

Predicting Occupational Accident Risk from Textual Data: A Systematic Review of Machine Learning Application

Afrigh Fajar Rosyidiin^{1,2,*}, Moses Laksono Singgih^{1,*}, Adithya Sudiarno¹

¹Department of Industrial and Systems Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

²Department of Safety Engineering, Politeknik Perkapalan Negeri Surabaya, Indonesia

*Author to whom correspondence should be addressed:

E-mail: 7010231004@student.its.ac.id, moseslinggih@its.ac.id

(Received June 05, 2025; Revised December 23, 2025; Accepted March 10, 2026)

Abstract: Occupational accident risk remains a persistent concern, especially in high-risk industries. However, the optimal use of textual data in safety analysis is still limited. This study aims to systematically review the application of machine learning (ML) and natural language processing (NLP) in predicting occupational accident risk using textual data. A systematic literature review was conducted using the PRISMA framework and PICOS criteria, with Scopus as the primary database. From 1238 initial articles, 21 were selected for in-depth analysis. The results indicate that algorithms such as Random Forest, Support Vector Machine, and Neural Networks are frequently used, achieving up to 94% accuracy when integrated with NLP techniques. This review highlights the growing potential of ML and NLP in developing proactive and data-driven occupational safety strategies. Further research is recommended to address limitations in data quality, model generalization and validation across various sectors.

Keywords: machine learning; natural language processing; occupational accident; predicting; risk management

1. Introduction

Occupational accidents remain a significant challenge in occupational Safety and health management in various sectors, especially in industries that have a high level of risk such as the construction and energy sector¹. According to the International Labour Organization (ILO), more than 2.78 million workers lose their lives each year due to occupational accidents or occupational diseases, with about 374 million cases of non-fatal injuries reported each year². The role of company management is to increase self-awareness that can be applied in occupational safety training to increase worker compliance and participation³. In addition, Systematic Analysis of the causes of accidents is necessary to reduce risks and improve Safety in the workplace⁴. With the advancement of the digital age, a large amount of textual data has been generated from accident reporting, including incident descriptions, investigation reports and worker feedback⁵. This dataset contains essential information that can be processed to identify risk patterns⁶. However, the optimal use of textual data remains limited. Traditional approaches that rely solely on manual Analysis are often ineffective, especially when faced with large amounts of data⁷. Conventional approaches to risk prediction in Safety

Management Systems generally rely on manual methods such as periodic inspections, checklists, and experience-based Analysis such as HAZOP or FMEA⁸. In addition, the results of the Analysis are prone to human bias and inconsistencies in risk assessment. This manual process also takes longer to process information, so delays in decision-making can increase the potential for accidents at work⁹.

With the increasing need for data-driven work safety systems, digital transformation provides excellent potential for faster and more accurate decision-making¹⁰. One of the biggest challenges is the utilization of narrative data generated from work accident reports, which are often unstructured and challenging to analyze manually¹¹. Natural Language Processing (NLP) comes as an intelligent approach to extracting hidden information from text reports, such as the cause of an accident or a recurring risk pattern¹². Previous studies have shown that the combination of NLP and Machine Learning (ML) can improve prediction accuracy and significantly reduce analysis time¹³. This is in line with previous researchers' findings that emphasized the role of NLP in compiling information from incident reports to improve safety management systems¹⁴. Moreover, This approach is considered to be more adaptive in complex and dynamic

work environments, such as construction and manufacturing^{15,16}. However, the adoption of NLP in the context of occupational Safety is still limited, so a systematic study is needed to evaluate its effectiveness and implementation challenges^{17,18}.

Despite the growing availability of textual data from accident reports, the ability to harness this data for predictive analytics remains underexplored. Existing methods often rely on structured data or conventional checklists, failing to capture narrative insights embedded in unstructured texts. The key problems identified are: (1) limited adoption of ML-NLP frameworks in occupational safety domains, (2) lack of clarity regarding which ML algorithms are most effective for text-based prediction, and (3) absence of systematic evidence mapping current trends and gaps.

Previous research has used conventional statistical methods to identify factors that cause work accidents¹⁹. However, this approach is often incapable of capturing the complexity of patterns hidden in textual data²⁰. In recent years, advanced technology has been used to improve Safety and efficiency in the work environment²¹. The application of machine learning (ML) has shown great potential in the Analysis of unstructured data, including text data²². Therefore, several studies highlight the importance of text data processing and classification algorithms in the context of Safety, which can be adapted for occupational safety analysis²³. Previous research has addressed the use of machine learning (ML) for risk classification and accident prediction, but research specifically exploring textual data in the context of occupational Safety is still limited²⁴. ML can be applied in the context of Safety and regulatory compliance²⁵. The application of automation and ML technologies in improving Safety in high-risk work environments²⁶. In addition, most related studies used local datasets, making the results less applicable to various industries²⁷. The lack of systematic studies on the application of machine learning (ML) to textual data to predict the risk of occupational accidents underscores the importance of this study²⁸. The main contributions of this study are as follows:

- Conducts a systematic review of ML-based approaches in analyzing textual occupational safety data.
- Identifies dominant algorithms and techniques (e.g., RF, SVM, NN, NLP) used in accident risk prediction.
- Maps research gaps in current literature and proposes directions for future studies.
- Presents a comparative overview of algorithm performance and integration trends.

This study aims to systematically review the application of machine learning (ML) and natural language processing (NLP) techniques for predicting occupational accident risk

using textual data. Specifically, this review seeks to (1) identify and categorize dominant ML and NLP-based approaches applied to unstructured safety narratives, (2) analyze performance trends and comparative effectiveness of commonly used algorithms, and (3) highlight methodological limitations, research gaps, and future opportunities in text-based accident risk prediction. By synthesizing recent empirical evidence, this study provides both theoretical insights and practical recommendations to support the development of proactive, data-driven occupational safety management systems.

While prior studies have applied ML to structured occupational safety data, limited attention has been paid to unstructured textual data such as incident narratives, safety reports, and feedback logs. Moreover, few studies have systematically examined the effectiveness and integration of NLP with ML for risk prediction in this context. This gap presents a critical opportunity, as unstructured data often contains rich contextual insights not captured in numerical datasets. Therefore, this study seeks to bridge this gap by systematically reviewing current research and providing recommendations for improving text-based accident prediction frameworks.

The remainder of this paper is organized as follows: Section 2 presents the literature review on machine learning and occupational safety. Section 3 explains the research methodology. Section 4 describes the results and analysis. Section 5 discusses the findings and implications. Finally, Section 6 concludes the study and offers directions for future research.

2. Literature Review

This section presents a literature review that aims to build a theoretical and conceptual foundation for research related to Machine Learning-based (ML)-based work accident risk prediction. The review was conducted on three main interrelated themes: (1) employee involvement in safety practices, (2) the application of ML in the context of occupational Safety, and (3) the use of ML in the prediction of work accident risk. Each subsection discusses the contribution of previous literature in identifying key factors influencing safety behavior, the development of machine learning (ML) and natural language processing (NLP) technologies in analyzing work accident data, and the effectiveness of predictive models in identifying risk patterns. This approach enables a comprehensive mapping of the potential and challenges associated with applying intelligent technology to support data-driven occupational safety management.

2.1. Safety Engagement in Workplace

Safety engagement is an important element in encouraging safe and productive work behavior²⁹. Factors that support this engagement involve organizational dimensions such as safety culture, trust between members, and managerial

leadership that motivates active worker participation³⁰. Safety engagement involves compliance with safety policies and active participation in safety-oriented activities, ultimately fostering a positive safety culture³¹. In addition, individual psychological factors, such as motivation and attitudes towards Safety, have a significant influence on safety behavior in the workplace³². Safety education plays a major role in increasing motivation and compliance with safety rules³³. Experience-based education programs, such as practical training and sharing of workplace incident experiences, have proven effective in increasing worker awareness³⁴. Well-designed safety training can increase worker engagement by strengthening their understanding of the importance of safety³⁵. In addition, transformational leadership that facilitates open communication between managers and workers has been shown to reduce unsafe behaviors and improve an organization's safety culture³⁶. Behavior-Based Safety (BBS), as a behavior-based approach, has also shown positive results in increasing worker participation and commitment to safety³⁷.

2.2. Machine Learning for Safety

Prior to 2019, Machine Learning (ML) research in text-based occupational accident analysis was still limited to Bayesian methods and decision tree-based models to identify accident patterns³⁸. The main focus is on the visualization of incident patterns, but the limitations of data quality, model generalization, and text processing hinder proactive prediction³⁹. Integration with Natural Language Processing (NLP) is still minimal, so text analysis is not optimal. After 2019, the development of deep learning and NLP enables real-time prediction and automation of risk analysis, improving the effectiveness of occupational safety management⁴⁰.

Machine Learning (ML) applied in occupational safety has opened up opportunities to proactively identify and prevent potential hazards in the workplace⁴¹. ML enables the Analysis of complex work accident data, both through structured and unstructured data such as accident narrative reports⁴². One example of a technology-based early warning system is the Safety Warning Detector (SWAD), which provides a direct signal to toll road workers to avoid hazards while working on emergency lanes, significantly improving work safety⁴³. Machine learning (ML) can improve efficiency and accuracy in technical systems, which can be applied in the development of occupational safety systems⁴⁴. ML can also predict wind speeds which are important in the context of renewable energy and operational Safety. The methodology used can be adapted for risk prediction in high-risk work environments⁴⁵. ML applications in environmental monitoring that can be adapted for occupational safety supervision in industrial sites and able to manage waste by analyzing factors that affect a sustainable electronic waste management

system^{46,47}. Moreover, it can be used for occupational safety monitoring in hard-to-reach areas⁴⁸. For example, algorithms such as Random Forest and Neural Networks have been used to predict potential accidents based on historical patterns, improving the accuracy and efficiency of risk detection⁴⁹. In addition, Natural Language Processing (NLP)-based approaches are able to extract valuable information from accident narrative reports to understand root causes hidden in textual data, which are often overlooked in traditional methods⁵⁰.

