Results in Engineering 28 (2025) 107762

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Key predictors of injury severity in occupational accidents involving construction-site vehicles

A Sánchez-Lite [a,*], J.L. Fuentes-Bargues [b], J.M. Geijo-Barrientos [a], C González-Gaya [c], A.Z. Sampaio [d]

[a] Department of Materials Science and Metallurgical Engineering, Graphic Expression in Engineering, Cartographic Engineering, Geodesy and Photogrammetry, Mechanical Engineering and Manufacturing Engineering, School of Industrial Engineering, Universidad de Valladolid, P° del Cauce 59, 47011 Valladolid, Spain
[b] Project Management, Innovation and Sustainability Research Center (PRINS), Universitat Politècnica de València, 46022 Valencia, Spain
[c] Construction and Manufacturing Engineering Department, National Distance Education University (UNED), C/Juan del Rosal 12, 28040 Madrid, Spain
[d] Department of Civil Engineering and Architecture, Higher Technical School, University of Lisbon, 1049-001 Lisbon, Portugal

ARTICLE INFO

Keywords:
Occupational accidents
Material agent
Construction
Vehicles
Accident statistics

ABSTRACT

Across national statistics, construction repeatedly ranks among sectors with the highest injury and fatality rates. Vehicle-related accidents constitute a modest share of minor injuries yet contribute a significant fraction of construction fatalities. This study analysed 16,781 Spanish construction vehicle–related accidents recorded from 2009 to 2022 (2.5% severe-fatal) to identify determinants of injury severity and develop predictive models. Records were retrieved from Delt@, the compulsory national electronic occupational injury reporting platform. Variables were structured into two domains (organisational, contextual) and five categories. Methods combined descriptive profiling, χ^2 association tests, mutual-information ranking and three machine-learning classifiers (Random Forest, XGBoost, multilayer perceptron). Seven predictors—hour block, worker age, job tenure, site zone, deviation pattern, injury type and body region—showed the strongest association with severity. Separate models were trained on contextual and organisational feature sets. The contextual model detected 87.1% of severe/fatal cases (balanced accuracy 88.1.%), while the organisational model detected 59.3% (balanced accuracy 62.1%). The findings emphasise the importance of scheduling (time-of-day exposure), targeted training for short-tenure and at-risk age groups (30–59 years old), and control of the site zone. These results provide practical guidance for managers, regulators, engineers and safety practitioners seeking to reduce the number of vehicle-related accidents on construction sites, particularly those with a high level of severity.

1. Introduction

Occupational accidents represent a significant challenge for companies, administrations, workers and society in general [1,2]. Among the productive sectors, the construction sector has had the highest total and fatal occupational accident rates in different countries around the world for many decades [3–7].

Researchers have linked this persistently high rate to the sector's dynamic work settings [8–9], other authors point to the lack of training of workers as one of the main causes [10], and other studies highlight the influence of inadequate risk perception by workers [11] and the lack of workers' awareness of occupational safety [12]. Indeed, a lack of occupational safety culture, in a dynamic and constantly changing environment, with constant movements and interactions between personnel, materials, and work equipment, means that safety risks are often overlooked, and more injuries (both minor and serious) are generated by occupational accidents [13–15].

Another line of work on accidents in the construction sector is that carried out according to the categories with the highest number of accidents, with falls, struck-by accidents, entrapment, and electric shocks being the most important typologies [6,16,17]. Specific work has been carried out on accidents caused by falls from heights [18–22], on accidents caused by electrical electrocutions [23–25], and on accidents caused by tower cranes [26–28].

Struck-by accidents by construction site vehicles and by falling objects represent a low percentage of minor accidents [6,29], but instead represent the second cause of fatal occupational accidents in construction [30–32], although, despite their importance, they have not been

* Corresponding author.
E-mail address: asanchez@uva.es (A. Sánchez-Lite).

https://doi.org/10.1016/j.rineng.2025.107762
Received 14 July 2025; Received in revised form 24 September 2025; Accepted 15 October 2025
Available online 21 October 2025

given due attention in the scientific literature to date [33–36].

