www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

A REVIEW OF MACHINE LEARNING FOR ANALYSING ACCIDENT REPORTS IN THE CONSTRUCTION INDUSTRY

SUBMITTED: February 2022
REVISED: March 2025
PUBLISHED: April 2025
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2025.019

May Shayboun, Ph.D. student,
Halmstad University, Entrepreneurship, Innovation and Sustainability
shayboun@hh.se

Dimosthenis Kifokeris, Associate Professor,
Chalmers University of Technology, Architecture and Civil Engineering
dimkif@chalmers.se

Christian Koch, Professor,
Halmstad University, Entrepreneurship, Innovation and Sustainability
christian.koch@hh.dk
University of Southern Denmark, Department of Technology and Innovation
ckoc@iti.sdu.dk

SUMMARY: *Recently, there has been a growth in the research interest on applied machine learning (ML) in safety analysis in the construction industry. The increased interest is part of a search for improved prevention of occupational accidents with a focus on text analysis and natural language processing (NLP). However, ML-based approaches have been less adapted compared to their perceived benefits due to barriers of implementation and challenges in analysing* safety records in the construction sector. *And the current literature has been criticized for a lack of clarity around the description of methodologies, interpretation, and the context of the application. Therefore, this work aims to review the latest developments in* research applying ML to accident report analysis in construction. *A review of the published literature on ML-based analysis of construction accident reports was carried out and organized in terms of the data pre-processing, algorithms, testing and implementation and further organized based on data structure. The results of the review found limitation related to data availability besides the manual structuring and the less use of unsupervised learning reflect complexity of handling textual accident data. Moreover, types of accidents happen in proportionally varying frequencies and need careful tackling outside basic assumptions of data pre-processing in addition to the general need for data pre-processing comparative studies and automated pipelines. The review also showed that data mining (DM) and unsupervised learning were less used especially with semi-structured and unstructured datasets reflecting maybe inefficient natural language processing (NLP) application with these types of learning. Among the reviewed articles, only four out of six prototypes were externally validated on construction environment thus we propose that future efforts would benefit from incorporating a standardized development method that also explicit how ML safety recommendation informs decision making. Future research should experiment and ascertain different choices in the pre-processing stage, validating the performance of the ML models and implementation in the construction practices. Finally, there are more advanced NLP methods that could be applied if domain specific repositories were available such as relation extraction and there are various advances that could be explored including large language models (LLMs).*

KEYWORDS: *Accident reports, construction, machine learning, natural language processing, safety.*

REFERENCE: *May Shayboun, Dimosthenis Kifokeris, Christian Koch (2025). A review of machine learning for analysing accident reports in the construction industry. Journal of Information Technology in Construction (ITcon), Vol. 30, pg. 439-460, DOI: 10.36680/j.itcon.2025.019*

1. INTRODUCTION

The global construction sectors are known for being both risky and having a high frequency of occupational accidents (Hoła and Szóstak 2015). Recently, there has been a notable increase in the literature on applying machine learning (ML)-based analysis on construction industry documentation (Xu et al. 2021a), including in the domain of safety (Hou et al. 2021, Yan et al. 2020). This involves using multiple technologies such as deep learning (DL) and ML (Elghaish et al. 2022) and exploiting these technologies to further advance knowledge on safety (Bhagwat and Delhi 2022). Accident-related data, such as those concerning circumstances, processes, and the people involved, can at least to some extent be found in accumulated accident reports in contracting companies and national registries. Moreover, the increased focus within the literature on ML-based analysis has been on extracting information from textual data on accidents, identifying contributing factors and classification (Hegde and Rokseth 2020), and identifying indicators for safety efforts (Garcia et al. 2022). In that vein, NLP applications for safety text analysis appear suitable for name entity recognition, text classification, information retrieval and text summarisation, aiming at accident causes analysis, hazard classification and accident information extraction (Liu et al. 2022, Ding et al. 2022). ML appear beneficial in research on underlying patterns especially with large volumes of accident-related data (Bilal et al. 2016, Hegde and Rokseth 2020, Vallmuur 2015).

Despite the above and although engaged industry professionals consider ML and AI to have a potential to improve safety performance (Dobrucali et al. 2024), a survey showed that the cost, practicability, and reliability of such emerging technologies constitute barriers to their adoption (Jan et al. 2021). Moreover, complexities in open- and long text-based information extraction and establishing dynamic information update are other barriers (Liu et al. 2022). Furthermore, there are several identified challenges in applying ML/ DL/ data mining (DM)/ text mining (TM) for analysing safety records in the construction sector. DM applications in construction are challenged by poor data quality and knowledge interpretation (Yan et al. 2020). Moreover, Khallaf and Khallaf (2021) found that limitations of DL for accident report analysis emanate from the need for manual labelling, and DL models are domain specific and require large data sets. Depending on the task and dataset, it may not be necessary that DL achieves superior levels of accuracy over ML; an accuracy and information extraction efficiency trade should then be established (Wu et al. 2022). Finally, the current literature has been criticized for analysing textual injury records while lacking sufficient descriptions of the methodologies used in processing the data and training the ML models. This has made it harder to understand the potential impact of ML applications on safety processes.

Given the above and building on previous reviews (Bhagwat and Delhi 2022, Elghaish et al. 2022) there is a need for a more focused review on ML-based analysis of accident reports. The sector's specific complexities and the need for a narrowed-down scope review led to the research question:

What is the state of the art of research applying ML to accident report analysis in construction?

The article focusses on the latest papers using ML in analysing accident reports and reflects on challenges and opportunities for possible implementation to analyse occupational accidents data within construction. A literature review to answer the research question and focuses on the applications to accident reports regarding its use of data preprocessing, algorithms, methods, testing and implementation. It is typical for a ML process to follow the steps of data exploration and pre-processing, and then model training, validation, and testing (Gopal 2018). Testing of a ML model corresponds to the internal performance of a ML model based on data that was set aside and not used for "training" i.e. feeding the model with data for growing its model (Gopal 2018). Training is followed by internal and external validation. External validation is "the use of independently derived datasets (hence, external), to validate the performance of a model that was trained on initial input data" (Ho et al. 2020). But we would extend this concept and suggest for this paper that external validation also includes any trials of ML-based models in real life situations with any forms of feedback indicating their accuracy, efficiency, and usability. The reviewed articles are therefore arranged into the following themes: data pre-processing, algorithms, testing and implementation.

2. METHOD

The literature review was conducted in iterations using the concept-centric framework (Webster and Watson 2002). Four databases were selected, namely Scopus, Science direct, Emerald and ProQuest, supplemented with a broader search in Google scholar. The review was conducted iteratively within the databases, with the broad search terms being "accident report," "construction," "machine learning." This search resulted in a high number of hits, therefore only the first appearing 100 were scanned because after that the search results were no longer relevant.

Then a more detailed search with a more specific search string including “building industry,” “accident cases,” “accident reports,” “accident records,” “accident data,” “machine learning,” “deep learning,” “data mining,” was conducted and the hits of each library are shown in Table 1. The literature search was thus targeted whilst still being sufficiently comprehensive (MacLure 2005).

Table 1: Number of hits and included articles in every database.

Database	**Num. hits**	**Num. included**
Scopus	18	4
Science direct	161	25
Emerald	42	6
ProQuest	55	2
Subtotal	**276**	**37**
Broad Serach	400	11
Total	**676**	**48**

The selection of papers using these criteria yielded a narrowly reference targeted list that is only related to the use of ML, DL, and DM on reported accidents’ data in construction. Any returned results in non-targeted contexts, e.g., the ones of chemical plants, steel plants, and car crashes were excluded. Articles analysing data from accident news or using web-crawling were also deemed out of scope. Articles analysing hazard reports and safety inspection documentation from construction sites were excluded too. Finally, considering the temporal dimension of the literature review, research on applying DL and DM in construction started to steadily increase from 2016 (Xu et al. 2021a), while accident analysis and injury investigation using past accidents cases increased sometime around 2017 (Bhagwat and Delhi 2022). Therefore, combining the above, the starting point for the review was set to 2017.

The literature search followed the process described in (Webster & Watson 2002). 276 publications (see Table 1) were abstract scanned and skimmed through if necessary. If the paper was a review itself, it was selected for consideration for the “Introduction”; otherwise, if it was an implementation paper (i.e., a ML model is developed in the paper) meeting the inclusion criteria, it was selected for consideration for the “Literature review” section. The collected articles were compared across the four databases to identify and remove duplicates (i.e., papers appearing in multiple databases). Finally, 37 articles were selected for including in-depth studies within a cross-section of the concepts and inclusion criteria mentioned above. During the in-depth studies common concepts and themes were identified and subsequently being allowed to organise the review (Webster & Watson 2002). Each article was ordered by the identified concepts. The organization of the review was done to units of analysis/concepts which became; data pre-processing, algorithm type and training the ML model, testing algorithm performance, and implementation of ML analysis.

3. LITERATURE REVIEW

The governing concept of the review derives from the understanding of a generic ML development process steps combined with the identification of main topics and wordings from the reviewed literature (Webster & Watson 2002). The review commences with analysing data preprocessing in three turns depending on data type; structured, manually structured, semi-structured and unstructured. Then it continues with algorithms to end with test and implementation.

3.1 Data pre-processing

Many of the reviewed studies highlighted data pre-processing as an essential step – i.e., trying to formulate and represent the data in a way that better fits the modelling and algorithmic structure of the ML system (Shehab et al. 2021). Data pre-processing can help in terms of improving the time taken for the analysis, the utilization of resources, storage, efficiency, and even the output-gained information (Shehab et al. 2021).