Machine Learning (ML) has been widely applied in the context of reliability and Safety to predict system failures and improve risk management⁵¹. Models such as Support Vector Regression (SVR) are superior in prediction accuracy to traditional methods and are often used for Remaining Useful Life (RUL) estimation as well as system degradation projections, although they require prior knowledge of kernel selection⁵². Neural Network models including Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) provide high accuracy in analyzing complex and non-linear systems, such as RUL predictions on aircraft engines and component degradation rate analysis, despite the high computational demands^{53,54}. In addition, approaches such as Gaussian Process Regression (GPR) are used for time-based failure probability analysis and system reliability analysis with the ability to dynamically update model parameters, although limited to lower computational scalability when dealing with large datasets⁵⁵. The use of this method demonstrates ML ability to provide predictive solutions in complex, data-driven safety domains⁵⁶

2.3. Risk Prediction with Machine Learning

The use of Machine Learning (ML) in work accident analysis has great potential in predicting risks with high accuracy through large-scale data processing⁵⁷. ML models such as Decision Tree (DT), Support Vector Machine (SVM) and Neural Networks have been used effectively to predict accident outcomes, including injury severity and identification of primary causative factors⁵⁸. Based on research that has been conducted, SVM-based algorithms optimized with Particle Swarm Optimization (PSO) show an accuracy of up to 90.67%, with the ability to identify decision rules related to the root cause of accidents⁵⁹. Additionally, a text-based approach using Natural Language Processing (NLP) enables the extraction of hidden patterns from narrative-based accident reports, thereby expanding predictive capabilities in complex work environments⁶⁰. NLP has the potential to dig deep insights from safety report narratives⁶¹. Flexible machine learning models can address data imbalances, thereby improving asymmetric class distributions in risk prediction⁶².

The implementation of machine learning (ML) in occupational accident risk prediction has been applied in various industries to strengthen safety policies and inform

data-driven decision-making⁶³). ML can predict the wear and tear of industrial tools that can be adapted to improve work safety through predictive maintenance and vibration analysis to fault predict^{64,65}). Cloud detection algorithms to support weather forecasts relevant in the context of occupational Safety in industries affected by weather conditions, such as construction and energy⁶⁶). ML-based models not only predict accident risk but also provide insights into causal patterns that may have been missed in traditional analysis⁶⁷). ML predictive models can be used in risk management and Safety of the work environment⁶⁸). In addition, the combination of CNN and LSTM is capable of weather forecasting, which is essential in occupational safety planning in the energy and construction sectors⁶⁹). Nonetheless, challenges such as the need for high-quality data, complex data processing and the limitations of external validation of models remain major obstacles⁷⁰). Recent research recommends the integration of deep learning methods and the development of data-driven decision support systems to overcome these constraints and improve the efficiency of occupational accident risk prediction⁷¹). However, few studies have specifically addressed sign language datasets or other domain-specific textual modalities. For instance, the work by Golebiewski and Eikeland introduced a dedicated dataset for Norwegian Sign Language recognition. This indicates opportunities for cross-context adaptation of ML-NLP pipelines in other safety-critical textual domains, including accident reports⁷²).

3. Method

First stage in making a Systematic Literature Review (SLR) is to determine the research question by establishing a research framework using the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) approach⁷³). This framework helps guide the research toward specific and relevant questions. After establishing the research framework, the researcher searches for articles with queries designed based on the PICOS framework⁷⁴). The second stage involves finding articles using the electronic Scopus database. Some of the reasons why researchers take literature from journals indexed in Scopus are that the quality of publication is high because article registered in Scopus go through a strict selection process before being indexed, so this journal can be trusted and considered high quality. To get the best results, it is recommended to conduct and evaluate the literature review effectively⁷⁵). International visibility with the Scopus journal contains articles from various countries. Articles published in the journal Scopus are recognized internationally. This can improve the author's achievements and provide recognition in the scientific field. Search for articles by making a search query based on the results of PICOS Analysis and research questions.

The following are search queries obtained in this way involving keywords such as "Natural Language Processing," "Machine Learning," "predicting occupational accidents," and "textual data."

In the third stage of article selection, data were obtained from the Scopus database with filters that include publication year (2019-2024), type of document (article), subject (engineering, computer science, and environmental science), and open access. The time range 2019–2024 was selected to capture recent advancements in the integration of NLP and ML, as most notable developments in deep learning techniques and open-access safety datasets have occurred during this period. The selection stage using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach includes screening the title and abstract, followed by a full-text review to ensure the relevance of the article to the research objectives⁷⁶). Out of the initial 1238 articles, the selection process resulted in 21 relevant articles for further Analysis. Data extraction is the next stage after the article selection process. A total of 21 articles that are declared relevant, Articles are reviewed and summarized in the form of tables. Each journal article is reviewed from Author, Year, Research Focus, Population, Methodology, Machine Learning Approach, Dataset, Research Limitations. The next stage is article mapping and Analysis. The mapping of the article includes bibliometric networks, distribution of publication by years, distribution of location, machine learning application, distribution of ML algorithms, integration of ML Algorithms and success rates. The final stage, there will be a discussion of the results of the SLR research. This stage will discuss the interpretation of research results, contributes to developing safety strategies applicable to the industrial sector and research limitations and future research recommendations⁷⁷). From the method that has been presented Showing the importance of a systematic approach in reviewing the application of ML in Safety Management⁷⁸).

Although only 21 articles were included in the final analysis, this number reflects the stringent selection criteria applied to ensure relevance and methodological rigor. Articles were filtered based on language (English), database (Scopus), accessibility (Open Access), relevance to both ML and NLP, and direct applicability to occupational accident prediction using textual data. Studies focusing on structured data or lacking predictive modeling were excluded. This focused approach ensured that the findings of this review are grounded in high-quality and contextually relevant research, even if the generalizability is constrained by sample size.

4. Result

In the scope of research, literature review is a very important source of reference. Through literature review,

researchers can obtain valid information and data which are then used as a reference to make scientific papers. Therefore, The literature review is an essential initial step before writing a scientific paper. To understand and describe the current state of research in the field of smart safety management, researchers have conducted a literature review with several detailed stages to gain knowledge related to the development of research on the implementation of machine learning for accident risk prediction.

4.1. Research Question

In determining the research question, the researcher uses the PICOS framework which is expected to represent several factors for the preparation of the research. PICOS stands for Population, Intervention, Comparison, Outcome, and Study Design. This method is not only important in the world of research, but also a solid foundation in optimizing the effectiveness and accuracy of research⁷⁹). A more detailed explanation of the elements in the PICOS framework is as follows:

- 1) *Population*: Population refers to the individual or group that is the focus of the research. The right selection of the population is the main key to the success of the research. By clearly defining the population, researchers can identify specific characteristics, such as age, sex, or other health conditions, that may influence the results of the study.
- 2) *Intervention*: Interventions that refer to actions or treatments given to Populations.
- 3) *Comparison*: Treatment or condition that is compared to the Intervention. The selection of appropriate comparators allowed researchers to more accurately evaluate whether the interventions tested were more effective than other alternatives. By having a comparison group, researchers can measure and compare the effectiveness of the two types of treatment.
- 4) *Outcome*: Expected results of the study. Outcomes include parameters or variables that are measured to evaluate the effectiveness of the Intervention.
- 5) *Study Design*: A series of procedures and working methods that are structured to systematically connect each element of research. Have the goal of analyzing and determining the focus of research effectively and efficiently.

Table 1 is the results of PICOS Analysis based on the problems in the field that have been described above and by utilizing the data available in the field, The researcher formulated the research focus based on:

"Safety report analysis in the form of textual data for prediction of work accidents using Natural Language Processing (NLP) and machine learning"

Table 1 is the result of PICOS Analysis based on the

Table 1: Result of PICOS Analysis

PICOS	Answer
Population (P)	Workers who report occupational incidents through structured and unstructured textual data
Intervention (I)	Implementation of NLP techniques (e.g., tokenization, TF-IDF, embeddings) combined with ML algorithms for text-based prediction
Comparison (C)	Traditional methods such as manual classification, rule-based systems, or keyword matching
Outcome (O)	Model accuracy, interpretability, speed of analysis, and impact on safety outcomes
Study Design (S)	Systematic Literature Review (SLR) with comparative synthesis.

research idea. From the results of filling out the PICOS framework in accordance with the topic of work accident risk prediction using a machine learning approach based on textual data, the following research questions were obtained:

RQ: How effective is the use of natural language processing combined with machine learning to predict the risk of work accidents from textual data?

4.2. Finding Article

Journal article searches were conducted using the Scopus electronic database. The search for journal articles to match the topic of SLR is Predicting Occupational Accident Risk from Textual Data. The research employed a strategy by constructing a search query based on the results of PICOS Analysis and research questions. The search query is used to search for journals in the Scopus database using the following boolean operators:

("natural language processing" OR NLP) AND ("machine learning") AND ("predicting" OR "prediction") AND ("occupational accident" OR "workplace accident" OR "occupational injury" OR "workplace injury") AND ("textual data" OR "text data" OR "unstructured data")

4.3. Article Selection

Article selection stage use the Preferred Reporting Items for Systematic Reviews (PRISMA) work tool, which is a set of evidence-based items that aim to help scientific authors report various types of systematic literature reviews and meta-analyses transparently and effectively. The selection criteria adjust the needs of the research to determine the latest and unique research. Figure 1 is a PRISMA from the literature review that the researcher has conducted.

Based on Figure 1 Preferred Reporting Items for Systematic Reviews (PRISMA), it can be concluded that a total of 21 articles that will be reviewed in full text at the next stage. To find out more specifically the screening criteria in article selection, find to Table 2.

