied to data from articles that met the PRISMA eligibility criteria and were included in the review. The study indicates an association between types of methods used and objectives, scope, type of data and resources under investigation. This association,

 doi:10.20944/preprints202405.0322.v1

based on correlation, was synthesised onto a single xy-plane (Dim1 and Dim2). The results of the PCA were consistent with those of the cluster analysis. Each of the two axes was assigned a label summarising the significance of the entire review. The x-axis (Dim1) was labelled “Supervised methods (dt, knn, svm and rf) applied to institutional data for classification and inference”. The y-axis (Dim2) was labelled "Not-supervised methods (lstm, word2vec, nlp, BIM) applied to projects, simulations, signals, interviews, literature or textual data to classify and make decisions". The meta-analysis with odds ratio (OR) of 0.71 and a confidence interval (CI) of 0.53 to 0.96 provides an overall estimate of the superior effectiveness of supervised methods compared to not-supervised ones.

Author Contributions: conceptualization, AP; method, AP, ML, AB; validation, AP, AB, ML, AC, DB; writing, AP, review and editing, AP, AB, AC, ML, BD. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Acronyms and dm method included in the review. Source: Author’s processing by Scopus and ProQuest archives, R.

dm method	dm method description	dm method	dm method description
3d motion	3d motion	it2f–ahp	interval type-2 (IT2) fuzzy-analytic hierarchy process
abc	approximate bayesian computation	it2fd	interval type-2 (IT2) fuzzy Delphi
adaboost	adaptive boosting (ensemble)	k-means	k-means clustering
afdd	automated fault detection and diagnostics	knn	k-nearest neighbour
ahp	analytic hierarchy process	ksvm	support vector machines in kernlab
al	ml-based active learning framework	lcca	ml based life-cycle cost analysis
anfis	adaptive neuro-fuzzy inference system	lda	latent dirichlet allocation
ann	artificial neural network	lin r	linear regression
ar	augmented reality	log r	logistic regression
autokeras	automl system based on keras	lstm	long short-term memory
auto-sklearn	automatic scikit-learn	mlaeld	machine learning architecture for excavators' location detection
bagging	bootstrap aggregating	mlp	multilayer perceptron
bbn	bayesian belief networks	monte carlo	montecarlo method
bert	bidirectional encoder represent. for transformers	mcda-c	multicriteria methodology for decision aiding-constructivist
bi-bert	binarized bidirectional encoder represent. for transformers	mosma	multi-objective slime mould algorithm
bi-lstm	bi-directional long short-term memory	nb	Naïve Bayes
bim	building information modeling	nbc	naive bayes classifier
bnn	binarized neural network	nlp	natural language processing
bns	bayesian networks	nltk	natural language toolkit

 doi:10.20944/preprints202405.0322.v1

bpnn	back propagation in neural network	onehotencoding	onehotencoding in scikit-learn
caml	customized automl	pca	principal components analysis
catboost	gradient boosting on decision trees	pca-ahp	analytic hierarchy process-principal component analysis)
c-bilstm	convolutional bi-directional long short-term memory	pls-sem	partial-least-squares structural-equation modeling
cbow	continuous bag of words	rf	random forest
chi-square	chi-square	rl	reinforcement learning
clustering	clustering	rnn	recurrent neural network
cnn	convolutional neural network	ros	robot operating system
cpbt	cognitive psychology and bloom's taxonomy	sae	sparse autoencorder
cramer's v	cramer's v	satellite-based meas.	satellite-based measurements
cv	computer vision process	scibert	scientific bidirectional encoder represent. for transformers
deepar	autoregressive recurrent networks	scikit-learn	key library for pyton programming language
dl	deep learning	sd	system dynamics
dnn	deep neural network	smote	synthetic minority over-sampling technique
dt	decision tree learning	sqp	sequential quadratic programming
ensemble	ensemble	srvm	smooth relevance vector machines
epsram	ensemble predictive safety risk assessment model	svm	support vector machines
faxtext	faxtext	svr	support vector regression
ffn	feed-forward neural network	swpl	smart work package learning
flac3d	flac3d	tf-idf	term frequency-inverse document frequency
fuz	fuzzy approaches	tokenitation	split sentences into small units
gbdt	gradient boosted decision trees	topsis	technique for order of preference by similarity to ideal solution
gcn	graph convolutional networks	vr	virtual reality
glove	global vectors for words representation	wbs-rbs	work breakdown structure-resource breakdown structure
gru	gated recurrent unit (recurrent neural network)	word2vec	word2vec (nlp)
hcpc	hierarchical clustering on principal components	yolo-v5	you only look once
ica	independent component analysis		