Based on the data collection of occupational accidents in each country's regulations, the category struck-by injuries includes collisions with equipment, private vehicles, falling materials, vertically lifted materials, horizontally transported materials, and trench collapses [33]. Such a broad category is not a productive approach for the analysis of outcomes at lower levels, as a level of specificity is lost [35].

The study conducted by Thomson (1996) [37] concluded that most struck-by accidents occurred due to a lack of compliance with Occupational Safety and Health Administration (OSHA) regulations and a lack of equipment maintenance. Hinze et al. (2005) [33] analysed struck-by accidents collected in the OSHA database between 1997 and 2000 and concluded that many workplace injuries and fatalities were a direct result of non-compliance with OSHA standards, particularly in the areas of signalling, materials handling, cranes, trenching, and material handling equipment. Other studies that have further broken down the struck-by-accidents category, such as Riaz et al. (2006) [17] using UK accident data, found that the percentage of total injuries caused by contact with machinery and being struck by a moving vehicle was 4 % and 2 % respectively. Zhu et al [6] in their study collected accident data for struck by vehicles, noting that in the USA there were 156 fatalities in 2012 in the private sector, while these values were even higher in public road construction works, accounting for 53 % of accidents, reaching a total of 442 fatal accidents between 2003 and 2010.

Woźniak and Bożena (2024) [38] in their study conducted in the Polish construction sector between 2015 and 2020 obtained that the category "Being run over by a vehicle/hit by a vehicle", which includes collisions in the man-machine, machine-machine and machine-object systems, were the most common dangerous events, with 5.48 % of the total accidents.

In Spain, Camino-López et al. (2008) [39] analysed construction accidents between 1990 and 2000 according to the type of accident, obtaining that 1.2 % of the total accidents were due to the category "Construction site vehicles", but significantly, the percentage of fatal accidents amounted to 15.9 %, with a total of 481 fatalities in the study period.

Based on the category "Being hit and/or run over by vehicles in the construction sector" is not a very well-studied aspect, and no previous studies have been carried out in Spain, together with the high mortality rate of this type of accident when they occur, this article aims to fill this gap in the scientific literature.

The studies analysed show that accidents involving vehicles on construction sites account for a significant proportion of serious and fatal injuries in construction. Run-overs/back-overs and, in particular, the manoeuvre of reversing dumpers and lorries, are recurrently listed as serious accidents in different work areas. As demonstrated in the works of Bunn et al. (2025) [40], Romano and Cassini (2014) [41] and Fan et al. (2019) [42].

The field of predictive modelling in construction safety has undergone substantial growth. A survey of the scientific literature reveals the application of supervised algorithms to predict severity or types of injuries [43–46]. Furthermore, natural language processing (NLP) has been employed to extract causal factors and typologies from accident reports [47]. Moreover, systematic reviews have been published that highlight the potential and acceptance of these techniques in occupational safety [48,49]. Despite the advances witnessed in severity prediction models in occupational safety over the past decade [50–52], significant limitations persist in the realm of accidents caused by vehicles on construction sites. The most important limitation is the marked imbalance between classes (minor accidents versus serious, very serious or fatal accidents), which renders it challenging to detect critical cases reliably [53,54]. In addition to the aforementioned points, the absence of external validation and the reliance on global metrics must be considered. It has been demonstrated that these metrics do not always reflect practical utility in contexts of high case asymmetry [55,56]. Moreover, there is a paucity of studies that compare different approaches according to the domain of the variables of interest. The present study contributes to overcoming these limitations by utilising a comprehensive national register of accidents involving construction vehicles, selecting the most influential predictors of severity, and developing two specific models (organisational and contextual). The integration of machine learning methodologies with balancing strategies and metrics tailored to critical cases facilitates the development of more effective predictive instruments that can be seamlessly integrated into preventive practices, aligning with the guidelines established by international organisations [57–59].

This study aims to explore the evolution of occupational accidents caused by construction site vehicles in the construction industry in Spain for 2009–2022, to analyse the relationship between the main key factors in the severity of the accidents, and to develop a classification model with the most influential variables from a data-driven perspective. For the development of the classification model, the use of neural networks and classification trees has been proposed.