3.1.1 Structured and manually structured data pre-processing

Structured datasets are "organized properly into a formatted repository" and in a tabular format (Solanki et al. 2019). A structured dataset can contain labels which often makes it suitable for using supervised learning (Han et al. 2011). Structured datasets are advantageous, since they are simpler to process and manage (Solanki et al. 2019). However, only few structured datasets exist in the reviewed literature while other unstructured datasets can become structured by manual feature extraction; the latter was found in 10 papers (see Table 6). Manually structured datasets are usually smaller, containing only hundreds of accident reports. This is limiting for ML- and DL-based analyses, as more available data can lead to more efficient learning especially for DL (Sun et al. 2017).

Table 2: Manually structured/Structured datasets pre-processing.

Main process	Subprocess	Technique	Paper
Data cleaning	Missing values	Imputation	Poh et al. (2018), Jiang et al. (2023), Halabi et al. (2022), Koc and Gurgun (2022)
		Removing	Ameri et al. (2017), Xu and Zou (2021), Koc and Gurgun (2022), Kang et al. (2022), Alkaissy et al. (2023)
		Replacing	Koc and Gurgun (2022), Kang et al. (2022)
	Outlier handling	Removing	Sabet et al. (2021), Alkaissy et al. (2023)
		Considering	Koc and Gurgun (2022)
		Upper and lower limit	Koc and Gurgun (2022)
	Duplicate removal		Ameri et al. (2017)
	Data binning		Koc and Gurgun (2022)
Data integration	Dimensionality reduction	Principal component analysis	Alkaissy et al. (2023)
Data reduction	Feature selection	Boruta feature selection function	Poh et al. (2018)
		Delphi method feature selection	Ayhan and Tokdemir (2019)
Data transformation	Discretization		Ameri et al. (2017), Koc and Gurgun (2022), Deng et al. (2024)
	One-hot encoding		Koc et al (2021), Koc et al. (2022a), Koc and Gurgun (2022), Zermane et al. (2023), Alkaissy et al. (2023)
	Label encoding		Koc et al. (2021), Sabet et al. (2021), Zermane et al. (2023)
	Weighted Adjacency Matrix		Assaad and El-adaway (2021)
	Fuzzification		Ameri et al. (2017)
	Normalization		Koc et al. (2021)
	Binarization		Ayhan and Tokdemir (2019)
	Scaling		Koc and Gurgun (2022)
Statistical methods	Chi-Square Test		Lee et al. (2020), Gholizadeh et al. (2021), Zhu et al. (2021)
	Cramer's V test		Lee et al. (2020), Gholizadeh et al. (2021)
	LCCA		Lee et al. (2020)
	Partial autocorrelation function (PACF)		Koc et al. (2022)
	Autocorrelation function (ACF)		Koc et al. (2022)
Data resampling	SMOTE		Poh et al. (2018), Zhu et al. (2021), Koc et al. (2022a)
	RUS		Kang and Ryu, (2019), Koc et al. (2022a)
	ROS		Choi et al. (2020), Koc et al. (2022a)
	Manual restructuring		Jiang et al. (2023)
	Resampling manually by regrouping target		Kang et al. (2022)
Manual feature extraction			Lee et al. (2020), Gholizadeh et al. (2021), Tang et al. (2021),Chen et al. (2023), Guo et al. (2022), Halabi et al. (2022), Shao et al. (2023), Alkaissy et al. (2023), Deng et al. (2024), Lu et al. (2023), Rafindadi et al. (2023)
NLP algorithm Tixier et al. (2016)			Tixier et al. (2017)

Table 2 shows and categorises data pre-processing techniques for manually structured and structured datasets. In this table, the two data types are organised together because we deemed the pre-processing steps to be very similar except manually organizing the data. We characterised the data pre-processing steps and types mainly based on the categorization by Han et al.'s (2011). Manual structuring may require a theoretical framework of accident information attributes. Such as stakeholder theory, system thinking, Construction Accident Causation (Lu et al. 2023), or a literature review on accident analysis in construction sites (Alkaissy et al. 2023, Tang et al. 2021) and possibly combined with experts' opinions (Rafindadi et al. 2023, Tang et al. 2021). Manual structuring could be combined with feature extraction for textual accident reports and then organize the extracted features into predefined accident attributes – e.g., extracting safety factors by manually assigning attributes to identified risk (i.e., location, work type, accident causes, and results) (Tang et al. 2021).

Manual labelling and structuring of accidents reports takes a considerable proportion of data-processing (Table 6). The dominant data pre-processing steps in structured-categorical datasets can also include label encoding, one-hot encoding, normalization, and dimensionality reduction (Koc et al. 2021). Moreover, steps in time series analysis include defining time related input and output (Koc et al. 2022). Typically, some data quality issues can surface in pre-processing, such as missing values and outliers which could result in loss of many accident cases (Koc and Gurgun, 2022). Jiang et al. (2023) explicitly tested with handling missing values using RF, KNN multiple imputation, single imputation, and not handling missing values. It was found that RF multiple imputation yielded good results if the important predictors did not have a high percentage of missing data (Jiang et al. 2023). There is less discussion about which type of data pre-processing methods are best used in relation to construction accident reports. However, Lee et al. (2020) proposed that reducing variables and elements of accident datasets is favourable for accident predictions and advise to perform data pre-processing by comparing different methods instead of one.

There had been multiple methods for managing the class imbalance for the ML model training step (see Table 2). Zhu et al. (2021) argued that SMOTE is better than random oversampling (ROS) as it adds artificially synthesized samples and does not risk overfitting. While ROS was chosen as the best method because it fits better with the categorical values in the dataset (Choi et al. 2020; Koc and Gurgun 2022). Kang and Ryu (2019) and Koc et al. (2022a) used random undersampling (RUS), which is a method that reduces the major classes, resulting in a reduction of the data sample from 9795 to 6374 accident reports (Kang and Ryu 2019). While Koc and Gurgun's (2022) study illustrated the need for an automatic data pre-processing pipeline by detailing an automatic scenario building and testing of various steps that could be taken to improve the ML data analysis.

3.1.2 Semi-structured datasets pre-processing

Semi-structured data "is not referred to the tabular format but… has some organizational property" (Solanki et al. 2019). In our case such data is found in reports containing textual descriptions of accident cases accompanied with structured data, such as the type of accident and severity level presented in columns or tables. Each of these organized features can be considered as a label.

Challenges related to data characteristics also tie with challenges in data handling and pre-processing. Two main themes were found in the reviewed literature; the first theme includes stop word removal, tokenization, text cleaning, and word embedding (Arumugam and Shanmugamani 2018). Typically, word embedding, and text vectorization are not considered as text processing procedures but in this paper, we classified these methods together with pre-processing for practically organizing the paper structure. The other theme is related to manual labelling, text segmentation, and using domain specific dictionaries and different NLP algorithms. In addition to textual data pre-processing, Li et al. (2021) also performed data cleaning of removing reports with missing values, removing accident types comprising of a small number of reports of certain accident types, and chi-square tests for feature selection.

Word embedding is one of the most important steps in processing textual data. Commonly used word embedding techniques are Word2Vec, Wikipedia Global Vector for Word Representation (GloVe), and Bidirectional Encoder Representations from Transformers (BERT) (see Table 3). These word embedding algorithms are usually trained using generic corpora but can also be retrained with domain-specific corpora such as accident reports that are not needed for training a ML classification model (Zhang 2019, Zhang et al. 2020, Fang et al. 2020, Baker et al. 2020b). Zhang et al. (2020) used BERT as a text pre-training algorithm instead of Word2vec, for being more efficient in handling text multi-meaning. Alternatively, data segments that were not used to train the ML model can be utilized to pre-train Word2Vec, which provides an advantage of using word embedding with domain-related vocabulary (Zhang 2019, Baker et al. 2020b). Goldberg (2022) tested with six different state-of-art word embedding algorithms (including BERT, Glove and Word2vec) with a baseline of "term frequency – inverse document frequency” (TF-IDF) vectorization and concluded that BERT performed best compared to Glove and Word2vec.

An important decision in text pre-processing concerns the used dictionaries and tokenization settings such as unigram, bi-gram and combinations of unigram, bi-gram, and trigram. Goh and Ubeynarayana (2017) found that unigrams provided the best classification results although TF-IDF should function better with bi-gram and trigram in capturing the context and leading to less misclassifications. Word embedding also exist in different modes, such as the continuous bag of words (CBOW) and the skip-gram model which Zhong et al. (2020a) found more suitable for accident report narratives because it may contain sparse text features.

Table 3: Semi-structured datasets pre-processing.