 doi:10.20944/preprints202405.0322.v1

Appendix B

Table A2. x1-x4 Obj. (classifying, decision making, monitoring, predicting), x5-x7 field (construction process, occupational accident, H&S management.), x8-x14 data (project, institutional data, interview, literature, text, signal & video, simulation), x15-x17 DM (supervised, unsupervised, other), x18-x20 source (construction, H&S resource, environment, machinery). Source: Author's processing by Scopus and ProQuest archives and R.

Reference	x1	x2	x3	x4	x5	x6	x7	x8	x9	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x2
Alateeq et al. 2023	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
Alhelo et al., 2023	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0
Alkaissy et al., 2023	0	0	0	1	0	1	0	0	1	0	0	0	0	0	3	2	1	1	0	0
Al-Kasasbeh et al. 2022	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0
Aminu Darda'u et al., 2023	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0
Ankit et al., 2023	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	2	2	1	0	0
Antwi-Afari et al., 2022	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	2	1	0	0
Bai et al., 2022	0	0	0	1	1	0	0	0	1	0	0	0	0	0	8	1	0	0	1	0
Choo et al., 2023	0	0	1	0	1	0	0	0	0	0	0	0	1	0	4	0	0	0	0	1
Dong et al., 2021	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0
Duan et al., 2023	1	0	0	0	0	0	1	0	0	0	0	0	1	0	2	0	0	0	0	1
Dutta et al., 2023	1	0	0	0	1	0	0	0	0	0	0	0	1	0	3	0	0	1	0	0
Ensslin et al., 2022	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0
Erzaij et al., 2021	0	0	0	1	1	0	0	0	1	0	0	0	0	0	2	0	0	1	0	0
Fernández et al., 2023	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2	0	0	1
Gao et al. 2022	0	0	0	1	0	0	1	0	1	0	0	0	0	0	4	0	0	1	0	0
Goh et al., 2017	1	0	0	0	0	1	0	0	1	0	0	0	0	0	6	1	1	1	0	0
Goldberg, 2022	0	0	0	1	0	1	0	0	0	0	0	1	0	0	4	0	1	1	0	0
Hasanpour et al., 2015	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1
Hoła et Szóstak, 2019	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
Jha et al., 2023	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0
Khairuddin et al., 2022	0	0	0	1	0	1	0	0	1	0	0	0	0	0	6	0	0	1	0	0
Kumari et al., 2022	0	0	0	1	1	0	0	0	0	0	0	0	1	0	4	0	1	0	0	1
Lee et al., 2020	0	0	0	1	0	1	0	0	1	0	0	0	0	0	3	5	0	1	0	0
Leng et al., 2021	0	0	0	1	1	0	0	0	1	0	0	0	0	0	2	0	1	0	0	1
Li et al., 2017	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
Li et al., 2023	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	1	0	0
Lim et al. 2022	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1
Lin et al., 2014	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1
Lin et al., 2023	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
Liu et al., 2022	0	0	0	1	1	0	0	0	1	0	0	0	0	0	4	1	1	0	0	1
Liu et al., 2023	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	2	0	1	0	0
Maqsoom et al. 2023	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
Mostofi et al., 2022	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	1	1	0	0
Muhammad et al., 2023	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	0
Numan et al., 2023	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0
Osa et al., 2023	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
Passmore et al., 2019	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0
Razi et al., 2023	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
Sadeghi et al., 2020	0	1	0	0	0	0	1	0	0	0	1	0	0	0	3	0	0	1	0	0
Sapronova et al., 2023	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0
Schindler et al., 2016	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
Shuang et al., 2023	0	0	0	1	0	1	0	0	1	0	0	0	0	0	7	0	0	1	0	0
Tixier et al., 2016	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0