The results of this study could provide employers, workers, legislators, prevention technicians, and researchers with valuable information to improve safety conditions on construction sites, thereby reducing the risk of workers suffering from both minor and serious accidents, as well as fatal accidents, and limiting the social and economic impacts of these incidents.

The rest of the article is structured as follows. Section 2 describes the materials and methods. The results and discussion are presented in Sections 3 and 4, respectively, and finally, Section 5 shows the research conclusions.

2. Materials and methods

2.1. Data scope, source and pre-processing

This research covers all Spanish construction vehicle-related occupational accidents recorded between 2009 and 2022. This is the most recent continuous period for which all national records had undergone administrative validation at the time the data was extracted. Events resulting in medical leave during working hours that are reportable under the law are included in the dataset. Accidents that occurred while commuting and incidents that are not reportable were excluded to ensure the data is comparable and consistent with the regulations.

All case records originate from the Delt@ national electronic declaration platform, which is a statutory system through which employers (or their designated occupational hazard prevention services) must report occupational injuries within legally specified deadlines. The Ministry of Labour provides anonymised, quality-controlled extracts once duplicate identifiers, inconsistent fields (e.g. an age–tenure combination that is not possible) and incomplete mandatory fields have been flagged and reconciled internally. This multi-stage validation process ensures high data completeness and reduces under-reporting bias for vehicle-related events [60].

Construction was defined according to CNAE divisions 41–43 (aligned with the European NACE Rev. 2 classification system) [61], and for more detailed analysis, we also retained a set of 12 aggregated activity subclasses (see Supplementary Table S1 for the classification) to allow for further stratification (e.g. building, civil engineering, and specialised trades).

A common methodological thread in earlier work is the extraction of predictor variables from compulsory occupational injury report forms to explain severity outcomes; this methodology has been adopted in construction studies [39–63], investigations in the metal sector [64], assessments within Andalusian public universities [65], and analyses of mining accidents [66]. A case was classified as vehicle-related by applying a structured filter to the 'contact form'/'mechanism' fields, retaining the collision/impact categories that corresponded to mobile construction plant, site logistics vehicles or material handling equipment interacting with a worker (whether they were standing, walking,

operating or being transported).

The severity of the condition was categorised according to the statutory four-level clinical/functional prognosis, which is divided into the following categories: light, serious, very serious and fatal. In the context of descriptive outputs, the four discrete levels were maintained. However, for modelling, a binarisation process was employed to categorise into non-critical (light) and critical (serious, very serious, and fatal) categories. This approach enables the prioritisation of machine learning capabilities towards the early identification of high-impact cases, thereby mitigating extreme class fragmentation. It should be noted that the percentages of very serious and fatal cases are individually less than 1 %. The overall proportion of critical cases in the study window is 2.5 %, producing a class imbalance ratio of approximately 1:39.

Predictor variables were organised into two analytical domains—(i) organisational (temporal scheduling, workforce and firm attributes) and (ii) contextual (task/site circumstances and injury manifestation)—and into five higher-level categories (time, worker, company, situation, injury consequence). Within these, 12 subgroups (weekday, hour block, shift hour, age group, job tenure, activity code, firm size, usual task, site zone, deviation pattern, injury type, body region) capture both exposure context and outcome morphology. This structure was outlined to separate relatively exogenous site/task conditions from organisational controllables, to enable a specific modelling to test whether contextual information alone can reach operationally relevant sensitivity, and to reduce possible multicollinearity problems in the predictive model development (Table 1).

Initial data quality screening was performed, resulting in the removal of records with missing or implausible key fields (<0.5 % of raw entries). As the number of exposed workers was unavailable, it was not possible to calculate incidence rates. Consequently, analyses were instead based on absolute counts and relative proportions within severity strata.

2.2. Analytical framework

The analytical workflow comprised three sequential phases with these main activities: (1) exploratory data profiling and preparation according to the constraints outlined in subSection 2.1; (2) bivariate association and information-theoretic ranking for predictor selection; and (3) classifier modelling and evaluation. As illustrated in Fig. 1, a simplified schematic is presented.