Main process	Subprocess or technique	Papers
Text processing	Stop word removal	Goh and Ubeynarayana (2017), Zhang (2019), Zhang et al. (2019), Cheng et al (2020), Shrestha et al. (2020), Zhang et al. (2020), Zhong et al. (2020a), Li et al. (2021), Ma et al. (2021), Gupta et al. (2022), Pan et al. (2022)
	Word stemming	Goh and Ubeynarayana (2017)
	Tokenization	Goh and Ubeynarayana (2017), Zhang (2019), Zhang et al. (2019), Baker et al. (2020b), Cheng et al (2020), Fang et al. (2020), Zhong et al. (2020a), Ma et al. (2021), Gupta et al. (2022)
	Convert to lower case	Baker et al. (2020b),Shrestha et al. (2020), Zhang et al. (2020), Gupta et al. (2022)
	Word segmentation	Baker et al. (2020b), Zhang et al. (2020), Li et al. (2021), Luo et al. (2023)
	Text cleaning	Baker et al. (2020b) Cheng et al (2020), Fang et al. (2020), Shrestha et al. (2020), Zhang et al. (2020), Goldberg (2022), Ma et al. (2021), Pan et al. (2022)
	Lemmatization	Zhang (2019), Zhang et al. (2019), Zhang et al. (2020), Shrestha et al. (2020), Gupta et al. (2022), Pan et al. (2022)
	Part of speech tagging	Zhang (2019), Zhang et al. (2019), Ma et al. (2021)
Vectorization	TF-IDF	Goh and Ubeynarayana (2017), Zhang et al. (2019), Goldberg (2022), Li et al. (2021), Ma et al. (2021) Gupta et al. (2022)"
Word embedding	Skip-gram word2vec	Zhang (2019), Baker et al. (2020b), Zhang et al. (2020), Zhong et al. (2020a), Ma et al. (2021), Goldberg (2022)
	GloVe	Cheng et al. (2020), Goldberg (2022), Gupta et al. (2022)
	BERT	Zhang et al. (2020), Goldberg (2022), Chen et al. (2022b)
	BioBERT	Goldberg (2022)
	SciBERT	Goldberg (2022)
	FastText	Goldberg (2022)
	Doc2vec	Li et al. (2021)
Manual labelling		Goh and Ubeynarayana (2017), Zhang (2019), Zhang et al. (2019), Zhong et al. (2020a), Shrestha et al. (2020), Luo et al. (2023)
Resampling	Inversed proportional weights	Goh and Ubeynarayana (2017), Baker et al. (2020b), Baker et al. (2020a)
	Manual labelling	Zhang (2019), Deng et al. (2020), Zhang et al. (2020)
	Data Augmentation	Gupta et al. (2022)
	SMOTE	Luo et al. (2023)
NLP algorithm Tixier et al. (2016)		Baker et al. (2020a)
Dimentionality reduction	t-SNE	Zhang (2019), Li et al. (2021)"
	PCA	Zhang (2019)
Multicollinearity analysis	Spearman correlation coefficient	Luo et al. (2023)

Data pre-processing might also include text feature extraction (Ma et al. 2021), or manually labelling the accident reports (Zhong et al. 2020a, Goh and Ubeynarayana 2017, Shrestha et al. 2020). Manual processing can also be combined with the extraction of safety factors by identifying the extracted keywords using TF-IDF and manually structuring attributes into a dictionary to the identified risk factors (i.e., location, work type, accident causes, and results) (Ma et al. 20201). Similarly, word segmentation can be performed statistically; then, the contextualization of accident attributes can be based on system engineering theory (Luo et al. 2023). While features have automatically been extracted from textual reports via an NLP algorithm especially developed for the industrial, infrastructure, and mining domains (Baker et al. 2020a).

Accident data can feature classes with a considerable variation in the number of included instances, most notably fatal accidents as opposed to other groups of accidents (so-called "unbalanced" classes). Such variation imposes a particular challenge in data pre-processing and ML-based analysis because the model's development tends to involve occurrences of misclassification of sparsely populated classes simply because they are harder to recognize compared to more populated classes. Frequency variation has been found in classes including injury severity, energy type involved, causes, accident types, and injured body parts, which affected the classification accuracy performance (see "Testing algorithm performance" section).

3.1.3 Unstructured data pre-processing

Unstructured dataset is one whose "structure can not be predicted … not organised into a pre-defined manner" (Solanki et al. 2019). In the case of accident reports, such unstructured data is textual (mostly continuous text passages). In the relevant references, the collected occupational accident reports included unlabelled textual report data (i.e., the instances were not initially attributed into the sets of specific classes).

For unstructured datasets, pre-processing is similar to the one used for semi-structured data (Table 3). Zhang (2019) and Zhang et al. (2019) performed experiments with labelled-unstructured and unlabelled-unstructured accident reports data. For the labelled dataset, data pre-processing included stop word removal, part of speech tagging (POS), and lemmatization. However, stop word removal is not performed with Word2vec, because it depends on the broader context of sentences (Zhang 2019).

Table 4: Unstructured datasets pre-processing.

Main process	Sub-process or technique	Papers
Text processing	Tokenization	Kim and Chi (2019), Zhang (2019), Zhang et al. (2019), Xu et al. (2021b)
	Stop word removal	Zhang et al. (2019)
	Convert to lower case	Zhang (2019)
	Text cleaning	Zhang (2019), Na et al. (2021)
	Part of speech tagging	Zhang et al. (2019), Xu et al. (2021b)
	Text segmentation	Na et al. (2021)
	Syntactic dependency parsing	Xu et al. (2021b)
Word embedding	Word2vec	Kim and Chi (2019), Zhang (2019)
Vectorization	Okapi BM25	Kim and Chi (2019)
	Text segmentation and TF-H	Na et al. (2021)
Resampling	Manual labelling	Zhang (2019)

As mentioned earlier, word embedding algorithms can be retrained with a domain-specific corpus which takes advantage of larger unlabelled datasets in case of lacking labels (Zhang 2019). Na et al. (2022) improved text mining by performing text segmentation via constructing a domain specific lexicon for text tokenization, based on defined safety and construction terms. They also proposed the use of the accumulative entropy weighted term frequency (TF-H) instead of TF-IDF, because for safety risks TF-H can account to a term's uniform distribution in the corpus (Na et al. 2021).

Domain and language related dictionaries were used in the tokenization of Korean accident reports (Kim and Chi 2019). The method developed by Kim and Chi (2019) consisted of developing a construction accident thesaurus to capture words and their synonyms, or the unique representation of words that are usually used in a construction context. This was done by combining a thesaurus of construction terms with accident cases and using Word2vec to expand queries in the accident reports database. Moreover, manual definitions of features and the associated semantic roles were defined for creating rules for accident information extraction (Kim and Chi 2019).

3.2 Algorithm type

The analysis of accident reports is mainly treated in the reviewed literature as an information extraction task or a classification task – i.e., classification of accident type, severity, and causes. Given this observation, the literature can be categorized according to the type of analytical algorithms used (DL, ML, DM), and the characterisation of the data type.

3.2.1 Algorithm type for structured datasets

The analysis of structured accident reports was performed by DM, ML, and DL algorithms (see Table 5). The table below breaks down the algorithms used in each article and the purpose of the ML model. Then, the best performing algorithms are highlighted in bold.

The data mining analysis included unsupervised algorithms such as association rule mining (ARM) (Amiri et al. 2017, Xu and Zou 2021), the Apriori algorithm (Sabet et al. 2021), and graph mining with hierarchal clustering principal components (HCPC) (Tixier et al. 2017). The purpose of the data mining algorithms was to extract risk rules and discover possible rules relating accident attributes. In some cases, experts check on discovered associations of safety attributes to separate the valid cases of new safety knowledge (Tixier et al. 2017). Moreover, using fuzzification for analysing accident reports helped in tackling data uncertainties and ambiguity (Amiri et al. 2017).

ML algorithms were also used for various prediction metrics in structured data, such as severity (Poh et al. 2018, Koc and Gurgun 2022), the likelihood of fatality (Choi et al. 2020), and the phase of occurrence of crane operation accidents (Jiang et al. 2023). The most frequently used algorithms are random forest (RF), XGBoost and AdaBoostb (see Table 5). Interestingly, tree-based models – whether it is RF (Poh et al. 2018, Choi et al. 2020) or XGBoost (Koc et al. 2021) seem to have an advantage over other algorithms. RF can even outperform DL algorithms such as MLP (Zermane et al. 2023) and ANNs (Koc et al. 2022a). In another effort, the AutoML model performed best (Zhu et al. 2021). Optimizing hyperparameters should also be considered for tree-based models due to their sensitivity to parameters such as the number of trees and learning rate. The relevant methods that were used are HyperOpt (Kang and Ryu 2019), Grid search OOB error (Jiang et al. 2023), and the GA (Koc et al. 2021).

Applied DL to structured data was mostly used to predict the severity and number of accidents (Koc et al. 2022). The approach of Koc et al. (2022) applied a hybrid wavelet transformation with different supervised algorithms and time series data, where the optimized ANN with Broyden-Fletcher-Goldfarb-Shanno (BFGS) outperformed other algorithms.

Table 5: Algorithms types of structured datasets. Text in bold highlights the best performing algorithm.

	Reference	**Training algorithm**	**Implementation**
Data mining	Amiri et al. (2017)	ARM, DT	Rule based risk assessment
	Tixier et al. (2017)	Graph mining, HCPC	Cluster accident attributes
	Sabet et al. (2021)	LR, RF, NB, **RNN-LSTM,** Apriori algorithm	Accident prediction and rule mining
	Xu and Zou (2021)	ARM	Rules between accident attributes
Machine learning	Poh et al. (2018)	SVM, LR, **RF**, DT, KNN	Predict severity
	Kang and Ryu (2019)	RF	Predict accident types
	Choi et al. (2020)	AdaBoost, LR, **RF**	Predict likelihood of a fatality
	Jiang et al. (2023)	RF	Predict phase of tower crane accident
	Koc et al. (2021)	RF, ET, AdaBoost, **XGBoost**	Predict permanent disability
	Kang et al. (2022)	RF	Predict injury lost days
	Koc and Gurgun (2022)	XGBoost	Predict severity.
Deep learning	Ayhan and Tokdemir	**ANNs**, multiple regression	Predict severity
	Zhu et al. (2021)	LR, DT, RF SVM, NB, KNN, MLP, **AutoML**,	Predict accident severity
	Koc et al. (2022)	**W-ANN**, W-MARS, W-SVR	Predict the number of accidents
	Koc et al. (2022a)	**RF**, NB, ANN, KNN	Predict severity
	Zermane et al. (2023)	XGBoost, **RF**, DT, MLP, KNN, SVM, LR	Predict fatal fall factors

3.2.2 Algorithm type for manually structured datasets

For manually structured datasets of accident reports, data mining and shallow learning techniques were applied (see Table 6). It is rather interesting to observe a tendency in the reviewed literature to apply unsupervised data mining algorithms (e.g., Apriori, ARM) and even clustering to manually structured and data. This might indicate that such type of learning is appreciated for understanding of what combination of different factors play a role (Assaad and El-adaway 2021) and it works well with small datasets (Rafindadi et al. 2023). Chen et al.'s (2023) approach was distinguished, as they considered spatial features of operations to measure cross-operation interaction (e.g., falling from heights was accompanied by collapse accidents and being hit by objects).