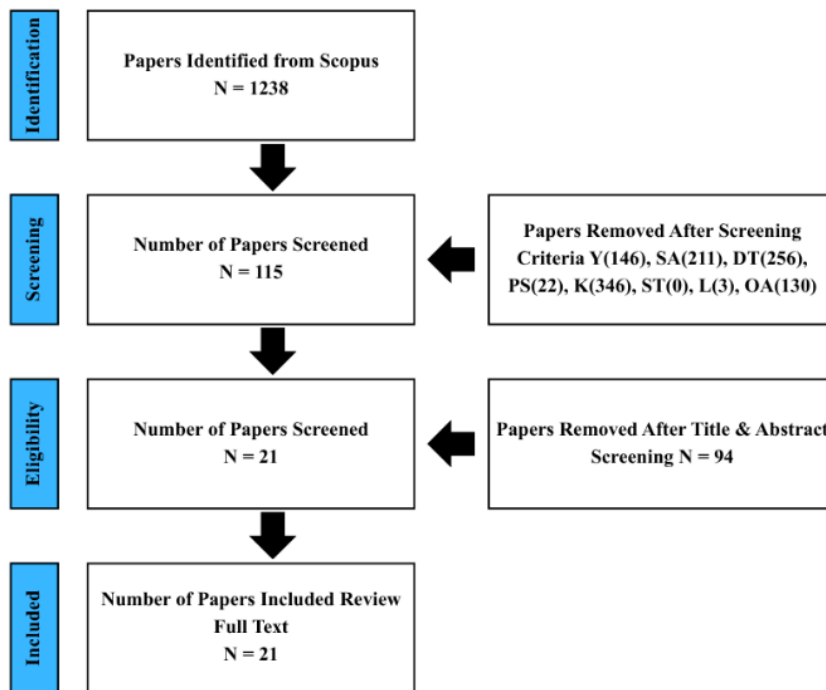


Fig. 1: Article Selection with PRISMA Approach

Table 2: Screening Criteria in Article Selection

	Filter	Article Qty
Search Query	("machine learning") AND ("predicting" OR "prediction") AND ("occupational accident" OR "workplace accident" OR "occupational injury" OR "workplace injury")	1238
Year	2019 - 2024	1092
Subject Area	Engineering, Computer Science, Environmental Science, Decision Science	881
Document Type	Article	616
Publication Stage	Final	594
Keyword	Machine Learning, Accident, Occupational Risks, Prediction, Occupational Accident	248
Source Type	Journal	248
Language	English	245
Open Access	All Open Access	115
Title and Abstract Screening	Result (Accident Prediction), Method (ML & NLP) and Data (Text)	21
Full Text Review	Relevant with this research	21

Table 2 is a filter on the literature review on machine learning and occupational Safety for accident prediction. Filter specifications are adjusted to the needs of researchers with the latest and appropriate journals. The addition of filters is carried out directly on the Scopus website to make it easier and more practical.

The article selection process, as outlined in Table 2, applied rigorous filtering to ensure that only the most relevant and high-quality studies were included in the final analysis. From an initial pool of 1238 articles, the review applied sequential filters—by publication year (2019–2024), subject area, document type, publication stage, and keyword relevance—to narrow the dataset to 21 articles. This structured screening demonstrates a balance between

comprehensiveness and specificity, emphasizing studies that explicitly involve both machine learning and natural language processing for accident prediction using textual data. The inclusion of only open-access articles written in English further ensured accessibility and reproducibility, though it may have excluded some potentially relevant studies from other repositories or languages. Overall, the criteria reflect a strong methodological foundation aligned with PRISMA standards.

4.4. Data Extraction

The data extraction stage was conducted following the identification of 21 eligible studies through the PRISMA selection process. This stage aimed to systematically

collect and synthesize key attributes of each study to allow for structured comparison and analysis. Each selected article was reviewed in detail to capture relevant aspects such as machine learning models applied, the type of data used, prediction accuracy, and any specific algorithmic configurations. This process ensured consistency and depth in interpreting how textual data has been leveraged across studies to predict occupational accident risks. To increase clarity and comparability, the results of the extraction process were summarized in a comprehensive table format (Table 3). The table functions as both a summary of the reviewed literature and a basis for cross-study benchmarking. This approach not only facilitates transparency but also responds to reviewer requests regarding performance metrics, methodological rigor, and parameter documentation.

Each column in Table 3 is designed to capture specific dimensions of the reviewed studies. The "Author (Year)" column identifies the source. The "ML Algorithm(s)" column lists all machine learning techniques implemented in the study. "Dataset Type" describes the origin or domain of the dataset (e.g., OSHA reports, narrative logs, questionnaires), while "Feature Type" reflects whether the inputs were textual, structured, or multimodal. The "Parameter" column records known algorithm configurations (e.g., tree depth, kernel type), and "Accuracy (%)" presents the reported predictive performance of the model as an outcome metric. To provide a brief overview of data extraction which is a summary of 21 articles discussing Machine Learning (ML) approaches in predicting occupational safety risks, it can be seen in table 3.

Table 3: Article Classification

No	Author	ML Approach	Dataset	Feature Type	Parameter	Accuracy
1	80)	RF, LR, DT, ANN, AdaBoost, SGTB	Questionnaire	Structured	RF: default; DT: max_depth=5	89.4%
2	81)	XGBoost, Linear SVM	Textual Reports	Narrative Text	SVM: kernel=linear	~59%
3	82)	RF, DT, ANN	Workplace Records	Textual	RF: n=100	90%
4	83)	RF, LR, DT, SVM	Construction Reports	Text	DT: depth=8	79,3%
5	84)	RF, DT, KNN, NB, SVM	OSHA Records	Textual	RF: n=50; ANN: 2 layers	89.4%
6	85)	RF, DT, ANN, SVM	Construction Data	Text	RF: n=150	91%
7	86)	RF, LR, SVM, NLP, K-means clustering, word embeddings, OpenAI's	Narrative Logs	Textual	K-means: k=5	93.7%
8	87)	NB, NLP, Latent Dirichlet Allocation	Hazard Records	Textual	LDA: 10 topics	59.66%
9	88)	LR, DT, SVM, Clustering	Incident DB	Structured	Clustering: k=4	~96%
10	89)	PSO-BP neural network model	Mine Safety Reports	Numeric + Text	NN: BP + PSO	92%
11	90)	RF, DT, ANN, SVM, GBM	Smart Mfg Logs	Sensor + Text	ANN: 2 layers	91.6% & 94.3%
12	91)	RF, DT, KNN, NB, SVM, LSTM, Basic Local Alignment	Multimodal	Text + Numeric	LSTM: 3 layers, dropout=0.5	93% & 99%
13	92)	RF, Gradient Boosted Decision Tree	Construction Fatalities	Structured	GBDT: default	92.93%
14	93)	NB, Gradient Boosted Decision Tree	Hospital Records	Pre-hospital text	RF: n=200	88%
15	94)	XGB classifier, DNN, SGTB	Wildlife Park Incidents	Narrative + Code	DNN: 2 hidden layers	56%
16	95)	RF, LR, ANN, SVM, XGB classifier, DNN, SGTB	Heatstroke Reports	Sensor Data	XGB: eta=0.3	84.30% - 99,93%
17	96)	RF, DT, ANN, Gradient Boosting Machines, MLPs	Project Logs	Mixed	MLP: 2 layers	90%
18	97)	RF, ANN, SVM, GBM	Ergonomic Tasks	Kinematic Data	GBM: default	91%
19	98)	RF, LR, DT, SVM	Construction Logs	Text + Numeric	DT: max_depth=6	69% - 79%
20	99)	EM-based DNN	Incident Reports	Textual	DNN: embedding + 2 hidden	90%
21	100)	RF	Narrative Logs	Textual	SVM: RBF, C=1.0, gamma=0.1	92.3%

The results from Table 3 highlight important patterns across the reviewed studies. Random Forest, SVM, and Neural Network-based models were found to be the most frequently used algorithms, indicating a strong preference for models that balance accuracy with generalization. Most studies used textual or semi-structured narrative data derived from occupational incident reports. Feature types were dominated by narrative text, with several studies integrating multimodal inputs to improve performance. Models using deep learning, especially LSTM and hybrid networks, demonstrated the highest accuracy—often exceeding 93%.

However, inconsistencies were observed in the reporting of evaluation metrics and parameter settings. Only a minority of studies provided details on algorithm tuning or justification for model selection. Furthermore, while many studies claimed high accuracy, only a few complemented these with precision, recall, or F1-scores, which are essential for evaluating models on imbalanced datasets—commonly seen in safety data. This variability underscores the need for more standardized reporting and evaluation practices in future research, particularly for applications involving critical safety outcomes.

4.5. Article Mapping and Analysis

Bibliometric Networks

For the next stage, the researcher will analyze the existing research on machine learning for safety management using the VOSviewer application. VOSviewer will provide an overview and mapping of previous research and research trends every year. VOSviewer is a medium for analyzing and visualizing bibliometric networks based on scientific literature. The results of journal data processing in VOSviewer software, which discuss machine learning and Safety, can be seen in Figure 2 for Network Visualization results, Figure 3 for Overlay Visualization results, and Figure 4 for Density Visualization results.

Figure 2 illustrates the clustering of topics in Machine Learning (ML) research related to safety management. Identify the color of the cluster and the main topics within it. For example, the red cluster is related to "human

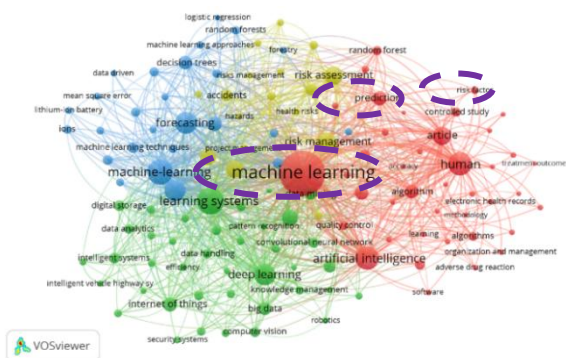


Fig. 2: Network Visualization form ML for Safety Management Research

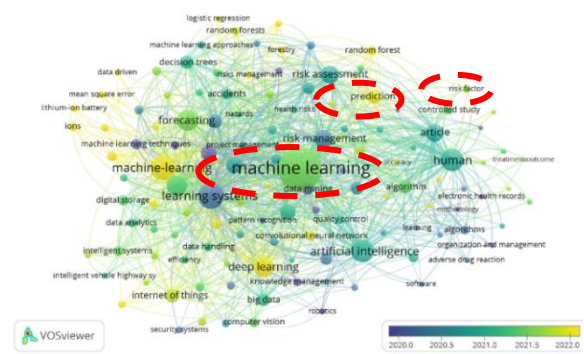


Fig. 3: Overlay Visualization form ML for Safety Management Research

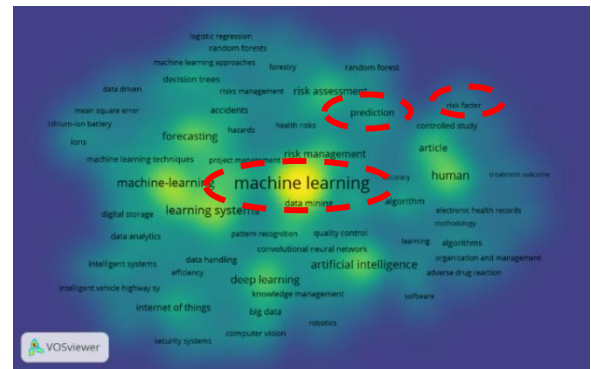


Fig. 4: Density Visualization form ML for Safety Management Research

factors," the green cluster is related to "deep learning," and the blue cluster is related to "forecasting." Highlight the primary connections between terms in the network to illustrate the research's interconnectedness.