Phase 1 – Data Curation and Baseline Analysis involves conducting exploratory studies, profiling, and data processing. Descriptive statistics were generated for each subgroup to characterise temporal trends (e.g., the contraction following the financial crisis, the pandemic-related decline) and structural patterns (e.g., changes in age and seniority distribution). The severity share metrics for light (L), serious (S), very serious (VS) and fatal (F) accidents were summarised both overall and stratified by each predictor. This provided a preliminary visual indication of potential differences among each group.

Data quality screening was conducted before analysis. The source dataset was the Spanish official registry of occupational accidents (Delt@, Ministry of Labour), which is mandatory and validated administratively. After filtering for construction-sector cases and accidents involving site vehicles, 16,781 valid records were retained for descriptive statistics. For predictive modelling, additional cleaning steps were taken: records with vague or residual categories (e.g., 'Other', 'Not specified') in key predictors were excluded to enhance interpretability. Duplicates and inconsistent records were checked and found to be minimal (<1 %). No systematic imputation was necessary. The final modelling sample was therefore slightly smaller than the full descriptive dataset. A detailed description of the data screening steps is available in Supplementary Table S2.

During Phase 2 – Statistical Screening, each categorical predictor was subjected to cross-tabulation with the four severity levels and was subsequently assessed through the implementation of chi-square (χ^2) tests of independence. To ascertain which specific category of severity cells were responsible for the significant chi-squared results, corrected standardised residuals (CSR) were examined. Those cells for which $|CSR| \geq 1.96$ were materially contributing at $\alpha = 0.05$, while those for which $|CSR| \geq 2.58$ were deemed to indicate stronger divergence at $\alpha = 0.01$. In instances where ordinal structure was present, such as in age group, job tenure, and shift hour, Somers' D was utilised to quantify the direction and magnitude of the ordered association. The effect sizes for the association tests were also reported using Cramér's V coefficient. This approach complemented the use of χ^2 analysis, which is insensitive to ordering.

To progress beyond the limitations of purely significance-based screening, the mean mutual information (MI) was computed between each predictor and severity (two-level: non-critical and critical, considering light cases as non-critical and the rest as critical). This approach was undertaken to capture the reduction in uncertainty about severity that is provided by the knowledge of the predictor category. Furthermore, a G-statistic (likelihood ratio) was considered for interpretive consistency. Predictors were then subjected to a three-tier ranking system, incorporating statistical significance, MI magnitude, and practical interpretability.

For each categorical predictor X and injury severity Y, mutual information (MI) is a quantitative metric that quantifies the average reduction in uncertainty about Y after observing a category of X. MI is computed as follows (Eq. (1)):

$$MI\ (Xi,\ Y) = \sum_i \sum_j Pij * log \frac{Pij}{Pi * pj} \quad (1)$$

where p_{ij} is the joint proportion for level i of X and severity level j, and p_i, p_j are the corresponding marginals. Higher MI indicates stronger dependence. In order to calculate the G-statistic, it is first necessary to compute the mutual information of each categorical predictor, Xi, with severity. When this value is multiplied by 2Nln2 (Eq. (2)), the

Table 1
Structure of the predictor variables: domain, groups, variables, coding and number of associated categories.

Domain	Group	Variable	Code	Description	#Category
Organisational	Time	Weekday	DOW	Day on which accident occurred	7
		Hour block	HB	One-hour interval (07–20 h)	14
		Shift hour	SH	Position within worker's shift	12
	Worker	Age group	AGE	Worker age bracket	8
		Job tenure	TEN	Time employed on site	8
	Company	Activity code (NACE)	NACE	Economic sector classification	12
		Firm size	FS	Headcount class	7
Contextual	Situation	Usual task	UT	Task classified as usual vs not	2
		Site zone	SZ	Location within site	3
		Deviation pattern (Deviation)	DEV	Abnormal initiating event	10
	Injury Consequences	Injury type	INJ	Lesion diagnostic category	15
		Body region	BR	Main anatomical site injured	9

*Full category labels are provided in Table S1 (Supplementary Information).