ML was performed for predicting accident severity (Lee et al. 2020), electrical accidents (Gholizadeh et al. 2021), the fatality of fall accidents (Halabi et al. 2022) and the similarity between accident cases (Lu et al. 2023). In this approach, several algorithms were used (Table 6), such decision trees (DT), LogR, support vector machines (SVM), Classification and Regression Trees (CART) and k-nearest neighbour (KNN) for clustering. Moreover, principal component analysis (PCA) is an unsupervised learning method usually utilized for dimensionality reduction in high dimensional data, but in the context of construction accidents it was used to visualise the severity level together with predictive attributes (year, type of accident, and injured part) to understand which variables are more correlated with the outcome (Lee et al. 2020).

Table 6: Algorithms types of structured datasets. Text in bold highlights the best performing algorithm.

	Reference	Training algorithm	Implementation
Data mining	Assaad and El-adaway (2021)	Spectral clustering, frequent pattern mining, Apriori	Association rule mining of fatal accident causes
	Tang et al. (2021)	Apriori	Safety risk knowledge extraction
	Chen et al. (2023)	Apriori	Cross operation risk visualization
	Guo et al. (2022)	Frequent pattern growth (ARM)	Unsafe behaviour factors correlation extraction
	Shao et al. (2023)	Apriori	Association rule mining for collapse accident types
	Deng et al. (2024)	Apriori	Association rule mining for accident types and causes
	Rafindadi et al. (2023)	Apriori	Association rule mining of factors
Machine learning	Lee et al. (2020)	**Ensemble**, SVM, PCA	Predict accident severity
	Gholizadeh et al. (2021)	CART	Predict electrical accidents
	Halabi et al. (2022)	LogR	Predict fall accident fatality
	Lu et al. (2023)	KNN	Case similarity measurement

3.2.3 Algorithm type for semi-structured datasets

The analysis of semi-structured datasets was performed by ML and DL algorithms (see Table 7). ML algorithms were used for multiple purposes, including prediction of collapse, accident severity (Luo et al. 2023), upstream precursors, energy source, accident type, injury severity (Goh and Ubeynarayana 2017, Shrestha et al. 2020), incident type, injury type, body part (Baker et al. 2020a). It is observed that SVM and tree-based models such as RF and XGBoost stand out among used algorithms either by choice of the authors or by outperforming other algorithms for the same task. SVM, XGBoost, and RF were used as state-of-art algorithms (Baker et al. 2020a) but a stacking model of the tree-based algorithms showed to perform better than each of the algorithms alone.

In the presented efforts below, we find that optimization algorithms were employed to increase accuracy and assign optimised hyperparameter values. Goh and Ubeynarayana (2017) used brute force exhaustive search for optimizing the C parameter related to SVM, while Luo et al. (2023) used the Gini index or the out-of-bag (OOB) error to measure the contribution of each feature to the DT, as well as RandomizedSearchCV for hyperparameter tuning of the RF. Moreover, Sequential Quadratic Programming (SQP) was used for optimizing an ensemble model (Zhang et al 2019).

It can be observed in Table 7 that DL was a common choice for analysing accident reports with a semi-structured format. This is probably expected as DL does not need feature extraction for performing classification using data such as images and text. DL was represented by using different variations of deep neural networks in what appears as an experimental pattern, such as the convolutional bidirectional long short-term memory (C-BiLSTM) (Zhang et al. 2020, Zhang 2019, Deng et al. 2020), and Context Connotative Network (hybrid C-LSTM) (Gupta et al. 2022). Graph-based neural networks such as the relational graph convolutional network (R-GCN) (Chen et al. 2022b) and Graph Convolutional Network (GCN) (Pan et al. 2022) were utilized as well. On top of modelling with different algorithms, Li et al. (2021) used K-NN as a meta-learner algorithm recommending the best performing algorithm for accident consequence prediction every time a new dataset is added to the model; this is done by comparing the new dataset to a meta-knowledge database which gives the model the possibility to update with new data. Moreover, topic modelling (Zhong et al. 2020a) and data mining (Ma et al. 2021) were combined with deep learning for processing unlabelled data and mine safety risk factors.

The choice of BiLSTM was based on its superior performance in extracting information from the text, as well as for examining information before and after the word for better context understanding (Zhang et al. 2020). In that vein, convolution layers are superior with extracting n-grams features from text and capture local correlations (Deng et al. 2020). SGRU is also a strong variant of long short-term memory (LSTM), but more computationally efficient and combined with parameter optimization of the neural network (Cheng et al. 2020).

Optimization is usually a very important part of training DL algorithms. The reviewed articles in this category used multiple optimization methods, such as Stochastic Gradient Descent (Baker et al. 2020b), Symbiotic Organisms Search (Cheng et al. 2020) and Adam (Gupta et al. 2022, Chen et al. 2022b).

Table 7: Algorithms types of semi-structured datasets. Text in bold highlights the best performing algorithm.

	References	Training Algorithm	Implementation
Machine learning	Goh and Ubeynarayana (2017)	**SVM**, LinR, RF, KNN, DT, NB	Predict accident types
	Zhang et al. (2019)	**Optimized Ensemble**, Ensemble, LR, SVM, NB, KNN, DT	Predict accident causes
	Baker et al. (2020a)	RF, XGBoost, **Linear SVM, (RF`+XGBoost)**	Predict accident severity, incident type, injury type, body part
	Shrestha et al. (2020)	SVM	Predict upstream precursors, energy source, accident type, injury severity
	Luo et al. (2023)	RF	Predict collapse accident severity
Deep learning	Baker et al. (2020b)	HAN, CNN, **(TF-IDF`+SVM)**	Predict accident severity, incident type, injury type, body part
	Zhang (2019)	**BiLSTM**, LogR, DT, KNN, NB, SVM	Classify accident types
	Cheng et al. (2020)	**SGRU**, NLP KNN, LSTM, DT, GRU, SVM, LR, NB	Predict accident causes
	Deng et al. (2020)	**C-BiLSTM**, SVM, NB, LR	Predict accident types
	Fang et al. (2020)	**Modified BERT**, Fast Text, (TextCNN + BiGRU), TextCNN, (BiGRU + Attention), TextRCNN	Predict near miss categories
	Zhang et al. (2020)	**C-BiLSTM**, SVM, NB, LR, CNN, LSTM, BiLSTM	Predict accident types
	Zhong et al. (2020a)	**CNN**, SVM, NB, KNN, LDA	Predict accident causes Topic mining
	Li et al. (2021)	KNN, DT, ANN, **RF**, NB, SVM, LR	Predict accident consequences
	Ma et al. (2021)	**CNN**, SVM, NB, KNN, Apriori algorithm	Safety risk assessment and risk prediction
	Chen et al (2022b)	**R-GCN**, DT, KNN, NB, LogR, SGRU, LSTM, GRU	Classify accident causes
	Goldberg (2022)	**Bi-LSTM**, SVM, MLP, LR, DT, RF, NB	Predict body part(s), source, event, hospitalization, and amputation.
	Gupta et al. (2022)	**Hybrid C-LSTM**, NB, DT, SVM, KNN, LogR	Accident causes classification
	Pan et al. (2022)	**GCN**, TextRank, SVM	Prediction of accident type and injury type

3.2.4 Algorithm type for unstructured datasets

Unstructured datasets are challenging, because they may require manual labelling, identification of rules in rule-based information extraction (Kim and Chi 2019), or identification of domain knowledge elements – based on experts' opinions – from the data; as an example, domain knowledge extraction (DKE) was performed by NLP rule-based extraction in Xu et al. (2021b). On the other end, validation and contextualization of extracted information are sometimes done by manually checking the results (Na et al. 2021).

Unstructured data were mostly analysed to extract information from accident reports (e.g., extracting common objects contributing to causing accidents by adopting the rule-based chunking approach) (Zhang et al. 2019). The conditional random fields (CRFs) algorithm was used for its effectiveness in labelling information from textual data and considering the sequence of a sentence (Kim and Chi 2019). This approach required definitions of tacit knowledge, where a tacit knowledge extraction (CRFs) model was trained and optimised for controlling the degree of overfitting by using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. This model worked as an information extraction system from accident cases after grouping the similar ones based on a query. While the process becomes more complicated with completely unstructured text, both general semantics roles and additional rules were created by manual data analysis (Kim and Chi 2019). Another approach for DKE was proposed by applying the semantic part-of-speech tokenization and the syntactic dependencies of word analysis (Xu et al. 2021b).

Na et al. (2021) performed a comparative analysis of different methods of term importance representation including document frequency (DF), TF-IDF, term frequency (TF), and information entropy weighted term frequency (TF-H). For that, an updated lexicon was constructed to manually capture the construction related terms through experts analysing the data to add the specific terms to a general lexicon. Na et al. (2021) found that TF-H – unlike other methods – accounts for the document distribution of terms, which is very important for safety risk factors.

Table 8: Algorithms types of unstructured datasets.