 doi:10.20944/preprints202405.0322.v1

Tixier et al., 2017	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	4	1	1	0	0
Toğan et al. 2022	0	0	1	0	0	0	1	0	1	0	0	0	0	0	6	0	2	1	0	0
Topal & Atasoylu, 2022	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0
Wang F. et al., 2016	0	0	0	1	1	0	0	1	0	0	0	0	0	0	2	1	0	0	0	1
Wang et al., 2022	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	3	0	0	0	1
Wei, 2021	0	0	0	1	1	0	0	0	1	0	0	0	0	0	6	6	2	0	1	0
Yan et al., 2022	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0
Yao et al., 2022	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Yedla et al., 2020	0	0	0	1	0	1	0	0	1	0	0	0	0	0	4	4	0	1	0	0
Yin et al., 2023	1	0	0	0	0	0	1	0	1	0	0	0	0	0	4	1	1	1	0	0
Yu et al., 2019	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0
Zermane et al. 2023	0	0	0	1	0	1	0	0	1	0	0	0	0	0	7	0	0	0	0	1
Zhang X. et al., 2020	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
Zhang F. et al., 2019	1	0	0	0	0	1	0	0	0	0	0	1	0	0	6	4	0	1	0	0
Zhang J. et al., 2020	1	0	0	0	0	1	0	0	0	0	0	1	0	0	3	2	5	1	0	0
Zhao et al., 2022	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0
Zhu&Liu, 2023	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0

References

1. , Muneerah M., Fathimathul Rajeena P.P., Ali, Mona, A.S. Construction Site Hazards Identification Using Deep Learning and Computer Vision. Sustainability. Basel. Vol. 15, Fasc. 3, (2023): 2358. DOI:10.3390/su15032358.
2. Alhelo, A.A., Radhi A., Hamad R. A. Framework Supporting Health and Safety Practices in the United Arab Emirates' Construction Projects. Sustainability. Basel. Vol. 15, Fasc. 2, (2023): 1587. DOI:10.3390/su15021587.
3. Alkaissy, M., Arashpour, M., Golafshani, E.M., Hosseini, M.R., Khanmohammadi, S., Bai, Y., Feng, H. (2023). Enhancing construction safety: Machine learning-based classification of injury types. Safety Science. Vol. 162. ISSN 106102. 9257535. DOI 10.1016/j.ssci.2023.106102.
4. Al-Kasasbeh, M, Mujalli, R., O., Abudayyeh, O., Liu, H., Altalhoni, A. Bayesian Network Models for Evaluating the Impact of Safety Measures Compliance on Reducing Accidents in the Construction Industry. Buildings. Basel Vol. 12, Fasc. 11, (2022): 1980. DOI:10.3390/buildings12111980.
5. Aminu D. R., Nasir S., Othman, I., Miljan M. Mechanism Models of the Conventional and Advanced Methods of Construction Safety Training. Is the Traditional Method of Safety Training Sufficient? International Journal of Environmental Research and Public Health. Basel Vol. 20, Fasc. 2, (2023): 1466. DOI:10.3390/ijerph20021466.
6. Ankit, S., Arashpour, M., Emadaldin M. G., Dwyer, T. Kalutara, P. Enhancing Safety Training Performance Using Extended Reality: A Hybrid Delphi–AHP Multi-Attribute Analysis in a Type-2 Fuzzy Environment. Buildings. Basel Vol. 13, Fasc. 3, (2023): 625. DOI:10.3390/buildings13030625.
7. Antwi-Afari, M.F., Qarout, Y., Herzallah, R., Anwer, S., Umer, W., Zhang, Y., Manu, P. Deep learning-based networks for automated recognition and classification of awkward working postures in construction using wearable insole sensor data. Automation in Construction. Vol. 136. (2022). ISSN 9265805. DOI 10.1016/j.autcon.2022.104181.
8. Arockiam, A., J., M., S., E., S., Irudhayaraj. Reclust: an efficient clustering algorithm for mixed data based on reclustering and cluster validation. Indonesian Journal of Electrical Engineering and Computer Science. Vol. 29, No. 1, January 2023, pp. 545-552. (2023). ISSN: 2502-4752, DOI: 10.11591/ijeecs. v29.i1.pp545-552.
9. Baas, J, Schotten, M, Plume, A, Côté, G. Karimi, R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quant Sci Stud*, *1*, 377–386, (2020). [CrossRef].
10. Bai, G., Hou, Y., Wan, B., An, N., Yan, Y., Tang, Z., Yan, M., Zhang, Y., Sun, D. Performance Evaluation and Engineering Verification of Machine Learning Based Prediction Models for Slope Stability. Applied Sciences (Switzerland). Vol. 12, Is. 15. (2022). ISSN 20763417. DOI10.3390/app12157890.
11. Bolasco, S. Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione. Roma, Italia: Carocci. ISBN: 8843014013. (1999). EAN: 9788843014019.