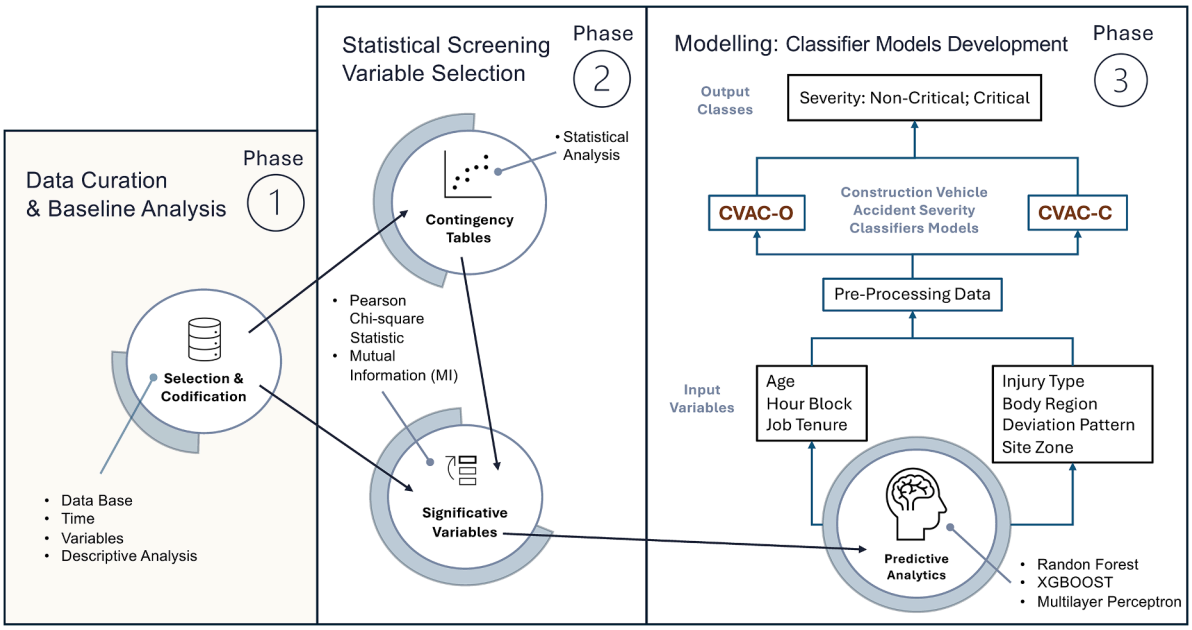

Fig. 1. Research framework. Source: own elaboration.

likelihood-ratio statistic G is obtained, which is used to test independence. It is evident that, under the condition that the expected cell counts are not sparse, G can be adequately approximated by the $\chi 2\,(n-1)*(r-1)$ distribution that possesses (n-1)(r-1) degrees of freedom; here n denotes the number of categories of Xi and r the number of distinct severity outcomes.

$$G = 2 * N * ln2 * MI\ (Xi,\ Y) \tag{2}$$

MI is a measure of symmetry and non-negativity. It is therefore a valuable tool for the analysis of data sets with a range of categories, as it allows for meaningful comparisons between predictors. In particular, MI enables an assessment of the deviation of the joint distribution from independence, thus providing a comprehensive analysis of the relationship between the predictor variable. Combined use of MI and the G-Chi test as a formal contrast is documented in the scientific literature [67]. Variables were retained if the null hypothesis of independence was rejected at $p < 0.05$. This approach allowed us to rank predictors by strength of association with severity and to exclude non-significant variables. The top predictors were identified as the fundamental elements that constituted the core feature sets for the classification models. The remaining variables were retained solely to provide descriptive context, to mitigate the impact of feature noise and the occurrence of overfitting.

Statistical analyses were carried out in IBM SPSS Statistics v31.0 (IBM Corp., Armonk, NY, USA; 2025) [68]

Phase 3 – Modelling includes the development of two classification models for construction accidents involving construction vehicles using three machine learning techniques (Table 2). Three algorithmic families were evaluated: Random Forest, Extreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP). It is evident that these algorithms, exhibiting disparate degrees of capability, are generally efficacious in reducing overfitting and processing large databases. However, notable variations exist in their susceptibility to noise, interpretability, and the capacity to propose non-linear models.