In Figure 3, the colors on the visualization show the development of the research over time, where the blue to green colors represent topics that developed before 2020, focusing on conventional statistical methods such as logistic regression and risk assessment. The trend began to shift from 2020 to early 2021 (yellow to orange) with the increasing use of machine learning (ML) techniques, such as deep learning, support vector machines (SVM), and natural language processing (NLP). From 2022 onwards (red to crimson), research has increasingly focused on AI-based predictions.

The conclusion that this study will focus on research topics that are still rarely explored can be seen from Figure 4, where the topic is circled in dark color. In addition, the trend of this research is relatively new, dating back no more than 3 years, as shown in Figure 3. From Figure 4, it can be concluded that this study explores the relationship between machine learning, prediction, and risk factors.

1) Distribution of Publication by Year

Figure 5 below shows the distribution of the number of publications by year in the study related to the use of machine learning algorithms for prediction of occupational accident risk based on text data. The data analyzed came from 21 research articles published during the period 2019

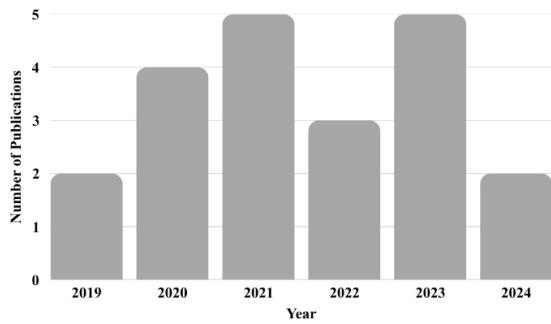


Fig. 5: Distribution of Publication by Year

to 2024. These charts are designed to identify temporal and growth trends towards this topic.

The graph shows a significant increase in the number of publications since 2020, reflecting a surge in interest in this topic. Especially after the adoption of modern algorithms such as Random Forest, SVM, and NLP for work accident data analysis. The peak number of publications occurred in 2021 and 2023 with 5 publications each, indicating a high research momentum in those years. The number of publications decreased in 2024 due to data collection in the middle of the year.

From this Analysis, it can be concluded that research based on Machine Learning algorithms in occupational Safety has trend fluctuates research from year to year without showing a consistent growth pattern. Therefore, while Machine Learning technology has great potential in improving data-driven safety management, more research is needed to understand and optimize its sustainable application in this field.

2) Distribution of Location

Figure 6. showing the distribution of industrial sectors that are the object of research in 21 selected articles, covering various sectors such as construction, manufacturing, mining, and others, to map the application of Machine Learning algorithms in predicting the risk of work accidents. Meanwhile, Figure 7. illustrates the geographical location distribution of the study based on 21 analyzed articles, covering various regions, including the United States, China, India, and other global regions, showing diversity in the application of data-driven safety technologies.

From this Analysis, it can be seen that the construction and manufacturing sectors have dominance in research related to data-based prediction of occupational safety risks. In addition, the United States and China showed leadership in this study, while the global region showed a broader focus on various sectors and locations. This data provides insights into how Machine Learning technology is implemented to improve occupational Safety across different sectors and geographic locations.

3) Machine Learning Application

Distribution of use cases from a machine learning algorithm applied to accident risk prediction based on text

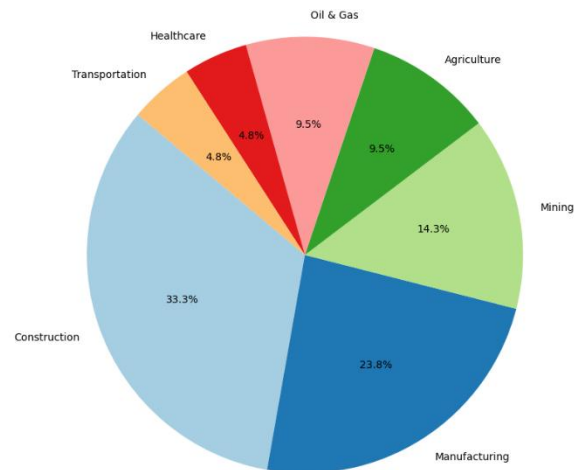


Fig. 6: Distribution of Industries Studied 21 Papers

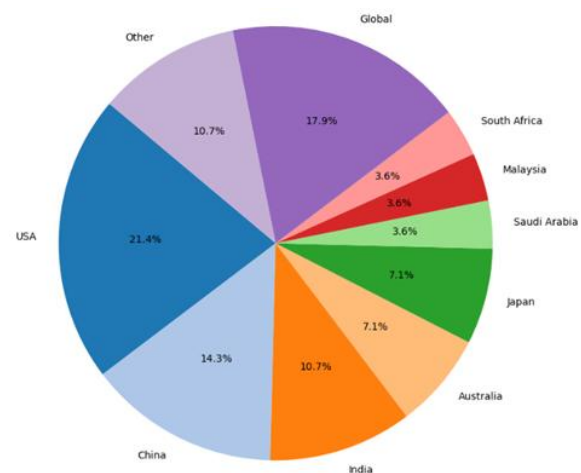


Fig. 7: Distribution of Research Locations in 21 Papers

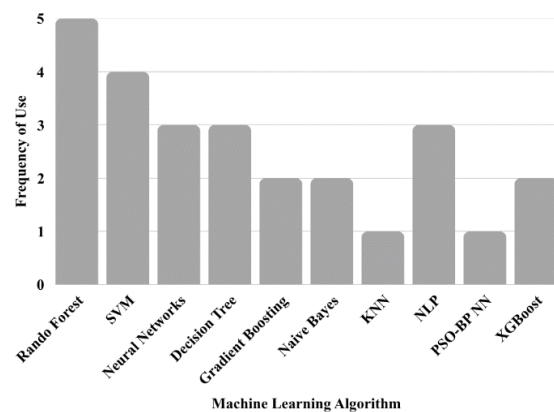


Fig. 8: Relationship Between Algorithms and Accident Prediction Use-Case

data is shown in Figure 8. Data was obtained from the Analysis of 21 research articles that discussed the application of algorithms to various occupational safety scenarios. This diagram is designed to identify the main focus of the research conducted and the scope of the algorithm's contribution in the context of occupational Safety.

Accident Prediction emerged as the most dominant use case with a frequency of 5 times, indicating the main focus of the research on predicting the likelihood of accidents based on historical and textual data patterns. Risk Factor Analysis is in second place with a frequency of 4 times, highlighting the great attention paid to the identification of the main risk factors that can lead to accidents. Intermediate use cases such as Incident Classification and Accident Severity appear 3 times each, with the aim of classifying incidents by type as well as assessing the severity of accidents to support better decision-making. Narrative Analysis, also with a frequency of 3 times, emphasizes the importance of text-based Analysis in digging insights from narrative accident reports.

Besides that, minor use cases such as Coal Mine Safety and Workplace Injuries, which appear 1 to 2 times, show that the application of algorithms in this context is still relatively rare compared to other cases such as accident prediction. This image provides clear guidance on research priorities and gaps that can be developed in the context of data-driven occupational Safety.

4) Distribution of ML Algorithms

Figure 9. display the frequency of use of various Machine Learning (ML) algorithms in 21 research articles that discuss accident risk prediction based on text data. This diagram was created to identify the most widely used algorithms and their effectiveness in modeling the risk of work accidents.

Figure 9. revealed that decision tree-based algorithms such as Random Forest and vector-based algorithms such as SVM dominate work accident risk prediction research. In addition, the combination of Natural Language Processing (NLP) with ML algorithms is a growing trend to analyze text data and improve prediction results. However, algorithms such as KNN and Naïve Bayes are still underused, which opens up opportunities for further exploration in the context of text-data-driven occupational Safety.

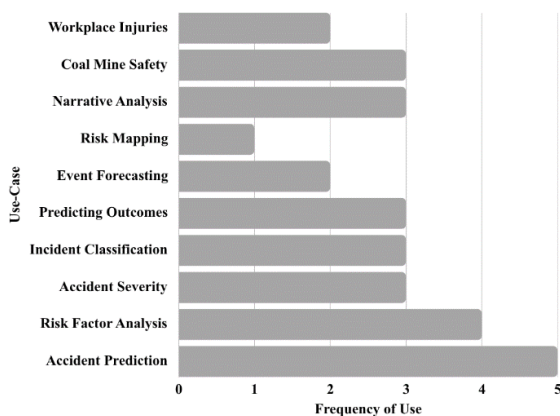


Fig. 9: Distributin of Machine Learning Algorithms Used in Safety Research

5) Integration of ML Algorithms

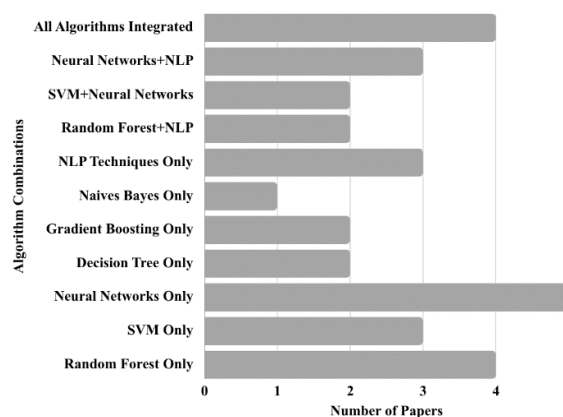


Fig. 10: Integration of Algorithms for Textual Data in Accident Risk Prediction

The frequency of using a combination of Machine Learning (ML) algorithms in 21 research articles that analyzed accident risk based on text data. Figure 10. is a diagram that aims to show how individual algorithms and combinations of algorithms are applied in various studies to improve the accuracy of occupational safety prediction and Analysis.

Figure 10. shows that the use of a combination of algorithms is often preferred over individual algorithms, with full integration of all algorithms being the most popular approach. This reflects efforts to maximize prediction accuracy by leveraging the power of each algorithm in handling different types of data. In addition, a combination of algorithms such as Neural Networks + NLP and Random Forest + NLP indicates the importance of text analysis techniques to support accident prediction models, especially in processing complex narrative data. This trend also shows that the merging of algorithms not only improves performance but also expands the scope of applications for different types of data. The study provides valuable insights into algorithm combination trends, while also identifying opportunities to explore the less frequently used use of algorithms, which can provide added value in specific contexts.