12. Bożena, H, Mariusz, S. Modeling of the Accidentality Phenomenon in the Construction Industry. Applied Sciences. Basel Vol. 9, Fasc. 9, (Jan 2019). DOI:10.3390/app9091878.
13. Chiang, M.MT., Mirkin, B. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. J Classif 27, 3–40 (2010). https://doi.org/10.1007/s00357-010-9049-5.
14. Choo, H., Lee, B., Kim, H., Choi, B. Automated detection of construction work at heights and deployment of safety hooks using IMU with a barometer. Automation in Construction. Volume 147, March 2023, Article number 104714. DOI: 10.1016/j.autcon.2022.104714.
15. Clarivate ProQuest Document Search Available online. https://www.proquest.com (Accessed: 16-18 November 2022).
16. Di Franco, G. Tecniche e modelli di analisi multivariata. Milano, Italia: Franco Angeli Editore. (2017). ISBN-13: 978-8891761064.
17. Dong, M., Wu, H., Hu, H., Azzam, R., Zhang, L., Zheng, Z., Gong, X. Deformation prediction of unstable slopes based on real-time monitoring and deepar model. Sensors (Switzerland). Vol. 21. IS. 1, 1-18. (2021). ISSN 14248220. DOI 10.3390/s21010014.
18. Duan, P., Zhou, J., Tao, S. Risk events recognition using smartphone and machine learning in construction workers' material handling tasks. Engineering, Construction and Architectural Management. Volume 30, Issue 8, 1 September 2023, Pages 3562-3582. DOI: 10.1108/ECAM-10-2021-0937.
19. Dutta, A., Breloff, S.P., Mahmud, D., Dai, F., Sinsel, E.W., Warren, C.M., Wu, J.Z. Automated Classification of the Phases Relevant to Work-Related Musculoskeletal Injury Risks in Residential Roof Shingle Installation Operations Using Machine Learning. Buildings. Vol. 13, Issue 6, June 2023, Article number 1552. DOI: 10.3390/buildings13061552.
20. Elsevier Scopus Document Search Available online. https://www scopus com (Accessed: 16-18 November 2022).
21. Ensslin, L. et al. Constructivist multi-criteria model to support the management of occupational accident risks in civil construction industry. PLoS One. San Francisco Vol. 17, Fasc. 6, (Jun 2022): e0270529. DOI:10.1371/journal.pone.0270529.
22. Erzaij, K.R., Burhan, A.M., Hatem, W.A., Ali, R.H. Prediction of the Delay in the Portfolio Construction Using Naive Bayesian Classification Algorithms. Civil and Environmental Engineering. Vol. 17, Is. 2, 673-680. (2021). ISSN 13365835. DOI 10.2478/cee-2021-0066.
23. Falagas, M E., Pitsouni, E. I., Malietzis, G A., Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses *FASEB J*, *22*, 338–342, (2007). [CrossRef]. https://doi.org/10.1096/fj.07-9492LSF.
24. Fernández, A., Rivera, F.M.L., Mora-Serrano, J. Virtual Reality Training for Occupational Risk Prevention: Application Case in Geotechnical Drilling Works. International Journal of Computational Methods and Experimental Measurements. Volume 11, Issue 1, March 2023, Pages 55-63. DOI: 10.18280/ijcmem.110107.
25. Gao, Y., González, V. A, Yiu, T. W., Cabrera-Guerrero, G., Deng, R. Predicting Construction Workers' Intentions to Engage in Unsafe Behaviours Using Machine Learning Algorithms and Taxonomy of Personality. Buildings. Basel Vol. 12, Fasc. 6. (2022): 841. DOI:10.3390/buildings12060841.
26. Goh, Y.M., Ubeynarayana, C.U. Construction accident narrative classification: An evaluation of text mining techniques. Accident Analysis and Prevention. Vol.108, 122-130. (2017). ISSN 14575. DOI 10.1016/j.aap.2017.08.026.
27. Goldberg, D.M. Characterizing accident narratives with word embeddings: Improving accuracy richness and generalizability. Journal of Safety Research. Vol. 80, 441-455.ISSN 224375. (2022). DOI 10.1016/j.jsr.2021.12.024.
28. Hasanpour, R., Rostami, J., Barla, G. Impact of Advance Rate on Entrapment Risk of a Double-Shielded TBM in Squeezing Ground. Rock Mechanics and Rock Engineering. Vol. 48, Is. 3, 1115-1130. (2015). ISN 7232632. DOI 10.1007/s00603-014-0645-2.
29. Hoła, B., Szóstak, M. Modeling of the Accidentality Phenomenon in the Construction Industry. Applied Sciences. (2019). DOI: 10.3390/app9091878.
30. ILOSTAT. International Labour Organization. Statistics on safety and health at work. https://ilostat.ilo.org/topics/safety-and-health-at-work/ (Accessed: 18 July 2023).
31. Jha, M.K., Wanko, N., Bachu, A.K. A Machine Learning-Based Active Learning Framework to Capture Risk and Uncertainty in Transportation and Construction Scheduling. 2nd International Conference on