The development of models was undertaken for each of the two domains that had been proposed for the predictor variables. Each domain enables an evaluation of whether its contextual or organisational data alone can provide actionable insights, and what incremental value the predictors add. The first model, Construction Vehicle Accident Severity Classifier – Organisational Factors (CVAC—O), classifies such accidents as either non-critical (minor) or critical (severe, very severe, or fatal), based on time-related, worker, and company group variables identified as statistically significant in terms of accident severity during Phase 2. In contrast, the second model, Construction Vehicle Accident Severity Classifier – Contextual Factors (CVAC—C), was built using the statistically significant variables found for the situation and injury consequences groups, also identified in Phase 2. Of the seven possible variables to be considered in the model CVAC—O, five are numerical and two are categorical. For the CVAC—C model, there are 5 possible variables to consider, all of which are categorical. Once the influential variables to be used in constructing the models have been selected, we work with the chosen numerical variables, processing their numerical data without grouping by field. The selected categorical variables will be processed using the One-Hot Encoding technique [69] to enable them to be used in numerical models.

Although the data have been grouped into two classes (No-Critical and Critical), the number of Critical class records is significantly lower than the number of No-Critical class records, resulting in an unbalanced data set. To address the severe class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) algorithm [70] was employed, which synthesises additional minority observations by interpolating in

Table 2
Comparison of construction vehicle accident classification models.

Feature	CVAC—O (Organizational Factors)	CVAC—C (Contextual Factors)
Full Name	Construction Vehicle Accident Classifier Organizational Factors	Construction Vehicle Accident Classifier Contextual Factors
Input Variable Groups	Temporal, Personal & Business	Accident Consequences & Circumstances
Output Classes	Non-critical (minor)- Critical (severe, very severe, fatal)	
Machine Learning Techniques Used	Random Forest, XGBOOST, Multilayer Perceptron (MLP)	
Main Objective	Identify severity based on internal & structural factors	Identify severity based on situational & impact factors
Derived From	Statistically significant variables in organizational domains (Phase 2)	Statistically significant variables in contextual domains (Phase 2)

feature space between each minority point and selected nearest neighbours. The expansion of the minority class in this manner serves to limit the dominance of the majority class, thereby facilitating the development of more comprehensive and less overfit decision boundaries by the classifier. The development of the models was centred on the reliable identification of critical accident cases.

The modelling of the supervised learning models utilised KNIME (Konstanz Information Miner. Version 5.5.0) [71], with its 5.0.0 open-source Analytics Platform release [72] providing a visual, low/no-code interface that has seen extensive uptake across multiple scientific disciplines (biology, chemistry, engineering, medicine, cybersecurity) [73].. The data set was partitioned into two distinct sets: 70 % was assigned for model training, while the remaining 30 % was reserved for the validation of the model's performance. Categorical variables were transformed into dummy variables using the One to Many (PMML) node in KNIME. To prevent data leakage, the transformation was fitted exclusively on the training sample after the 70/30 data partition. The resulting PMML model was then applied to the validation/test data using the PMML Transformation Apply node, ensuring identical dummy-variable structures between both subsets (Figure S1). This approach avoids incorporating information from the validation set into the model during training, thereby preserving model generalizability and avoiding overly optimistic performance estimates [74]. The missing values were not incorporated into the model. One-hot encoding produces n dummy variables for a categorical variable with n categories, which can introduce redundancy into the prediction model. This redundancy does not lead to overfitting in the models used. Although tree-based models such as Random Forest can tolerate multicollinearity effectively, as they can select relevant features and ignore redundant ones during training and empirical comparisons with linear models in high-multicollinearity datasets support this [75], we retained (n-1) dummy variables without compromising model performance. The SMOTE technique [76] was exclusively implemented on the training sample. SMOTE node was configured with five nearest neighbours, oversampled the minority Class (the Critical Class) with fixed random seeds set to 0, and applied it only to the train branch (after encoding). The SMOTE parameters are also provided in Supplementary Table S2.

The optimisation of the model was conducted using Parameter Optimisation Loop nodes. The optimisation hyperparameters for each model are presented in Table 3.