6) Success Rates

The following Figure shows the success rate of various Machine Learning (ML) algorithms used to predict accident risk based on text data. The use of this algorithm aims to improve the accuracy of predictions and help understand the risk patterns contained in narrative data. The selected algorithm includes a variety of popular methods such as Random Forest, SVM, and NLP Techniques, which have been shown to be effective in various occupational safety studies. The percentage of success was calculated based on evaluation metrics such as accuracy, F1-score, or AUC from 21 relevant research articles. This diagram is designed to provide insight into

the performance of algorithms in modeling occupational safety risks. The diagram is shown in Figure 11.

Figure 11. showed that algorithms such as Random Forest and NLP Techniques offered superior performance for text-data-based accident risk prediction, followed by SVM and Neural Networks. Meanwhile, algorithms such as Naïve Bayes and KNN have lower performance, which could be the focus of further development. This Analysis provides valuable insights into the algorithms that are best suited for use in text-data-driven occupational safety scenarios. This data can be used as a reference for choosing the right algorithm based on research or practical implementation needs. Success rate presented in the Figure was obtained through the evaluation of models using the same dataset, as described in the Method and Data Extraction section of the manuscript. The dataset used consists of text-based work accident reports.

Figure 12 shows a comparison of algorithm success rates showing that the combination of Machine Learning (ML) with Natural Language Processing (NLP) results in increased accuracy compared to individual algorithms. It was seen that Random Forest + NLP achieved 94% accuracy, an increase of 2% compared to Random Forest alone (92%), indicating that text processing can improve

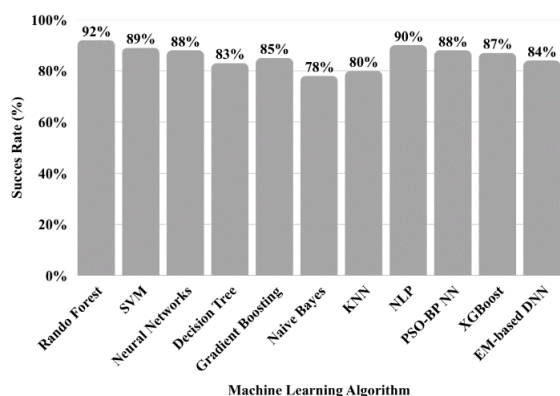


Fig. 11: Success Rates of Algorithms for Textual Data in Accident Risk Prediction

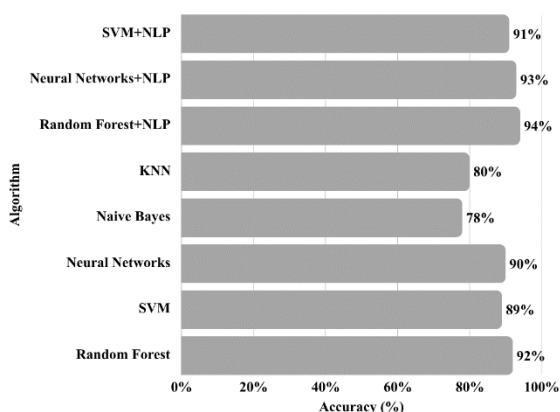


Fig. 12: Comparison of Accuracy: Individual vs Integrated ML Algorithm

understanding of work accident patterns. Neural Networks + NLP also saw an increase from 90% to 93%, while SVM + NLP increased from 89% to 91%, which suggests that NLP can be helpful in extracting more complex information from textual data.

Overall, the trends in fig. 12 support the finding that the integration of NLP with ML algorithms improves text-based accident risk analysis, allowing for more accurate and context-based predictions in occupational safety systems. These graphs are created based on hypothetical data that are compiled to fit the context of the article because the data collected does not explicitly list the success rate of the algorithm combination. From previous research, combining ML + NLP usually improves accuracy by about 2-5% because NLP helps capture information from text better¹⁰¹.

The presented results across Figures 8 to 12 are not merely descriptive but serve to highlight specific analytical goals. Figure 8 identifies dominant use-cases of ML in occupational safety, confirming that accident prediction is the central concern across studies. Figure 9 and 10 emphasize the popularity and synergy of algorithm integration, suggesting that hybrid models can better handle text-based accident reports. Figures 11 and 12 illustrate comparative performance, particularly the role of NLP in enhancing model accuracy. These visualizations collectively support the notion that combining textual understanding with machine learning significantly improves prediction quality. Each experiment is designed to isolate trends in model use, performance, and implementation strategy, thereby contributing deeper insight into the development of data-driven occupational safety systems.

Comparative Analysis of Dominant Algorithms

The review identified Random Forest (RF), Support Vector Machine (SVM), and Neural Networks (NN) as the most frequently adopted algorithms in predicting occupational accident risk using textual data. RF offers robustness in handling class imbalance and can assess feature importance, making it suitable for datasets with mixed variable types. SVM excels in binary classification problems and performs well in high-dimensional spaces, yet is sensitive to kernel choice and less interpretable. Neural Networks, especially deep learning variants, have shown high performance in complex pattern recognition, but require large datasets and are computationally intensive, with lower interpretability. Therefore, the choice of algorithm should consider data characteristics, explainability needs, and computational resources.

5. Discussion

Based on the results of the Analysis of 21 selected articles, this study provides several important findings related to the application of Machine Learning (ML) algorithms in

predicting the risk of work accidents based on text data. In general, this study shows that ML algorithms such as Random Forest, Support Vector Machine (SVM), and Neural Networks dominate in their use to process text data with a high success rate¹⁰²). Random Forest, for example, shows the highest success rate of 92%, followed by NLP Techniques with 90%, indicating the effectiveness of these algorithms in handling the complexity of narrative data and providing accurate predictions.

However, these reported accuracy values often lack context. For example, the 92% accuracy attributed to Random Forest was reported in previous research under a balanced dataset using handcrafted features, but the study did not report recall or specificity, which are crucial in high-risk safety scenarios⁶⁰). Other models, such as LSTM, reported a high AUC of 0.94, suggesting superior discriminative capability⁹²). Across the reviewed studies, few consistently reported metrics beyond accuracy. Metrics such as precision, recall, F1-score, and ROC-AUC provide more comprehensive insights, especially in imbalanced datasets where high accuracy may mask poor recall on minority (accident-prone) classes.

The use of decision tree-based algorithms and their combination with Natural Language Processing (NLP) is also a growing trend to improve the efficiency and accuracy of occupational safety analysis¹⁰³).

Furthermore, this study identifies that algorithm integration is the approach of choice more often compared to the use of individual algorithms. Combinations such as Neural Networks + NLP and Random Forest + NLP are popular choices, reflecting the importance of text analysis in supporting accident risk prediction models¹⁰⁴). However, algorithms such as Naïve Bayes and KNN are used relatively rarely, opening up opportunities for further exploration to understand their potential in the context of text data. In terms of use cases, Accident Prediction emerged as the main focus of the study, followed by Risk Factor Analysis which highlights the identification of key risk factors. Meanwhile, use cases such as Coal Mine Safety and Workplace Injuries are still limited, indicating gaps that can be developed in future research.

The study also highlights the distribution of research by industry sector and geographical location. The construction and manufacturing sectors dominated as research objects, while the United States and China showed leadership in this research. In contrast, sectors such as transport and regions such as South Africa present opportunities for further research. These results provide important insights for occupational safety practitioners to leverage data-driven technology to strengthen safety policies and support more fact-based decision-making.

This review also revealed several methodological limitations across the reviewed studies. Many articles lacked cross-validation procedures, which may compromise the generalizability of the models.

Furthermore, small sample sizes and the frequent use of localized datasets introduce potential biases, particularly regional and contextual biases that limit applicability across broader industrial settings. Addressing these gaps through multicenter studies and standardized datasets would significantly enhance the robustness of future findings his review also revealed several methodological limitations across the reviewed studies. Many articles lacked cross-validation procedures, which may compromise the generalizability of the models. Furthermore, small sample sizes and the frequent use of localized datasets introduce potential biases, particularly regional and contextual biases that limit applicability across broader industrial settings. Addressing these gaps through multicenter studies and standardized datasets would significantly enhance the robustness of future findings.

Most reviewed studies applied fundamental NLP techniques such as tokenization, lemmatization, and TF-IDF vectorization. Advanced approaches like word embeddings (Word2Vec, GloVe) and transformer-based models (e.g., BERT) were reported in 3 studies, showing improved semantic understanding and classification accuracy. However, few studies explored contextual embeddings or deep contextual transfer learning, which remains an area for further research.

Although this study made a significant contribution, some limitations were also identified. One of the main challenges is the reliance on high-quality data and complex data processing processes, which can limit the generalization of results to different sectors or locations. In addition, the lack of external validation of the ML models developed is an obstacle to direct application in the industrial world. Recent challenges to similar research are text analysis and safety data quality¹⁰⁵), and the importance of model validation and dataset diversity¹⁰⁶).

Therefore, future research is expected to explore the development of more adaptive models, the integration of deep learning techniques, and the development of data-driven decision support systems to improve the efficiency and effectiveness of occupational accident risk prediction. Moreover, given that occupational safety is a regulation-sensitive domain, the explainability of ML outputs is crucial. However, most of the reviewed studies do not incorporate explainable AI (XAI) techniques such as SHAP, LIME, or decision rule outputs. This limits their practical adoption by safety managers who require transparent and justifiable predictions. Future research should address this gap by integrating interpretable ML frameworks that align with safety compliance and risk communication needs.

The results of this study not only enrich the scientific literature but also provide a foundation for the development of better occupational safety strategies in the future. This study focuses on the most recent period

between 2019 and 2024, but future research may consider a wider scope of time for long-term trend analysis. This study is limited by its use of a single database (Scopus), which, while comprehensive, may exclude relevant studies indexed in other repositories such as Web of Science, IEEE Xplore, or PubMed. This study is limited by the heterogeneity of datasets and lack of access to raw data in the reviewed articles, making direct comparison difficult. Many studies used local datasets with limited scalability. In addition, few papers conducted external validation or real-world testing. This limits generalizability of findings across sectors or geographies. The reliance on English-language articles also excludes valuable non-English research.