Transportation Infrastructure Projects: Conception to Execution, TIPCE 2022. Haridwar, India. 14 September 2022 through 17 September 2022. Code 297359. DOI: 10.1007/978-981-99-2556-813.

32. Khairuddin, M. Z. F., Puat, L. H., Hasikin, K., Nasrul, A. A. R., Khin, W. L., et al. Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. International Journal of Environmental Research and Public Health. Basel Vol. 19, Fasc. 21, (2022): 13962. DOI:10.3390/ijerph192113962.
33. Kim, Y.-S., Lee, J.Y., Yoon, Y.-G., Oh, T.-K. Effectiveness Analysis for Smart Construction Safety Technology (SCST) by Test Bed Operation on Small- and Medium-Sized Construction Sites. Int. J. Environ. Res. Public Health 2022, 19, 5203. https://doi.org/10.3390/ijerph19095203.
34. Kumari, S., Siwach, V., Singh, Y., Barak, D., Jain, R. A Machine Learning Centered Approach for Uncovering Excavators' Last Known Location Using Bluetooth and Underground WSN. Wireless Communications and Mobile Computing. (2022). DOI 10.1155/2022/9160031.
35. Lee, J.Y., Yoon, Y.G., Oh, T.K., Park, S., Ryu, S.I. A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry. Applied Sciences (Switzerland). Vol. 10, Is. (2020). DOI 10.3390/app10217949.
36. Leng, S., Lin, J.-R., Hu, Z.-Z., Shen, X. A Hybrid Data Mining Method for Tunnel Engineering Based on Real-Time Monitoring Data from Tunnel Boring Machines. IEEE Access. Vol. 8. 30-49. (2020). ISSN 21693536. DOI10.1109/ACCESS.2020.2994115.
37. Li, L., Tao, J.-F., Yu, H.-D., Huang, Y.-X., Liu, C.-L. Online Condition Monitoring of Gripper Cylinder in TBM Based on EMD Method. Chinese Journal of Mechanical Engineering (English Edition). Vol. 30, Is. 6, 1325-1337. (2017). ISSN 10009345. DOI 10.1007/s10033-017-0187-0.
38. Li, X., Zeng, J., Chen, C., Chi, H.-L., Shen, G.Q. Smart work package learning for decentralized fatigue monitoring through facial images. Computer-Aided Civil and Infrastructure Engineering. Volume 38, Issue 6, April 2023, Pages 799-817. DOI: 10.1111/mice.12891.
39. Lim, J., Jung, D, G, Park, C, Kim, D, Y. Computer Vision Process Development regarding Worker's Safety Harness and Hook to Prevent Fall Accidents: Focused on System Scaffolds in South Korea Advances in Civil Engineering. New York Vol. 2022, (2022). DOI:10.1155/2022/4678479.