	Reference	Training algorithm	Implementation
Data mining	Kim and Chi (2019)	CRFs	Information retrieval system
	Na et al. (2021)	TF-H, TF, DF, TF-IDF	Risk factors extracting
	Xu et al. (2021b)	POS, DOW	Rule based knowledge extraction
Information extraction	Zhang et al. (2019)	Rule based chunking	Information extraction
	Zhang (2019)	Train word embedding	Information extraction

3.3 Testing algorithm performance

The generalization capability of a ML algorithm is usually evaluated using an unseen split of data (Riccio et al. 2020). This is referred to as the test split and is not used during model training or hyperparameter tuning and validation (Riccio et al. 2020). The testing of a classification algorithm is used to evaluate how well the algorithm classifies the target after the training step and distinguishes which of the compared algorithms performs better against each other for performing the same task. On the other hand, external validation is "the use of independently derived datasets (hence, external), to validate the performance of a model that was trained on initial input data" (Ho et al. 2020). Here we extended this concept to also include any trials of ML-based models in real life situations with any forms of feedback indicating their accuracy, efficiency, and usability. In this section, aspects of algorithmic performance are presented, along with the ML model's purpose.

DL was used for both structured and semi-structured datasets (See Tables 6 and 7). In the case of semi-structured datasets, advanced DL algorithms almost always outperformed supervised ML algorithms (ex. SVM, NB, LogR, KNN, DT) and elementary DL ones (LSTM, GRU, CNN) (Zhang 2019, Cheng et al. 2020, Deng et al. 2020, Zhang et al. 2020, Zhong et al. 2020a, Goldberg 2022, Gupta et al. 2022). Even more, R-GCN outperformed LSTM, SGRU and GRU (Chen et al. 2022b). It is important to note that algorithmic performance depends on which metrics are used for testing, the choices of parameter tuning, and other factors. For example, the weighted average F1 score is a better performance metric than a single F1 score, primarily when the data is characterized by class imbalance (Zhang et al. 2020). Moreover, algorithmic performance can depend on the parameter tuning of the word embedding step using unigrams, bigrams of two different dimensionalities, and the results showed that bigrams are constantly superior (Zhang 2019). Furthermore, testing BERT compared to other DL text classification algorithms for near-misses classification had the advantage of a pre-trained bi-directional network with an altered architecture (Fang et al. 2020). For two different experiments about body part classification in the same dataset with Baker et al' (2020a), HAN had the best performance but, TF-IDF+SVM had an overall better performance than HAN. In this case, the performance of DL was explained as it can sometimes use part of the narrative that indicates prediction outcomes, which is misleading the learning; instead, speech parts that are potential predictors (such as circumstances or work environment) could be used (Baker et al. 2020a). Furthermore, DL requires large data volumes and other data sources such as success cases may be needed to learn what goes well instead of safety related attributes and incidents only (Baker et al. 2020a).

In the structured datasets which were analysed with DL, ANNs performed with the fatalities being predicted with 100% accuracy, then testing accuracy dropped by 50% (Ayhan and Tokdemir 2019), and the W-ANN for time series data performed better than other algorithms. Nonetheless, AutoML and RF outperformed MLP and ANNs in other cases (Zhu et al. 2021, Koc et al 2022a, Zermane et al. 2023). In addition to testing with different algorithms, Zhu et al. (2021) evaluated both the original and the adapted SMOTE datasets (see data pre-processing section) for comparing the evaluation metrics in accident severity prediction. The results showed that SMOTE demonstrated a slight improvement in the testing, but the authors found the method prone to overfitting (Zhu et al. 2021). The best F1-score was achieved by AutoML) (Zhu et al. 2021).

NLP was applied to extract the causes of accidents and the objects which contributed to them (Zhang et al. 2019). Multiple classification algorithms were tested, and the performance was considered low (Zhang et al. 2019), and the authors attributed that to natural language not being precise, and the difficulty of developing comprehensive rules to cover all meanings of different expressions (Zhang et al. 2019, Xu et al. 20201b).

ML was used for all types of datasets, except for the unstructured ones (see Tables 6 and 8). For the structured datasets, the RF outperformed other classification algorithms, such as, indicatively, SVM, KNN, DT, LR, and AdaBoost (Poh et al. 2018, Choi et al. 2020). In another example, XGBoost outperformed RF, Extra Trees (ET), and AdaBoost (Koc et al. 2021).

For the manually structured datasets, Lee et al. (2020) tested with nested cross validation for an ensemble model and a SVM one, with the ensemble model outperforming the SVM. However, prediction performance measured by multiple metrics showed to be low for predicting dates, which is kind of expected to be less correlated with an accident (Lee et al. 2020). But even for the severity and other accident-related attributes, the predictions were not very satisfactory, since the data was concentrated on certain elements and affected prediction (Lee et al. 2020).

For the semi-structured datasets that were analysed with ML (see Table 7), the best performance was obtained in classifying the injury severity with the SVM (Baker et al. 2020a). A combination of SVM and RF was applied as a single algorithm to classify injury severity (Shrestha et al. 2020) and collapse accident severity (Luo et al. 2023).

An observed pattern in the reviewed literature relates to the misclassification of the less populated classes or the narratives of different types of accidents that contain similarities in description, such "struck by moving objects" and "collapse of object" (Goh and Ubeynarayana, 2017). Additionally, misclassifications for accident types can happen because of labelling, as it is difficult even for humans to label accident types correctly (Gupta et al. 2022). Although the averaged F1 scores might indicate acceptable performances for the tested algorithms, the low F1 scores of the single classes are as important. The misclassification of minority classes of accident severity, type, and causes was a challenge for a few of the reviewed research efforts. For a single cause classification (collapse of an object), the model did not perform as well (Zhang et al. 2020). because these instances were found to have unique occurrences and characteristics when individually checked (Zhu et al. 2021). Another example was found with confusing "falling object" with "moving object" because of the similarity in the word vectors, while the "electrocution" category had lower accuracy score compared to other classes because of the less frequent accident data in it (Ma et al. 2021). It is often advised to use a resampling technique for unbalanced datasets, but it is not always known how efficient these are. Koc et al. (2022a) tested with different resampling techniques including RUS, ROS, and SMOTE and found significant improvements when ROS and RUS algorithms were applied compared to not resampling or SMOTE; at the same time, NB performed better without the data being subjected to any resampling. Alternatively, one approach suggested a solution for facing smaller accident type data by a meta-learner knowledge base (Li et al. 2021).

3.4 Implementation of ML-based analysis

Model usage is one of the most central ML development processes (Bilal and Oyedele 2020). The reviewed literature presents two types of propositions to use ML-based models in the analysis of accident reports. Models within the literature which suggested a possible use of the ML analysis results might be labelled as conceptual propositions, those suggesting a precise implementation of ML-based data analytics were termed prototypes, and if they tested the ML model's use, they were termed as externally validated.

One prototype was proposed for information retrieval and knowledge extraction (Kim and chi, 2019). In this case, the authors assigned semantic roles for the elements that characterize the accident and defined the roles as predicates (i.e., "accident result,") and effector ("hazard object," "location," "hazard position," and purpose "work process”). The rule-based dataset extraction was tested against experts tasked with labelling and was found very close to the experts’ labelling (Kim and Chi, 2019). Another model was presented as an integrated framework of accident type classification, feature ranking and calculating the cascading effect on project tasks based on the effect the risk factors have on the time of a task (Ma et al. 2021). A rule-based safety recommendation was also developed based on the safety management and known causes (identified by literature review) integrated with data mining of accident causes (Tang et al. 2021). The latter requires the interpretation of experts for the recommended safety instruction, while an expert survey showed a positive response in terms of benefit to safety management (Tang et al. 2021). One of the examples that Kim and Chi (2019) presented showed that there is more than one consequence other than the injury– such as damage to infrastructure and material– while only one consequence was extracted from the accident report. Moreover, lower recall values compared to F1 score because duplicate features of classes exist and overlap with different classes in the learning process (Kim and Chi 2019). This indicates that the labelling and classification of accident components are highly dependent on the labels that specialists assign to them. The same observation can be made in rule-based data mining examples (Tang et al. 2021, Xu et al. 2021b, Ma et al. 2021).

When it comes to testing of ML models in real life, the material is also limited. We refer to this as external validation to clarify the distinction from the often-used internal validation, done within a pre-existing data set. Therefore, we rarely have insights into validation criteria and challenges of implementations. Lu et al. (2023) presented the concept of a hybrid risk response model by classifying factors based on importance. Then, they suggested those as response strategies for risk assessments based on a similarity search through the accident database and connected this system to existing real-time data collection streams on site (Lu et al. 2023). This system was validated for a one-month observation period in a construction project and in combination with a video alert module (Lu et al. 2023). Another system was tested with three large construction sites in terms of applicability, comprehensiveness, and accuracy (Amiri et al. 2017). At the same time, validation can be performed by transfer learning to demonstrate generalisability via testing the ML model in two unseen datasets from the construction, metal, and mining industries (Goldbeg 2022). The only mentioned limitation for the test is that a situation was misclassified as noncompliant with wearing protective equipment while it was a false negative alarm (Lu et al. 2023). The prediction model for tower crane procedures was also validated and one of the main challenges was found in the collection of on-site crane data, as some features were not observable (e.g., design errors, structure status, operators’ behaviour, and installation preparations sufficiency) (Jiang et al. 2023).