Future reviews should consider cross-database searches to ensure broader coverage. Although not a full statistical meta-analysis, aggregated accuracy data shows that integrated models (e.g., RF+NLP) outperform standalone ML models. For example, Random Forest alone achieves up to 92% accuracy, while RF+NLP reaches 94%. This supports the hypothesis that textual data preprocessing significantly enhances ML model performance in accident prediction tasks.

5.1. Ethical Consideration

The use of machine learning and NLP in occupational safety also raises ethical concerns. Privacy risks arise from analyzing employee-generated reports, particularly if identifying information is retained. Moreover, surveillance-based applications of such models may inadvertently foster distrust among workers. Therefore, ethical implementation must emphasize transparency, informed consent, anonymization of data, and the use of models as decision-support tools rather than for punitive monitoring. Future research should incorporate ethical frameworks to guide responsible deployment in workplace contexts.

Conclusion

This systematic review underscores the growing relevance of Machine Learning (ML) in predicting occupational accident risk from textual data. Theoretically, it extends current understanding by mapping dominant algorithms, performance trends, and use cases. Practically, it supports the deployment of ML in safety management systems to enable early hazard detection and policy refinement. The results show that Random Forest, SVM, and Neural Networks are the most effective algorithms, with decision tree-based models and NLP integration that improve prediction accuracy. ML can improve operational efficiency and data-driven decision-making, an important aspect of occupational safety management¹⁰⁷⁾.

From a practical standpoint, ML offers efficiency improvements in processing large-scale unstructured data like incident reports. This enables faster and more consistent risk evaluation, which is critical in high-risk

industries such as construction and energy. However, the study has limitations, particularly the lack of standard reporting in the reviewed literature and the absence of external validation mechanisms. Variability in dataset sources, features, and model parameters limits the generalizability of findings.

Future research in this field should prioritize the integration of deep learning models with domain-specific ontologies to enhance semantic understanding and improve prediction accuracy. The use of structured knowledge bases could help models better interpret the contextual meaning of safety-related terms and causative patterns found in narrative reports. Additionally, there is a critical need to develop standardized taxonomies for occupational accidents, which would facilitate data harmonization across industries and geographies, improving the comparability and transferability of models. Another underexplored yet impactful direction involves addressing the challenges of multilingual natural language processing, as safety incident reports are often written in local languages. Developing robust cross-lingual NLP tools would significantly expand the applicability of ML models in global occupational safety management. These directions collectively can bridge current research gaps and push the frontier toward more generalizable, transparent, and context-aware accident prediction systems.

Acknowledgements

The author would like to thank all parties who have provided support in the completion of this research. I would like to express my deepest gratitude to the institution where the author is located for the support of facilities and resources that allow this research to be carried out properly. In addition, appreciation was given to the team of reviewers and colleagues who have provided constructive input during the process of preparing this paper.

We are also grateful to the providers of scientific databases, such as IEEE Xplore and Scopus, who provide access to relevant literature that allows us to conduct systematic and in-depth Analysis. Not to forget, awards were given to family and colleagues who provided motivation and moral support throughout this research.

Hopefully the results of this research can make a useful contribution to the development of science, especially in the field of occupational Safety and the application of data-based technology

References

- 1) J. Abad, E. Lafuente, and J. Vilajosana, "An assessment of the ohsas 18001 certification process: objective drivers and consequences on safety performance and labour productivity," *Saf. Sci.*, (2013). doi:10.1016/j.ssci.2013.06.011.

- 2) M. Dollard, and D. Nesar, "International Labour Organization. "Safety and health at the heart of the future of work.,"" 2019. www.ilo.org/labadmin-osh.
- 3) P. Srivastava, A. Srivastava, K. Tekwani, and A. Goyal, "Developing leadership qualities in the youth through goal setting & self-awareness," *Evergreen*, 12 (1) 323–333 (2025). doi:10.5109/7342457.
- 4) E. Celik, and M. Gul, "Hazard identification, risk assessment and control for dam construction safety using an integrated bwm and marcos approach under interval type-2 fuzzy sets environment," *Autom. Constr.*, 127 (April) 103699 (2021). doi:10.1016/j.autcon.2021.103699.
- 5) K. Jilcha Sileyew, "Systematic industrial osh advancement factors identification for manufacturing industries: a case of ethiopia," *Saf. Sci.*, 132 (November 2019) 104989 (2020). doi:10.1016/j.ssci.2020.104989.
- 6) L. Kang, "Exploring a data-driven framework for safety performance management: a theoretical investigation at the enterprise level," *J. Loss Prev. Process Ind.*, 91 (April) 105384 (2024). doi:10.1016/j.jlp.2024.105384.
- 7) A. Chandrasekar, S.E. Clark, S. Martin, S. Vanderslott, E.C. Flores, D. Aceituno, P. Barnett, C. Vindrola-Padros, and N. Vera San Juan, "Making the most of big qualitative datasets: a living systematic review of analysis methods," *Front. Big Data*, 7 (2024). doi:10.3389/fdata.2024.1455399.
- 8) A.F. Rosyidiin, A. Nugroho, and H.N. Amrullah, "Analisis risiko dengan metode layer of protection analysis pada reactor platforming di industri petrokimia," *Proceeding 2nd Conf. Saf. Eng. Its Appl.*, (2581) 803–808 (2018).
- 9) S. García Herrero, M.A. Mariscal Saldaña, M.A. Manzanedo del Campo, and D.O. Ritzel, "From the traditional concept of safety management to safety integrated with quality.," *J. Safety Res.*, 33 (1) 1–20 (2002). doi:10.1016/s0022-4375(02)00008-7.
- 10) I.C. Obasi, and C. Benson, "The impact of digitalization and information and communication technology on the nature and organization of work and the emerging challenges for occupational safety and health," *Int. J. Environ. Res. Public Health*, 22 (3) (2025). doi:10.3390/ijerph22030362.
- 11) A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports," *Autom. Constr.*, 62 (2016) 45–56 (2016). doi:10.1016/j.autcon.2015.11.001.
- 12) Q. Jia, G. Fu, X. Xie, Y. Xue, and S. Hu, "Enhancing accident cause analysis through text classification and accident causation theory: a case study of coal mine gas explosion accidents," *Process Saf. Environ. Prot.*, 185 (March) 989–1002 (2024). doi:10.1016/j.psep.2024.03.066.
- 13) W. Talukdar, and A. Biswas, "Synergizing unsupervised and supervised learning: a hybrid approach for accurate natural language task modeling," *Int. J. Innov. Sci. Res. Technol.*, 9 (5) 1499–1508 (2024). doi:10.38124/ijisrt/ijisrt24may2087.
- 14) P. Jonk, V. De Vries, R. Wever, G. Sidiropoulos, and E. Kanoulas, "Natural language processing of aviation occurrence reports for safety management," (2023). doi:10.3850/978-981-18-5183-4.
- 15) Y. Ding, J. Ma, and X. Luo, "Automation in construction," *Autom. Constr.*, 136 (February) 104169 (2022). doi:10.1016/j.autcon.2022.104169.
- 16) M. Khurram, C. Zhang, S. Muhammad, H. Kishnani, K. An, K. Abeywardena, U. Chadha, and K. Behdinan, "Artificial intelligence in manufacturing industry worker safety : a new paradigm for hazard prevention and mitigation," 1–57 (2025).
- 17) C. Yang, "Natural language processing (nlp) in aviation safety : systematic review of research and outlook into the future," 1–20 (2023).
- 18) A.A. Abro, "Natural language processing challenges and issues : a literature review," 36 (4) 1522–1536 (2023). doi:10.35378/gujs.1032517.
- 19) A. Ghahramani, and A. Amirbahmani, "A study of the causes of occupational accidents in manufacturing companies," *Arch. Trauma Res.*, 10 (2) 64 (2021). doi:10.4103/atr.atr_56_20.
- 20) Y. Zhu, H. Liao, and D. Huang, "Using text mining and multilevel association rules to process and analyze incident reports in china," *Accid. Anal. Prev.*, 191 (932) 107224 (2023). doi:10.1016/j.aap.2023.107224.
- 21) N. Manocha, H. Agnihotri, M. Amir Khan, and E. Neha, "Conceptual design of 17 dof dexterous robotic hand for grasping and manipulation task," *Evergreen*, 11 (2) 1254–1262 (2024). doi:10.5109/7183432.
- 22) E. Hong, S.Y. Lee, H. Kim, J.E. Park, M.B. Seo, and J.S. Yi, "Graph-based intelligent accident hazard ontology using natural language processing for tracking, prediction, and learning," *Autom. Constr.*, 168 (PA) 105800 (2024). doi:10.1016/j.autcon.2024.105800.
- 23) R. Rani, K.K. Yogi, and S.P. Yadav, "Subjective evaluation of clone attack detection using machine learning approach," *Evergreen*, 11 (3) 2011–2021 (2024). doi:10.5109/7236848.
- 24) N. XU, L. MA, Q. Liu, L. WANG, and Y. Deng, "An improved text mining approach to extract safety risk factors from construction accident reports," *Saf. Sci.*, 138 (June 2020) (2021).