40. Lin, P., Li, Q., Fan, Q., Gao, X., Hu, S. A Real-Time Location-Based Services System Using WiFi Fingerprinting Algorithm for Safety Risk Assessment of Workers in Tunnels. IEEE Access. (2014). http://dx.doi.org/10.1155/2014/371456.
41. Lin, P., Wu, M., Zhang, L. (2023). Probabilistic safety risk assessment in large-diameter tunnel construction using an interactive and explainable tree-based pipeline optimization method. Applied Soft Computing. Volume 143, August 2023, Article number 110376. DOI: 10.1016/j.asoc.2023.110376.
42. Liu, Y., Huang, S., Wang, D., Zhu, G., Zhang, D. Prediction Model of Tunnel Boring Machine Disc Cutter Replacement Using Kernel Support Vector Machine. Applied Sciences (Switzerland). Vol.12. Is. 5. ISSN 20763417. (2022). DOI 10.3390/app12052267.
43. Liu, Y., Wang, J., Tang, S., Zhang, J., Wan, J. Integrating Information Entropy and Latent Dirichlet Allocation Models for Analysis of Safety Accidents in the Construction Industry. Buildings. Basel Vol. 13, Fasc. 7, (2023): 1831. DOI:10.3390/buildings13071831.
44. Maqsoom, A., Hassan, A., Wesam S. A., Salman, A., Ullah, F. et al. (2023). The Relationship between Error Management, Safety Climate, and Job-Stress Perception in the Construction Industry: The Mediating Role of Psychological Capital. Buildings. Basel Vol. 13, Fasc. 6, (2023): 1528. DOI:10.3390/buildings13061528.
45. Mostofi, F., Togan V., Ayözen Y.E., Tokdemir, O.B. Construction Safety Risk Model with Construction Accident Network: A Graph Convolutional Network Approach. Sustainability (Switzerland). Vol. 14-23, (2022). ISSN 20711050. DOI 10.3390/su142315906.
46. Muhammad Ali Musarat, Wesam Salah Alaloul, Muhammad Irfan, Sreenivasan, Pravin, Muhammad Babar Ali Rabbani. Health and Safety Improvement through Industrial Revolution 4.0: Malaysian Construction Industry Case. Sustainability. Basel Vol. 15, Fasc. 1, (2023): 201. DOI:10.3390/su15010201.
47. Numan, Khan, Syed F. A. Z., Yang, J., Park, C., Lee, D. Construction Work-Stage-Based Rule Compliance Monitoring Framework Using Computer Vision (CV) Technology. Buildings. Basel. Vol. 13, Fasc. 8, (2023): 2093. DOI:10.3390/buildings13082093.
48. Osa, T., Osajima, N., Aizawa, M., Harada, T. Document details - Learning Adaptive Policies for Autonomous Excavation Under Various Soil Conditions by Adversarial Domain Sampling. IEEE Robotics and Automation Letters. 2023, Pages 1-8. DOI: 10.1109/LRA.2023.3296933.