Fuzzy logic was applied either through fuzzy data mining for severity index and frequency rate rule-based expert systems (Ameri et al. 2017), or as a decision-making scheme, based on the ANNs predictions and expert opinions (Ayhan and Tokdemir, 2019). In this case, the decision categories depended on the predicted severity; if the decision-making scheme predicted a fatality, Ayhan and Tokdemir (2019) suggested that the construction should stop until an investigation is thoroughly done to eliminate the danger. Another model was suggested to predict risks for projects and individuals based on information about gender, age, experience, construction type, employment count, day of the week, and month that should be available from the construction sites and/or participating companies (Choi et al. 2020). Choi et al (2020) assumes that this information could be retrieved at the construction site entrance terminal through personal key cards and then Choi et al (2020) posits that predictions would automatically be done. Zhu et al. (2021) recommended using the RF prediction rules for accident severity prediction to assess occupational risks and prevent injuries. A usage suggestion for a RF model prediction was as a leading indicator for high-risk projects in the company (Poh et al. 2018). Baker et al. (2020a) proposed using severity and accident type classification by practitioners who do safety planning by identifying the task, tools or equipment, and working circumstances (Baker et al. 2020a), and by extracting safety precursors (Baker et al. 2020b). Another proposition was to use the classification of incident reports as part of a digital strategy to help managers extract information about near misses and increase awareness and learning on-site (Fang et al. 2020). Kang and Ryu (2019) proposed the ML model as a prediction model for accident types, but few ML models considered within the literature were developed to classify the accident reports with labels related to the accident causes (Zhang 2019, Zhang et al. 2019, Zhang et al. 2020, Zhong et al. 2020a). Shrestha et al. (2020) formulated the ML classification into a framework of upstream precursors linked to accident type, energy source, and severity. The latter factors are supposed to be solved accident hazards at the design phase.

The literature showcases promising suggested implementation for ML in analysing construction accident reports. However, only a few efforts propose a concise, tested and externally validated prototype. Most of these prototypes originated from a data mining model and less frequent classification and clustering. There is very promising usage proposals combining different accident attribute rules and advising safety instructions based on those rules. Classification of severity and accident types can be used in connection to a model's ranked predicates. There are common limitations such as more risk factors and consequences could be hidden within the text but not extracted which is related to several factors such as the limitations in manual labelling (Zhong et al. 2020a), defined extraction rules (Tang et al. 2021, Xu et al. 2021b), or the technology of NLP (Zhang et al. 2019, Xu et al. 2021b). Moreover, the complexity and constraints of the context for implementing ML-based analytical models should be considered together with the domain specific complexities associated with the construction industry – but this is scarcely done in the existing literature. For example, solving hazards at the design stage would probably influence the design process and the involved professionals. Moreover, from an application point of view, the framework of converting text classification and keyword extraction into actionable prevention strategies is a challenge, especially in systems of information retrieval or question-answer tasks (Pan et al. 2022). Otherwise, external validation gives valuable insights into metrics such as the applicability, comprehensiveness, and accuracy of the system, and the misclassification and availability of the relevant data.

4. DISCUSSION

This section analyses the literature review, especially in answering the research question:

What is the state of the art of research applying ML to accident report analysis in construction?

The analysis was structured in the same sequence of the paper's sections above. By analysing the literature, it can be noted that the data can be different in the structure (i.e., labelled or unlabelled), the degree of injury (near misses, accidents, fatality), or the data source (single or multiple companies, national registries). For data pre-processing, there are brief justifications for choosing the pre-processing methods. There is no doubt that the data is one of the legitimate limitations such as data being manually labelled or structured which leads to less volumes of data. There are suggestions for tackling manual labelling by using distance supervision as solution (Wu et al. 2022). Moreover, we have found that the data is confined by the missingness of factors outside of the organization such as governmental safety initiatives and regulations (Na et al. 2021). For that, data pre-processing which induces data loss might be not preferable for better safety severity prediction models (Koc and Gurgun, 2022). One proposal to improve data availability would be to add accident reports across multiple companies in comparable domains and even industries. Furthermore, the further introduction of digital systems in construction can create the potential to enrich available text data in larger volumes and integrate it with other sensory or imagery data (Baek et al. 2021, Yan et al. 2022, Ding et al. 2022). And, for smaller data sets, Ding et al. (2022) suggest the application of semi-supervised learning, few-shot learning, transfer learning, knowledge distillation, federated transfer learning, and model pre-training. It is also important to be mindful when using pre-structured data or predefined accident attributes, to not repeat the pattern of "what you look for is what you find" (Lundberg 2009). In a similar note, data being manually structured by some theory or framework, could also lead to a similar effect, unless the chosen method takes note of what lies outside the framework's boundaries. Nonetheless, the limited use of semi-structured or unstructured data reflects the limitations of analysing and understanding this data. However, we need to more experiments with this kind of data with unsupervised ML.

Another accident data limitation was the unbalanced characteristic. The critique of the notion of "unbalanced" when resampling lies in the implicit assumption that a phenomenon should generate balanced datasets, but this is not the case in the causes, types, and consequences of accidents. The different methods to balance the dataset imply that the ML designer moves into an unknown ground by assuming similarities in different parts of the studied phenomenon. Zhu et al. (2021) insight into the proneness of the SMOTE method to overfit indicate such consequences especially at the performance evaluation stage. The identified problems in data pre-processing are specific to the context of accidents in the construction industry and require further investigation, even so some evidence suggests that predicting fatality being sensitive to ROS, outliers, and data scaling (Koc and Gurgun, 2022). Moreover, in pre-processing, word embedding algorithms (BERT, Word2vec, GloVe, LTP) and pre-training with domain-related data (Baker et al. 2020b) stand out as variations of the chosen methods. Domain specific language is an essential characteristic that can contribute to advancing NLP feature extraction, so creating domain specific lexicons and dictionaries is a fundamental data pre-processing asset that should become more available. Such resources allow for applying a wider range of NLP methods such as relation extraction and knowledge graphs.

In the reviewed literature, many different algorithms have been adopted in studies with broadly the main purpose to enable improved prevention of accidents. The series of adopted algorithms is not consistently accompanied with proper justification of the selected algorithm but appears to be a result of experimentation. This approach and using a combination of algorithms is a shared property with other machine learning studies (e.g., Portugal, 2018). However, there is a tendency to focus less on prevention measures and more on accident factors like a prioritised situation or location (Halabi et al. 2022). This is coupled with challenges related to the data but also ones related to NLP and information extraction not being comprehensive in capturing all expressions. Not to forget to mention construction vocabulary lexicons and tailored algorithms for word embedding like in Tixier et al. (2016) play a crucial role. This might be a general observation in text-based research in the construction industry that generic NLP pre-processing tools do not capture construction-specific terminology unless ontologies of semantic meaning were identified by the knowledge of domain experts (Baek et al. 2021).

It is challenging to argue for a clear and consistent algorithm choice. But the review showed a promising application of deep learning algorithms (e.g., BERT, C-BiLSTM, SGRU, HAN) although the benefit of applying DL as opposed to other ML algorithms is yet to be established. We have seen previous authors raising concerns about uncertainty in the model or the data especially in DL (Zarei et al. 2023), but we also have seen approaches attempting to reduce that by applying DM in combination with DL and the use of fuzzy theory. Regarding ML, RF, XGboost and SVM show to be performing better than other algorithms. While DM and unsupervised learning is the type of learning that needs to be investigated further especially where it is less used (i.e. semi-structured and unstructured datasets). This might reflect the complexity of analysing text by NLP combined with data mining only. Moreover, there is much to be explored with text reasoning and advanced trends of Neural Module Networks combined with BERT and GPT (Ding et al. 2022). We would add to that the potentials of data analytics through large language models which exploded since late 2022.

By analysing the current literature on the application of ML-based analysis of accident reports within the construction industry context, it is found that several opportunities were presented. ML-based analytical models were suggested for prediction/classification, clustering, and information retrieval. Only six prototypes were presented (Kim and chi 2019, Ma et al. 2021, Tang et al. 2021, Lu et al. 2023, Amiri et al. 2017, Goldberg 2022, Jiang et al. 2023). While only four of them have been externally validated which makes room for spotting potential improvements or any functionality problems. Some of the limitations are misclassification, the availability of relevant data, labelling dependencies and the efficiency of NLP extraction. It is important to note that data availability is a major drawback, especially given the high share of small construction companies in the market. These smaller companies may need to rely on larger firms to invest in ML solutions. Nonetheless, larger companies could encourage smaller firms to adopt technological solutions if reliable and cost-effective use cases are established.

Other implications involve the utilization of safety recommendations that are extracted from the data. For example, the decision to stop construction to investigate a predicted accident (Ayhan and Tokdemir, 2019), or to make sure that the safety management-level is sufficient (Tang et al. 2021). To stop construction and investigate, is worth it even if it is estimated that there is a small chance to prevent a fatal accident but, in such case, an algorithm that allows for factors ranking is better preferred to for example DL. Furthermore, predictions involving worker's information might raise ethical concerns that need to be taken in consideration (Choi et al. 2020). Generally, the recommendation provided by knowledge extraction of task-safety-risk-factors is one important ML contribution. But to expect that the ML-based recommendations for safety as intuitive and easily implemented by practitioners, is not realistic. Elements of digitalization and change management research are options to evaluate the adoption of ML models. Thus, more external validation efforts should be taken in consideration for understanding how such ML powered systems can be integrated into existing safety procedures and functions.

So, in general we see low level of implementation of such ML models in construction. Even more, in hazard identification tasks, we see that accident reports data or safety tasks planning is not one of the current conversational AI applications in construction (Saka et al. 2023), which could be a future development for accident database analysis. So, the implementation of ML-based models would benefit from feasibility and implementation analysis, the involvement of practitioners, task definition and its expected use, together with listing all assumptions are crucial to the successful implementation of ML-based analytical models (Bilal and Oyedele 2020).