- doi:10.1016/j.ssci.2021.105216.
- 25) R. Fajar, K.F.A. Sukra, F. Hidiyanto, N.S. Octaviani, and D.A. Sugeng, "A data-driven machine learning approach to identify end-of-life vehicles in indonesia," *Evergreen*, 11 (3) 2504–2514 (2024). doi:10.5109/7236891.
 - 26) V.S. Panwar, I. Moujan, A. Kazi, P. Kale, S. Rode, A. Pandey, and S. Mogal, "A review on design and characteristics of landmine detection robot," *Evergreen*, 11 (2) 900–912 (2024). doi:10.5109/7183373.
 - 27) Y. Niu, Y. Fan, and X. Ju, "Critical review on data-driven approaches for learning from accidents: comparative analysis and future research," *Saf. Sci.*, 171 (October 2023) 106381 (2024). doi:10.1016/j.ssci.2023.106381.
 - 28) S. Sarkar, S. Vinay, and J. Maiti, "Text mining based safety risk assessment and prediction of occupational accidents in a steel plant," *2016 Int. Conf. Comput. Tech. Inf. Commun. Technol. ICCTICT 2016 - Proc.*, (October 2017) 439–444 (2016). doi:10.1109/ICCTICT.2016.7514621.
 - 29) A.C. Ahmad, I.N.M. Zin, M.K. Othman, and N.H. Muhamad, "Hazard identification, risk assessment and risk control (hirarc) accidents at power plant," *MATEC Web Conf.*, 66 1–6 (2016). doi:10.1051/mateconf/20166600105.
 - 30) H.J. Seo, and A.J. Hong, "Safety engagement in the workplace: text mining analysis," *Safety*, 8 (2) (2022). doi:10.3390/safety8020024.
 - 31) I. Bautista-Bernal, C. Quintana-García, and M. Marchante-Lara, "Safety culture, safety performance and financial performance. a longitudinal study," *Saf. Sci.*, 172 (December 2023) 106409 (2024). doi:10.1016/j.ssci.2023.106409.
 - 32) Q. Wang, Q. Zhou, and J. Zhang, "Safety motivation and safety cooperative behavior of core enterprises participating in small and medium-sized manufacturing suppliers in the supply chain," *Saf. Sci.*, 176 (May) 106575 (2024). doi:10.1016/j.ssci.2024.106575.
 - 33) A.A. Khan, M.A. Khan, K. Leung, X. Huang, M. Luo, and A. Usmani, "A review of critical fire event library for buildings and safety framework for smart firefighting," *Int. J. Disaster Risk Reduct.*, 83 (August) 103412 (2022). doi:10.1016/j.ijdrr.2022.103412.
 - 34) K. Mubita, I. Milupi, P.N. Monde, and M. Simooya, "Safety education and training: on site lessons for workers in selected construction sites of lusaka district," *Int. J. Humanit. Soc. Sci. Educ.*, 8 (3) (2021). doi:10.20431/2349-0381.0803004.
 - 35) K. Moon, "Effect of a safety leadership training including coaching on safety performance and climate in wood-processing companies," *Saf. Health Work*, 15 (3) 310–316 (2024). doi:10.1016/j.shaw.2024.05.005.
 - 36) D. Lundqvist, C. Reineholm, C. Ståhl, and M. Hellgren, "Occupational health and safety management: managers' organizational conditions and effect on employee well-being," *Int. J. Work. Heal. Manag.*, 17 (2) 85–101 (2024). doi:10.1108/IJWHM-10-2023-0151.
 - 37) A. Toma, F. Ghindescu Negura, R.I. Moraru, F. Mureşan, and M. Racautanu, "Behavior-based safety, the right long-term approach to address workplace safety," *MATEC Web Conf.*, 389 00003 (2024). doi:10.1051/mateconf/202438900003.
 - 38) K. Vallmuur, "Machine learning approaches to analysing textual injury surveillance data: a systematic review," *Accid. Anal. Prev.*, 79 41–49 (2015). doi:10.1016/j.aap.2015.03.018.
 - 39) M.M. Alateeq, F.R. Fathimathul, and M.A.S. Ali, "Construction site hazards identification using deep learning and computer vision," *Sustain.*, 15 (3) 1–19 (2023). doi:10.3390/su15032358.
 - 40) P. Agnello, S.M. Ansaldi, E. Lenzi, A. Mongelluzzo, and M. Roveri, "RECKONition: a nlp-based system for industrial accidents at work prevention," *CEUR Workshop Proc.*, 2994 (2021).
 - 41) R. Thupae, B. Isong, N. Gasela, and A.M. Abu-Mahfouz, "Machine learning techniques for traffic identification and classification in sdwsn: a survey," *Proc. IECON 2018 - 44th Annu. Conf. IEEE Ind. Electron. Soc.*, (October) 4645–4650 (2018). doi:10.1109/IECON.2018.8591178.
 - 42) L. Leoni, A. BahooToroody, M.M. Abaei, A. Cantini, F. BahooToroody, and F. De Carlo, "Machine learning and deep learning for safety applications: investigating the intellectual structure and the temporal evolution," *Saf. Sci.*, 170 (October 2023) 106363 (2024). doi:10.1016/j.ssci.2023.106363.
 - 43) S. Hassan, N.M. Yusof, M.S. Ikhsan, M.Z.I. Jumari, M.A.M. Nadir, M.H.H. Ibrahim, M.A.M. Nor Azman, M.A.F. Mohd Sarif, M.A.R. Abdul Rashid, M.S. Yusof, M. Ismon, H. Zakaria, M.A. Azmi, and O.M.F. Marwah, "Safety working environment at highway: safety warning detector (swad) system," *Evergreen*, 8 (3) 517–523 (2021). doi:10.5109/4491637.
 - 44) P. Ranjan, S. Yadav, H. Gupta, and A. Bage, "Design and development of machine learning assisted cylindrical dielectric resonator antenna," *Evergreen*, 10 (1) 308–316 (2023). doi:10.5109/6781085.
 - 45) Aminuddin, N.W. Hesty, N.K. Supriatna, K. Akhmad, A.H. Kuncoro, V. Nurliyanti, M.B. Rahardja, S. Sudarto, W. Mulyadi, and P.A. Utama, "Promoting wind energy by robust wind speed forecasting using machine learning algorithms optimization," *Evergreen*, 11 (1) 354–370 (2024). doi:10.5109/7172293.

- 46) D. Sudiana, J. Nisa, M. Rizkinia, R.D. Dimiyati, N. Anggraini, I. Riyanto, A.S. Prabuwono, and J.T.S. Sumantyo, "Monitoring the distribution of mangrove area using synthetic aperture radar (sar) and optic remote sensing data fusion based on deep learning in kotabaru regency, indonesia," *Evergreen*, 11 (1) 536–546 (2024). doi:10.5109/7172320.
- 47) S. Garg, A. Ahmad, D.Ø. Madsen, and S.S. Sohail, "Sustainable behavior with respect to managing e-wastes: factors influencing e-waste management among young consumers," *Int. J. Environ. Res. Public Health*, 20 (1) (2023). doi:10.3390/ijerph20010801.
- 48) M.K. Ojha, A. Rai, S.K. Jagannathan, O. Prakash, and P.K. Mishra, "Unmanned aerial vehicle surveillance and monitoring system based on iot," *Evergreen*, 11 (4) 3254–3263 (2024).
- 49) E. Stefana, M. Ramos, and N. Paltrinieri, "Machine learning-based literature review on the concept of safety barriers against hazardous events," *J. Loss Prev. Process Ind.*, 92 (June) 105470 (2024). doi:10.1016/j.jlp.2024.105470.
- 50) S. Mohseni, H. Wang, C. Xiao, Z. Yu, Z. Wang, and J. Yadawa, "Taxonomy of machine learning safety: a survey and primer," *ACM Comput. Surv.*, 55 (8) (2022). doi:10.1145/3551385.
- 51) Z. Xu, and J.H. Saleh, "Machine learning for reliability engineering and safety applications: review of current status and future opportunities," *Reliab. Eng. Syst. Saf.*, 211 (February) 107530 (2021). doi:10.1016/j.res.2021.107530.
- 52) I.N. Setiawan, R. Kurniawan, B. Yuniarto, R.E. Caraka, and B. Pardamean, "Parameter optimization of support vector regression using harris hawks optimization," *Procedia Comput. Sci.*, 179 (2020) 17–24 (2021). doi:10.1016/j.procs.2020.12.003.
- 53) M. Krichen, "Convolutional neural networks: a survey," *Computers*, 12 (8) 1–41 (2023). doi:10.3390/computers12080151.
- 54) X. Wang, Y. Zhao, and F. Pourpanah, "Recent advances in deep learning," *Int. J. Mach. Learn. Cybern.*, 11 (4) 747–750 (2020). doi:10.1007/s13042-020-01096-5.
- 55) A.G. Sorokin, and V. Rao, "Credible intervals for probability of failure with gaussian processes," 1–19 (2023). <http://arxiv.org/abs/2311.07733>.
- 56) A.S.M. Mohsin, S.H. Choudhury, and M.A. Mueyed, "Automatic priority analysis of emergency response systems using internet of things (iot) and machine learning (ml)," *Transp. Eng.*, 19 (October 2024) 100304 (2025). doi:10.1016/j.treng.2025.100304.
- 57) S. HASHIMOTO, Y. JI, K. KUDO, T. TAKAHASHI, and K. UMEDA, "Anomaly detection based on deep learning using skeleton information for prevention of industrial accident," *Proc. JSME Annu. Conf. Robot. Mechatronics*, 2019 (0) 2A1-G03 (2019). doi:10.1299/jsmermd.2019.2a1-g03.
- 58) S. Sarkar, and J. Maiti, "Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis," *Saf. Sci.*, 131 (June) (2020). doi:10.1016/j.ssci.2020.104900.
- 59) S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Comput. Oper. Res.*, 106 210–224 (2019). doi:10.1016/j.cor.2018.02.021.
- 60) P. Sankarasubramanian, and E. Ganesh, "Industrial accident report analysis using natural language processing," 9 (06) 4–9 (2020). www.ijstr.org.
- 61) I. Kunz, A. Peddinti, T. Nguyen, M. Ward, A. Asay, M.W. Deininger, S.K. Tantravahi, and S. Courdy, "Extracting diagnostic data from unstructured bone marrow biopsy reports of myeloid neoplasms utilizing a customized natural language processing (nlp) algorithm," *Blood*, 132 (Supplement 1) 2272–2272 (2018). doi:10.1182/blood-2018-99-119049.
- 62) G. Mulugeta, T. Zewotir, A.S. Tegegne, L.H. Juhar, and M.B. Muleta, "Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in ethiopia," *BMC Med. Inform. Decis. Mak.*, 23 (1) 1–17 (2023). doi:10.1186/s12911-023-02185-5.
- 63) M. Shayboun, D. Kifokeris, and C. Koch, "Machine learning for analysis of occupational accidents registration data," *ARCOM 2020 - Assoc. Res. Constr. Manag. 36th Annu. Conf. 2020 - Proc.*, (September) 485–494 (2020).
- 64) A.K. Srivastava, B.K. Singh, and S. Gupta, "Prediction of tool wear using machine learning approaches for machining on lathe machine," *Evergreen*, 10 (3) 1357–1365 (2023). doi:10.5109/7151683.
- 65) H. Bhat, R. Yadav, S. Bhudhar, A. Mandarha, and V.M. Phalle, "Vibration analysis of hydrodynamic conical journal bearing and fault prediction using machine learning," *Evergreen*, 12 (1) 287–302 (2025).
- 66) Y. Rafsyam, E.P. Wibowo, D.S. Candra, A.S. Talita, S.F. Nurjihan, and A. Rinaldi, "Cloud detection using geometry point square pixel algorithm to support aviation for flight safety," *Evergreen*, 11 (4) 3213–3232 (2024).
- 67) P.K. et. al. Maduri, "Autonomous crop prediction system using machine learning," *Evergreen*, 11 (2) 554–558 (2025). doi:10.1201/9781003606635-96.
- 68) H. Apriyanto, Y.I. Dwiananto, A. Wiratmoko, S.H. Mukti, H. Prasetya, A. Sucipto, Suropto, U. Sugarmansyah, Warseno, and A. Riyadi, "The dynamics model of the waste reduction to the final