49. Passmore, D., Chae, C., Borkovskaya, V., Baker, R., Yim, J.-H. Severity of U.S. Construction Worker Injuries 2015-2017. E3S Web of Conferences. Vol. 97. (2019). ISSN 25550403. DOI 10.1051/e3sconf/20199706038.
50. Prisma Flow Diagram. Available online: https://www prisma-statement org//PRISMAStatement/FlowDiagram (Accessed: 21 March 2023).
51. Razi, P.Z., Sulaiman, S.K., Ali, M.I., Ramli, N.I., Saad, M.S.H., Jamaludin, O., Doh, S.I. "How Artificial Intelligence Changed the Construction Industry in Safety Issues. IOP Conference Series: Earth and Environmental Science. ISSN 17551307. (2023). DOI 10.1088/1755-1315/1140/1/012004.
52. 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. Data mining in occupational safety and health: a systematic mapping and roadmap. Production, 31, e20210048. (2021). https://doi.org/10.1590/0103-6513.20210048.
53. Sadeghi, H., Mohandes, S.R., Hosseini, M.R., Banihashemi, S., Mahdiyar, A., Abdullah, A. Developing an ensemble predictive safety risk assessment model: Case of Malaysian construction projects. International Journal of Environmental Research and Public Health. Vol. 17. ISSN 16617827. (2020). DOI 10.3390/ijerph17228395.
54. Sapronova, A., Unterlass, P.J., Dickmann, T., Hecht-Méndez, J., Marcher, T. Prediction of Geological Conditions Ahead of the Tunnel Face: Comparing the Accuracy of Machine Learning Models Trained on Real and Synthetic Data. 3rd International Conference of International Society for Intelligent Construction, ISIC 2022. Guimarães, Portugal. 6 September 2022 through 9 September 2022. Code 287369. Lecture Notes in Civil Engineering. Vol. 306 LNCE, 2023, Pages 76-86. DOI: 10.1007/978-3-031-20241-4_6.
55. Schindler, S., Hegemann, F., Koch, C., Konig, M., Mark, P. Radar interferometry-based settlement monitoring in tunnelling: Visualisation and accuracy analyses. Visualization in Engineering. Vol. 4, Is.1. (2016). ISSN 22137459. DOI 10.1186/s40327-016-0034-x. 2016.
56. Scimago Compare Journal Available online: https://www scimagojr com/comparejournals php?ids[]=29593 (Accessed: 18 November 2022).
57. Scimago Journal Rank Available online: https://www scimagojr com/journalrank php?category=2603 (Accessed: 18 November 2022).
58. Scimago Viztools Available online: https://www scimagojr com/viztools php (Accessed: 18 November 2022)
59. Shuang, Q., Zhang, Z. Determining Critical Cause Combination of Fatality Accidents on Construction Sites with Machine Learning Techniques. Buildings. Vol. 13. ISSN 20755309. (2023). DOI 10.3390/buildings13020345.
60. Tixier, A.J.-P., Hallowell, M.R., Rajagopalan, B., Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction. Vol. 62, 45-56. (2016). ISSN 9265805. DOI 10.1016/j.autcon.2015.11.001.
61. 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. Vol. 74, Is. 39. ISSN 9265805. (2017). DOI 10.1016/j.autcon.2016.11.001.
62. Toğan, V., Mostofi, F., Yunus E. A., Tokdemir, O. B. Customized AutoML: An Automated Machine Learning System for Predicting Severity of Construction Accidents. Buildings. Basel. Vol. 12, Fasc. 11, (2022): 1933. DOI:10.3390/buildings12111933.
63. Topal, S., Atasoylu, E. A Fuzzy Risk Assessment Model for Small Scale Construction Work. Sustainability. Basel Vol. 14, Fasc. 8, (2022): 4442. DOI:10.3390/su14084442.
64. Van Eck, N J, Waltman, L. (2014). Visualizing bibliometric networks. In *Measuring Scholarly Impact: Methods and Practice*, Ding, Y, Rousseau, R, Wolfram, D, Eds, Springer: Cham, Switzerland, pp 285–320 [CrossRef].
65. Vosviewer. Visualizing scientific landscapes. https://www.vosviewer.com/ (Accessed: 21 March 2023).
66. Wang, F., Ding, L., Love, P.E.D., Edwards, D.J. Modeling tunnel construction risk dynamics: Addressing the production versus protection problem. Safety Science. Vol. 87, 101-115. ISSN 9257535. (2016). DOI 10.1016/j.ssci.2016.01.014.
67. Wang, J., Mohammed, A.S., Macioszek, E., Ali, M., Ulrikh, D.V., Fang, Q. A Novel Combination of PCA and Machine Learning Techniques to Select the Most Important Factors for Predicting Tunnel Construction Performance Buildings. Vol. 12, Is.7. (2022). DOI 10.3390/buildings12070919.