5. CONCLUSION AND FUTURE RESEARCH

This article has endeavoured to review the latest developments in research applying ML to accident report analysis in construction. The review is an in-depth analysis organized into the themes: data pre-processing, algorithms, testing and implementation. The analysis of the literature on applied ML on accident reports show that using ML-based models includes the classification of accident causes, accident type and consequences, prediction of severity, and extracting information from textual accident reports.

Regarding the data, there are limitations related to data availability, not covering broad safety causes in addition to difficulty combining other sources of data. Moreover, we found a limited utilisation of semi-structured and unstructured data sets while 10 out of reviewed articles manually structured the data which reflects challenges in feature extraction and could explain why authors opt out of using unsupervised learning. The ML development is also dependent on data pre-processing. There are challenges concerning the differentiated frequencies relating to accident type, severity and causes which are handled with techniques that implicitly assumes that the occupational accident phenomenon should generate balanced datasets. With the pre-processing step the outcome of the ML model is not always improved, and more comparative studies like (Koc and Gurgun, 2022) are needed to optimise the choices of this step. Moreover, it is essential that more effort is given to increasing availability of domain specific NLP resources for the distinctive accident records in construction.

In terms of algorithm choice, it is challenging to conclude which type of learning seems like the best approach in this domain, but DL, ML, and DM showed promising models. However, DM and unsupervised learning are less used with semi-structured and unstructured datasets which reflects the challenge of applying NLP combined with data mining only.

The targeted focus of this research has resulted in a limited number of papers to be reviewed, which can be seen as a limitation. However, six prototypes were found, involving the extraction of accident elements, risk assessment, and rule-based safety recommendations. While four of the presented efforts were externally validated. Such validation revealed issues of applicability, comprehensiveness, and accuracy as well as challenges of misclassification and availability of relevant data. Conceptual propositions were also presented as decision-making schemes when the ML-based models were aimed at predicting severe accidents which showed promises of improved prevention. Nevertheless, there are several factors related to the models and the characteristics of the accident reports that challenge the development of accurate and informative ML analysis. Most importantly, applying ML-based analytical models in the construction industry for accident prevention purposes depends upon a clear definition of the ML task, its intended use and associated ethical concerns. This type of analysis risks to lead to the extraction of accident causes and risks that are already known, as a result of pregiven framing and concepts. Moreover, ML algorithms like RF are better suited for understanding safety recommendations but regardless of the used algorithm it is generally so that information is still hidden within text and not efficiently extracted with NLP. The application of NLP is complicated by the need of labelling the extracted factors and because it often fails to capture all possible expressions added to that the domain knowledge terminology repositories constraints.

As the literature on ML analysis demonstrates, developing ML-based analytical models without careful feasibility study and the involvement of relevant stakeholders has a noticeable negative impact on the methodological choices of data pre-processing, algorithm choice, ML performance evaluation, and context constraints. While this is probably a general requirement that ML applied models in any domain should satisfy, but from a safety angle, this requirement is strictly necessary in the domain of construction. Moreover, it is not particularly established how safety professionals' benefit from the model and what implications learning from accident reports has on site safety. Applying ML in the construction industry, would benefit from incorporating a standardized development method that take in consideration practical aspects through explication of how ML-based recommendations for safety informs decision making, and through external validation.

Future research should focus on tackling the latter issues. The construction safety domain needs building domain specific NLP resources such as lexicons and word embedding algorithms. Future directions of text-based research should aim to create customised/domain specific NLP lexical ontologies for construction (Baek et al. 2021, Yan et al. 2022, Ding et al. 2022). Such developments might lead to a clarification of methodologies in terms of pre-processing and ML algorithm employment and the applying methods such as relation extraction and knowledge graphs. Thus far, DM and unsupervised learning are applied, but they could be explored further in future research.

Previous reviews of text-based research advocated information extraction can also be facilitated by knowledge representation learning and attention mechanisms for improving relation extraction, distant supervision instead of manual labelling, and rule-based or graph neural networks for knowledge-based development (Wu et al. 2022).

Research could also usefully carry out implementation studies on safety processes and management in construction companies. Of particular interest in such studies would be frameworks of integrated centralized computing platforms that are constantly fed with real time accident report data. From a methodological point of view, more experimental studies are needed to in order to draw firmer conclusions about the best ML methods to fit the domain specific context. Finally, an end-to-end text reasoning network based on Neural Module Networks or combined with BERT and GPT were proposed for interpretable text reasoning rather than just information extraction (Ding et al. 2022) and large language model present new analytic opportunities in thus domain.

REFERENCES

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, 162, 106102.

Amiri, M., Ardeshir, A., & Zarandi, M. H. F. (2017). Fuzzy probabilistic expert system for occupational hazard assessment in construction. Safety science, 93, 16-28.

Arumugam, R., & Shanmugamani, R. (2018). Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications. Packt Publishing Ltd.

Assaad, R., & El-adaway, I. H. (2021). Determining critical combinations of safety fatality causes using spectral clustering and computational data mining algorithms. Journal of Construction Engineering and Management, 147(5), 04021035.

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

Baek, S., Jung, W., & Han, S. H. (2021). A critical review of text-based research in construction: Data source, analysis method, and implications. Automation in Construction, 132, 103915.

Baker H., Hallowell M.R. and Tixier A.J.P. (2020a). AI-based prediction of independent construction safety outcomes from universal attributes. Automation in Construction, 118, p.103146.

Baker H., Hallowell M.R. and Tixier A.J.P. (2020b). Automatically learning construction injury precursors from text. Automation in Construction, 118, p.103145.

Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., ... & Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. Automation in Construction, 141, 104440.

Bell E., Bryman A., & Harley B. (2018). Business research methods. Oxford university press.

Bhagwat, K., & Delhi, V. S. K. (2022). A systematic review of construction safety research: quantitative and qualitative content analysis approach. Built Environment Project and Asset Management, 12(2), 243-261.

Bilal M., Oyedele L.O., Qadir J., Munir K., Ajayi S.O., Akinade O.O., ... and Pasha M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced engineering informatics, 30(3), 500-521.

Bilal M., and Oyedele L.O. (2020). Guidelines for applied machine learning in construction industry—A case of profit margins estimation. Advanced Engineering Informatics, 43, 101013.

Chen K. (2020). Enhancing construction safety management through edge computing: Framework and scenarios. Journal of Information Technology in Construction (iTcon), 25(25), 438-451.

Chen, Q., Tian, Z., Lei, T., & Huang, S. (2023). An association rule mining model for evaluating the potential correlation of construction cross operation risk. Engineering, Construction and Architectural Management, 30(10), 5109-5132.

Chen, Z., Huang, K., Wu, L., Zhong, Z., & Jiao, Z. (2022b). Relational graph convolutional network for text-mining-based accident causal classification. Applied Sciences, 12(5), 2482.

Cheng M.Y., Kusoemo D., and Gosno R.A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. Automation in Construction, 118, 103265.

Choi J., Gu B., Chin S., and Lee J.S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974.

Deng, D., Hou, Y., Zhang, J., & Wang, M. (2020, October). Analysis of Construction Accident Reports Based on C-BiLSTM Model. In 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 37-40). IEEE.

Deng, Y., Liu, Z., Song, L., Ni, G., & Xu, N. (2024). Exploring the metro construction accidents and causations for improving safety management based on data mining and network theory. Engineering, construction and architectural management, 31(9), 3508-3532.

Ding, Y., Ma, J., & Luo, X. (2022). Applications of natural language processing in construction. Automation in Construction, 136, 104169.

Dobrucali, E., Demirkesen, S., Sadikoglu, E., Zhang, C., & Damci, A. (2024). Investigating the impact of emerging technologies on construction safety performance. Engineering, Construction and Architectural Management, 31(3), 1322-1347.

Elghaish, F., Matarneh, S. T., & Alhusban, M. (2022). The application of “deep learning” in construction site management: Scientometric, thematic and critical analysis. Construction Innovation, 22(3), 580-603.

Fang W., Luo H., Xu S., Love P.E., Lu Z., and Ye C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. Advanced Engineering Informatics, 44, 101060.

Garcia, J., Villavicencio, G., Altimiras, F., Crawford, B., Soto, R., Minatogawa, V., ... & Yepes, V. (2022). Machine learning techniques applied to construction: A hybrid bibliometric analysis of advances and future directions. Automation in Construction, 142, 104532.

Gholizadeh P., Esmaeili B., and Memarian B. (2018, April). Evaluating the performance of machine learning algorithms on construction accidents: An application of ROC curves. In Proceedings of the Construction Research Congress (CRC 2018), ASCE, New Orleans, LA, USA (pp. 2-4).

Gholizadeh, P., Onuchukwu, I. S., & Esmaeili, B. (2021). Trends in catastrophic occupational incidents among electrical contractors, 2007–2013. International Journal of Environmental Research and Public Health, 18(10), 5126.

Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., and Kagal L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

Goh, Y. M., & Ubeynarayana, C. U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. Accident Analysis & Prevention, 108, 122-130.

Goldberg, D. M. (2022). Characterizing accident narratives with word embeddings: Improving accuracy, richness, and generalizability. Journal of safety research, 80, 441-455.

Gopal, M. (2019). Applied machine learning. McGraw-Hill Education.

Guo, S., Zhao, Y., Luoren, Y., Liang, K., & Tang, B. (2022). Knowledge discovery of correlations between unsafe behaviors within construction accidents. Engineering, Construction and Architectural Management, 29(4), 1797-1816.

Gupta, A. K., Pardheev, C. G. V. S., Choudhuri, S., Das, S., Garg, A., & Maiti, J. (2022). A novel classification approach based on context connotative network (CCNet): A case of construction site accidents. Expert Systems with Applications, 202, 117281.