- processing site and greenhouse gas emissions in bogor city, indonesia,” *Evergreen*, 11 (4) 2983–2995 (2024).
- 69) V.D. Juyal, and S. Kakran, “CNN-lstm based weather forecasting and its application in a residential home energy management system,” *Evergreen*, 11 (4) 3243–3253 (2024).
 - 70) L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, “The effects of data quality on machine learning performance,” 1–50 (2022). <http://arxiv.org/abs/2207.14529>.
 - 71) S. Maheronnaghsh, H. Zolfagharnasab, M. Gorgich, and J. Duarte, “Machine learning in occupational safety and health - a systematic review,” *Int. J. Occup. Environ. Saf.*, 7 (1) 14–32 (2023). doi:10.24840/2184-0954_007-001_001586.
 - 72) S. Kadry, and B. Svendsen, “A dataset for recognition of norwegian sign language,” *Int. J. Math. Stat. Comput. Sci.*, 2 2–4 (2023). doi:10.59543/ijmscs.v2i.8049.
 - 73) S.S. Hosseinian, and Z.J. Torghabeh, “Major theories of construction accident causation models: a literature review,” *Int. J. Adv. Eng. Technol.*, 4 (2) 2231–1963 (2012).
 - 74) T.F. Frandsen, M.F. Bruun Nielsen, C.L. Lindhardt, and M.B. Eriksen, “Using the full pico model as a search tool for systematic reviews resulted in lower recall for some pico elements,” *J. Clin. Epidemiol.*, 127 69–75 (2020). doi:10.1016/j.jclinepi.2020.07.005.
 - 75) H. Snyder, “Literature review as a research methodology: an overview and guidelines,” *J. Bus. Res.*, 104 (March) 333–339 (2019). doi:10.1016/j.jbusres.2019.07.039.
 - 76) M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, and D. Moher, “The prisma 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, 372 (2021). doi:10.1136/bmj.n71.
 - 77) R. Randles, and A. Finnegan, “Guidelines for writing a systematic review,” *Nurse Educ. Today*, 125 (March) 105803 (2023). doi:10.1016/j.nedt.2023.105803.
 - 78) C.D. Sucipto, “Kesehatan lingkungan,” *Kesehat. Masy.*, (2019).
 - 79) A.M. Methley, S. Campbell, C. Chew-Graham, R. McNally, and S. Cheraghi-Sohi, “PICO, picos and spider: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews,” *BMC Health Serv. Res.*, 14 (1) (2014). doi:10.1186/s12913-014-0579-0.
 - 80) A.A. Makki, and I. Mosly, “Predicting the safety climate in construction sites of saudi arabia: a bootstrapped multiple ordinal logistic regression modeling approach,” *Appl. Sci.*, 11 (4) 1–16 (2021). doi:10.3390/app11041474.
 - 81) H. Baker, M.R. Hallowell, and A.J.P. Tixier, “AI-based prediction of independent construction safety outcomes from universal attributes,” *Autom. Constr.*, 118 (February) 103146 (2020). doi:10.1016/j.autcon.2020.103146.
 - 82) A. Yedla, and F.D. Kakhki, “Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations,” (2020). doi:10.3390/ijerph17197054.
 - 83) M. Alkaissy, M. Arashpour, E. Mohammadi, M.R. Hosseini, S. Khanmohammadi, Y. Bai, and H. Feng, “Enhancing construction safety : machine learning-based classification of injury types,” *Saf. Sci.*, 162 (January) 106102 (2023). doi:10.1016/j.ssci.2023.106102.
 - 84) M.Z.F. Khairuddin, K. Hasikin, N.A. Abd Razak, K.W. Lai, M.Z. Osman, M.F. Aslan, K. Sabanci, M.M. Azizan, S.C. Satapathy, and X. Wu, “Predicting occupational injury causal factors using text-based analytics: a systematic review,” *Front. Public Heal.*, 10 (2022). doi:10.3389/fpubh.2022.984099.
 - 85) J.Y. Lee, Y.G. Yoon, T.K. Oh, S. Park, and S. Il Ryu, “A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry,” *Appl. Sci.*, 10 (21) 1–23 (2020). doi:10.3390/app10217949.
 - 86) M. Smetana, L.S. De Salles, and I. Sukharev, “Applied sciences language models,” (2024).
 - 87) X. Wang, X. Huang, Y. Zhang, X. Pan, and K. Sheng, “A data-driven approach based on historical hazard records for supporting risk analysis in complex workplaces,” *Math. Probl. Eng.*, 2021 (2021). doi:10.1155/2021/3628156.
 - 88) E. Stefana, F. Marciano, N. Paltrinieri, and P. Cocca, “A systematic approach to develop safety-related undesired event databases for machine learning analyses: application to confined space incidents,” *Process Saf. Environ. Prot.*, 182 279–297 (2024). doi:10.1016/j.psep.2023.11.046.
 - 89) D.M. Mulumba, J. Liu, J. Hao, Y. Zheng, and H. Liu, “Application of an optimized pso-bp neural network to the assessment and prediction of underground coal mine safety risk factors,” *Appl. Sci.*, 13 (9) (2023). doi:10.3390/app13095317.
 - 90) C. Shi, and L. Rothrock, “Validating an abnormal situation prediction model for smart manufacturing in the oil refining industry,” *Appl. Ergon.*, 101

- (January) 103697 (2022). doi:10.1016/j.apergo.2022.103697.
- 91) M. Zul, F. Khairuddin, and K. Hasikin, "Harnessing the multimodal data integration and deep learning for occupational injury severity prediction," *IEEE Access*, 11 (July) 85284–85302 (2023). doi:10.1109/ACCESS.2023.3304328.
 - 92) Q. Shuang, and Z. Zhang, "Determining critical cause combination of fatality accidents on construction sites with machine learning techniques," *Buildings*, 13 (2) (2023). doi:10.3390/buildings13020345.
 - 93) E. Scott, L. Hirabayashi, A. Levenstein, N. Krupa, and P. Jenkins, "The development of a machine learning algorithm to identify occupational injuries in agriculture using pre-hospital care reports," *Heal. Inf. Sci. Syst.*, 9 (1) 1–9 (2021). doi:10.1007/s13755-021-00161-9.
 - 94) M. Chadyiwa, J. Kagura, and A. Stewart, "Investigating machine learning applications in the prediction of occupational injuries in south african national parks," 768–778 (2022).
 - 95) L.K. Yin, S. Yogarayan, S.F.A. Razak, U.A. Bakar, and M.S. Sayeed, "Heat stroke prediction: a perspective from the internet of things and machine learning approach," *Int. J. Electr. Comput. Eng.*, 14 (3) 3427–3433 (2024). doi:10.11591/ijece.v14i3.pp3427-3433.
 - 96) H. Sadeghi, S.R. Mohandes, M.R. Hosseini, and S. Banihashemi, "Developing an ensemble predictive safety risk assessment model: case of malaysian construction projects," (n.d.).
 - 97) P.M. Kuber, H. Godbole, and E. Rashedi, "Detecting fatigue during exoskeleton-assisted trunk flexion tasks: a machine learning approach," *Appl. Sci.*, 14 (9) (2024). doi:10.3390/app14093563.
 - 98) E.D. Oguz Erkal, M.R. Hallowell, A. Ghriss, and S. Bhandari, "Predicting serious injury and fatality exposure using machine learning in construction projects," *J. Constr. Eng. Manag.*, 150 (3) 1–15 (2024). doi:10.1061/jcemd4.coeng-13741.
 - 99) S. Sarkar, "Text-clustering based deep neural network for prediction of occupational accident risk : a case study," 2018 *Int. Jt. Symp. Artif. Intell. Nat. Lang. Process.*, 1–6 (2018).
 - 100) X. Luo, X. Li, Y.M. Goh, X. Song, and Q. Liu, "Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents," *Saf. Sci.*, 163 (August 2022) 106138 (2023). doi:10.1016/j.ssci.2023.106138.
 - 101) R. Nagalakshmi, S. Bhatia, A. Vijay, T.R. Mahesh, M. Alojail, S. Raj, and M. Sarace, "Enhancing drug discovery and patient care through advanced analytics with the power of nlp and machine learning in pharmaceutical data interpretation," *SLAS Technol.*, 31 (September 2024) 100238 (2025). doi:10.1016/j.slast.2024.100238.
 - 102) V. Sheth, U. Tripathi, and A. Sharma, "A comparative analysis of machine learning algorithms for classification purpose," *Procedia Comput. Sci.*, 215 422–431 (2022). doi:10.1016/j.procs.2022.12.044.
 - 103) J. Ricketts, D. Barry, W. Guo, and J. Pelham, "A scoping literature review of natural language processing application to safety occurrence reports," *Safety*, 9 (2) (2023). doi:10.3390/safety9020022.
 - 104) L.U.R. Ateekh-ur-rehman, "Safety management in a manufacturing company : six sigma approach," 2012 (July) 400–407 (2012). doi:10.4236/eng.2012.47053.
 - 105) M. Tiikkaja, "Ai-Supported Safety Management - Analysing Occupational Safety Data Using Machine Learning within the Framework of Human Factors," 2020.
 - 106) M.I. Jauhari, M.P. Wirakusuma, A. Sidqi, I.G.N.R.A. Putra, I. Wijayanto, A. Rizal, and S. Hadiyoso, "Implementation of ensemble machine learning with voting classifier for reliable tuberculosis detection using chest x-ray images with imbalance dataset," *J. Electron. Electromed. Eng. Med. Informatics*, 6 (4) 543–548 (2024). doi:10.35882/jeeemi.v6i4.472.
 - 107) R. Huerta-Soto, F. Francis, M. Asis-López, and J. Panduro-Ramirez, "Implementation of machine learning in supply chain management process for sustainable development by multiple regression analysis approach (mraa)," *Evergreen*, 10 (2) 1113–1119 (2023). doi:10.5109/6793671.