68. Wei, C.C. Collapse warning system using LSTM neural networks for construction disaster prevention in extreme wind weather. Journal of Civil Engineering and Management. Vol. 27, Is. 4, 230-245. (2021). ISSN 13923730. DOI 10.3846/jcem.2021.14649.
69. Yan, R., Yi, W., Wang, S. Predicting Maximum Work Duration for Construction Workers. Sustainability (Switzerland). Vol. 14, Is. 17. (2022). ISSN 20711050. DOI 10.3390/su141711096.
70. Yao, G, Sun, W T, Yang, Y. Analysis and Identification of Building Construction Accident Risk in China basing Exclusively Database IOP Conference Series. Earth and Environmental Science. Bristol. Vol. 1101, Fasc. 7, (Nov 2022): 072009. DOI:10.1088/1755-1315/1101/7/072009.
71. Yedla, A., Kakhki, F.D., Jannesari, A. Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. International Journal of Environmental Research and Public Health. Vol. 17, Is. 19, 1-17. (2020). ISSN 16617827. DOI: 10.3390/ijerph17197054.
72. Yin, S., Wu, Y., Shen, Y., Rowlinson, S. Development of a Classification Framework for Construction Personnel's Safety Behavior Based on Machine Learning. Buildings. Basel. Vol. 13, Fasc. 1, (2023): 43. DOI:10.3390/buildings13010043.
73. Yu, Y., Li, H., Yang, X., Kong, L., Luo, X., Wong, A.Y.L. An automatic and non-invasive physical fatigue assessment method for construction workers. Automation in Construction. Vol. 103, Is. 1. (2019). ISSN 9265805. DOI 10.1016/j.autcon.
74. Zermane, A., Mohd Tohir, M.Z., Zermane, H., Baharudin, M.R., Mohamed Yusoff, H. Predicting fatal fall from heights accidents using random forest classification machine learning model. Safety Science. Vol. 159, March 2023, Article number 106023. DOI: 10.1016/j.ssci.2022.106023.
75. Zhang, F., Fleyeh, H., Wang, X., Lu, M. Construction site accident analysis using text mining and natural language processing techniques. Automation in Construction. Vol. 99, 238-248. (2019). ISSN 9265805. DOI 10.1016/j.autcon.2018.12.016.
76. Zhang, J., Zi, L., Hou, Y., Deng, D., Jiang, W., Wang, M. A C-BiLSTM approach to classify construction accident reports. Applied Sciences (Switzerland). Vol. 10, Is. 17, 54-76. (2020). DOI10.3390/APP10175754.
77. Zhang, X., Huang, S., Yang, S., Tu, R., and Jin L. Safety Assessment in Road Construction Work System Based on Group AHP-PCA. Mathematical Problems in Engineering (2020). Vol. 2020. https://doi.org/10.1155/2020/6210569 (32).
78. Zhao, F., Zhang, G., Wang, Z., Hao, X. Construction of Higher Education Management Data Analysis Model Based on Association Rules. Scientific Programming. ISSN 5414238. (2022). DOI 10.1155/2022/5414238.
79. Zhu, T., Liu, G. A Novel Hybrid Methodology to Study the Risk Management of Prefabricated Building Supply Chains: An Outlook for Sustainability. Sustainability (Switzerland). Vol. 15. (2023). ISSN 20711050. DOI 10.3390/su15010361.