Halabi, Y., Xu, H., Long, D., Chen, Y., Yu, Z., Alhaek, F., & Alhaddad, W. (2022). Causal factors and risk assessment of fall accidents in the US construction industry: A comprehensive data analysis (2000–2020). Safety science, 146, 105537.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

Hegde J., and Rokseth B. (2020). Applications of machine learning methods for engineering risk assessment–A review. Safety science, 122, 104492.

Ho, S. Y., Phua, K., Wong, L., & Goh, W. W. B. (2020). Extensions of the external validation for checking learned model interpretability and generalizability. Patterns, 1(8), 100129.

Hoła B., and Szóstak M. (2015). Analysis of the state of the accident rate in the construction industry in European Union countries. Archives of Civil Engineering, 61(4).

Hou L., Chen H., Zhang G.K., and Wang X. (2021). Deep Learning-Based Applications for Safety Management in the AEC Industry: A Review. Applied Sciences, 11(2), 821.

Jang, Y., Kim, K., Leite, F., Ayer, S., & Cho, Y. K. (2021). Identifying the perception differences of emerging construction-related technologies between industry and academia to enable high levels of collaboration. Journal of Construction Engineering and Management, 147(10), 06021004.

Jiang, L., Zhao, T., Feng, C., & Zhang, W. (2023). Improvement of random forest by multiple imputation applied to tower crane accident prediction with missing data. Engineering, construction and architectural management, 30(3), 1222-1242.

Kang K., and Ryu H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120, 226-236.

Khallaf, R., & Khallaf, M. (2021). Classification and analysis of deep learning applications in construction: A systematic literature review. Automation in construction, 129, 103760.

Kim T., and Chi S. (2019). Accident case retrieval and analyses: using natural language processing in the construction industry. Journal of Construction Engineering and Management, 145(3), 04019004.

Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2021). Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. Automation in Construction, 131, 103896.

Kang, K. S., Koo, C., & Ryu, H. G. (2022). An interpretable machine learning approach for evaluating the feature importance affecting lost workdays at construction sites. Journal of Building Engineering, 53, 104534.

Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2022). Accident prediction in construction using hybrid wavelet-machine learning. Automation in Construction, 133, 103987.

Koc, K., & Gurgun, A. P. (2022). Scenario-based automated data preprocessing to predict severity of construction accidents. Automation in Construction, 140, 104351.

Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2022a). Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Engineering, Construction and Architectural Management, 30(9), 4486-4517.

Lee, J. Y., Yoon, Y. G., Oh, T. K., Park, S., & Ryu, S. I. (2020). A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry. Applied Sciences, 10(21), 7949.

Liang, H., & Shi, X. (2022). Exploring the structure and emerging trends of construction health management: a bibliometric review and content analysis. Engineering, Construction and Architectural Management, 29(4), 1861-1889.

Li, X., Zhu, R., Ye, H., Jiang, C., & Benslimane, A. (2021). MetaInjury: Meta-learning framework for reusing the risk knowledge of different construction accidents. Safety science, 140, 105315.

Liu, J., Luo, H., & Liu, H. (2022). Deep learning-based data analytics for safety in construction. Automation in construction, 140, 104302.

Lundberg, J., Rollenhagen, C., & Hollnagel, E. (2009). What-You-Look-For-Is-What-You-Find–The consequences of underlying accident models in eight accident investigation manuals. Safety science, 47(10), 1297-1311.

Luo, X., Li, X., Goh, Y. M., Song, X., & Liu, Q. (2023). Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents. Safety Science, 163, 106138.

Lu, Y., Yin, L., Deng, Y., Wu, G., & Li, C. (2023). Using cased based reasoning for automated safety risk management in construction industry. Safety Science, 163, 106113.

Ma G., Wu Z., Jia J., and Shang S. (2021). Safety risk factors comprehensive analysis for construction project: Combined cascading effect and machine learning approach. Safety science, 143, 105410.

MacLure* M. (2005). 'Clarity bordering on stupidity': where's the quality in systematic review? Journal of Education Policy, 20(4), 393-416.

Milch V., and Laumann K. (2016). Interorganizational complexity and organizational accident risk: A literature review. Safety science, 82, 9-17.

Morais, C., Yung, K. L., Johnson, K., Moura, R., Beer, M., & Patelli, E. (2022). Identification of human errors and influencing factors: A machine learning approach. Safety science, 146, 105528.

Na, X. U., Ling, M. A., Liu, Q., Li, W. A. N. G., & Deng, Y. (2021). An improved text mining approach to extract safety risk factors from construction accident reports. Safety science, 138, 105216.

Nnaji, C., Awolusi, I., & Gambatese, J. (2023). A systematic review of technology acceptance models and theories in construction research. Journal of Information Technology in Construction (ITcon), 28(3), 39-69.

Pan, X., Zhong, B., Wang, Y., & Shen, L. (2022). Identification of accident-injury type and bodypart factors from construction accident reports: A graph-based deep learning framework. Advanced Engineering Informatics, 54, 101752.

Poh C.Q., Ubeynarayana C.U., and Goh Y.M. (2018). Safety leading indicators for construction sites: A machine learning approach. Automation in construction, 93, 375-386.

Rafindadi, A. D. U., Shafiq, N., Othman, I., Ibrahim, A., Aliyu, M. M., Mikić, M., & Alarifi, H. (2023). Data mining of the essential causes of different types of fatal construction accidents. Heliyon, 9(2).

Regona, M., Yigitcanlar, T., Xia, B., & Li, R. Y. M. (2022). Opportunities and adoption challenges of AI in the construction industry: A PRISMA review. Journal of open innovation: technology, market, and complexity, 8(1), 45.

Riccio V., Jahangirova G., Stocco A., Humbatova N., Weiss M., and Tonella P. (2020). Testing machine learning based systems: a systematic mapping. Empirical Software Engineering, 25(6), 5193-5254.

Saka, A. B., Oyedele, L. O., Akanbi, L. A., Ganiyu, S. A., Chan, D. W., & Bello, S. A. (2023). Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities. Advanced Engineering Informatics, 55, 101869.

Sabet, M. F. A., Dahroug, A., & Hegazy, A. F. (2021, December). A Proposed model for field workers Injuries' prevention based on machine learning. In 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 383-388). IEEE.

Shao, L., Guo, S., Dong, Y., Niu, H., & Zhang, P. (2023). Cause analysis of construction collapse accidents using association rule mining. Engineering, construction and architectural management, 30(9), 4120-4142.

Shehab N., Badawy M., and Arafat H. (2021). Big Data Analytics and Preprocessing. In Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges (pp. 25-43). Springer, Cham.

Shrestha S., Morshed S.A., Pradhananga N., and Lv X. (2020, November). Leveraging Accident Investigation Reports as Leading Indicators of Construction Safety Using Text Classification. In Construction Research Congress 2020: Safety, Workforce, and Education (pp. 490-498). Reston, VA: American Society of Civil Engineers.

Solanki, A., Kumar, S., & Nayyar, A. (Eds.). (2019). Handbook of Research on Emerging Trends and Applications of Machine Learning.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 843-852).

Tang N., Hu H., and Xu F. (2021, January). Personalized Safety Instruction Recommendation for construction workers Based on Apriori Algorithm. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1197-1203). IEEE.

Tixier, A. J. P., & Hallowell, M. R. (2023). Safer Together: Machine Learning Models Trained on Shared Accident Datasets Predict Construction Injuries Better than Company-Specific Models. arXiv preprint arXiv:2301.03567.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2017). Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining. Automation in Construction, 74, 39-54.

Vallmuur K. (2015). Machine learning approaches to analysing textual injury surveillance data: a systematic review. Accident Analysis & Prevention, 79, 41-49.

Webster J., and Watson R.T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, xiii-xxiii.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. Automation in Construction, 134, 104059.

Xu, X., & Zou, P. X. (2021). Discovery of new safety knowledge from mining large injury dataset in construction. Safety science, 144, 105481.

Xu Y., Zhou Y., Sekula P., and Ding L. (2021a). Machine learning in construction: From shallow to deep learning. Developments in the Built Environment, 100045.

Xu N., Ma L., Wang L., Deng Y., and Ni G. (2021b). Extracting domain knowledge elements of construction safety management: Rule-based approach using Chinese natural language processing. Journal of Management in Engineering, 37(2), 04021001.

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction, 119, 103331.

Yan, H., Ma, M., Wu, Y., Fan, H., & Dong, C. (2022). Overview and analysis of the text mining applications in the construction industry. Heliyon, 8(12).

Zarei, E., Khan, F., & Abbassi, R. (2023). How to account artificial intelligence in human factor analysis of complex systems?. Process safety and environmental protection, 171, 736-750.

Zhang F. (2019). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. International Journal of Construction Management, 1-21.

Zhang F., Fleyeh H., Wang X., and Lu M. (2019). Construction site accident analysis using text mining and natural language processing techniques. Automation in Construction, 99, 238-248.

Zhang J., Zi L., Hou Y., Deng D., Jiang W., and Wang M. (2020). A C-BiLSTM Approach to Classify Construction Accident Reports. Applied Sciences, 10(17), 5754.

Zermane, A., Tohir, M. Z. M., Zermane, H., Baharudin, M. R., & Yusoff, H. M. (2023). Predicting fatal fall from heights accidents using random forest classification machine learning model. Safety Science, 159, 106023.

Zhong B., Pan X., Love P.E., Ding L., and Fang W. (2020a). Deep learning and network analysis: Classifying and visualizing accident narratives in construction. Automation in Construction, 113, 103089.

Zhu R., Hu X., Hou J., and Li X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety and Environmental Protection, 145, 293-302.