Model hyperparameters were optimised once on the full 70 % balanced training partition. The resulting configuration was then evaluated with k-fold cross-validation within the training set to assess stability and mitigate potential overfitting. After fixing the optimal configuration, models were retrained on the entire training partition. Their generalisation performance was finally assessed on the untouched 30 % external test set, which was not used at any stage of training or parameter tuning. To further check robustness, model performance was also calculated on the unbalanced version of the 70 % training sample, to confirm that results will be consistent across balanced and unbalanced conditions. This sequential two-step strategy (internal cross-validation for stability and external test set for generalisation) prioritises computational efficiency, preserves an unbiased test set for final evaluation, and is consistent with recommended practices in predictive modelling [77].

Table 3
Optimisation parameters for building the classification model.

Machine Learning Technique	Parameter	Value
Random Forest	tree depth	6 - 10
Extreme Gradient Boosting (XGBoost)	eta	0.1 - 0.3
	maximum depth	4 - 6
	minimum depth	1 - 4
Multilayer Perceptron (MLP)	Hidden layer	1 - 2
	number of hidden neuronal	6 - 12
	Activation function	sigmoid

Supplementary Figure S1 shows the layout of the nodes used in the construction of the neural network classification models with the KNIME tool. Supplementary Table S2 details node configurations. To guarantee reproducibility, Partitioning, SMOTE, Random Forest, XGBoost, MLP Learner, and X-Partitioner nodes were configured with fixed random seeds. The default seed value was set to 0 across the workflow, ensuring identical results upon independent execution. This is also provided in Supplementary Table S3

Model evaluation was carried out on the untouched 30 % external test set. The primary selection criterion was the absolute number of true positives in the critical severity class, given its greater preventive relevance. In cases where models achieved the same number of critical detections, the F2-score for the critical class was used, since this metric emphasises recall over precision. To ensure balanced performance across both classes, we also defined a Global Performance Score (GPS) (Eq. (3)) [78], calculated as the harmonic mean of the F2-scores for the critical (F2Cc) and non-critical classes(F2NCc) (Eqs. (4) and 5). This combination of criteria prioritises the reliable detection of severe accidents while maintaining overall model robustness. This global index GPS is based on the four terms of the confusion matrix (true positives (TP), true negatives (TN), false positives (FN) and false negatives (FN))

$$GPS = \frac{2}{\frac{1}{F2Cc} + \frac{1}{F2NCc}} \tag{3}$$

Where F2 is:

$$F2Cc = \frac{5 * TP}{5 * TP + FP + 4 * FN} \tag{4}$$

$$F2NCc = \frac{5 * TN}{5 * TN + 4 * FP + FN} \tag{5}$$

3. Results

3.1. Data curation & baseline analysis

A total of 16,781 accidents by construction site vehicles occurred in Spain between 2009 and 2022, 15,845 accidents in men (94.4 %) and 936 in women (5.6 %). Fig. 2 presents a consolidated time series, illustrating both the aggregate number of construction vehicle-related accidents and the subset resulting in fatalities across the study period. Firstly, the total accidents decreased to the minimum values between 2012 and 2014, coinciding with the peak of the Spanish economic crisis. Subsequently, there is a progressive increase in the number of accidents, which was only interrupted in 2020 as a result of the decrease in activity caused by the confinement due to the COVID 2019 pandemic. As for the number of fatal accidents, there is no clear trend. Of the 86 fatalities, only 2 were women.

Accidents by construction site vehicles mostly involve Spanish workers, both in minor accidents (15,128 accidents; 90.1 % of the total) and fatal accidents (74 fatalities; 86 % of the total). With very low values, the second nationality in terms of minor accidents is Romanian (320 accidents; 1.9 %), and in terms of fatal accidents, it is Moroccan (4 deaths; 4.7 %).

Most vehicle-related accident cases occurred while workers were moving between work areas (≈81 % of light events; ≈85 % of fatal cases), whereas only about one in six total cases originated at the usual workstation, and <3 % in other locations

Figs. 3 and 4 show a summary of accidents for the variables Age group, Hour block, Firm Size, Job tenure, Deviation pattern, Body region, Workplace Area, and Site zone based on severity

3.2. Statistical screening: variable selection

This subsection presents the analyses of the χ^2 contingency tables, which were employed in the study to outline the relationships between

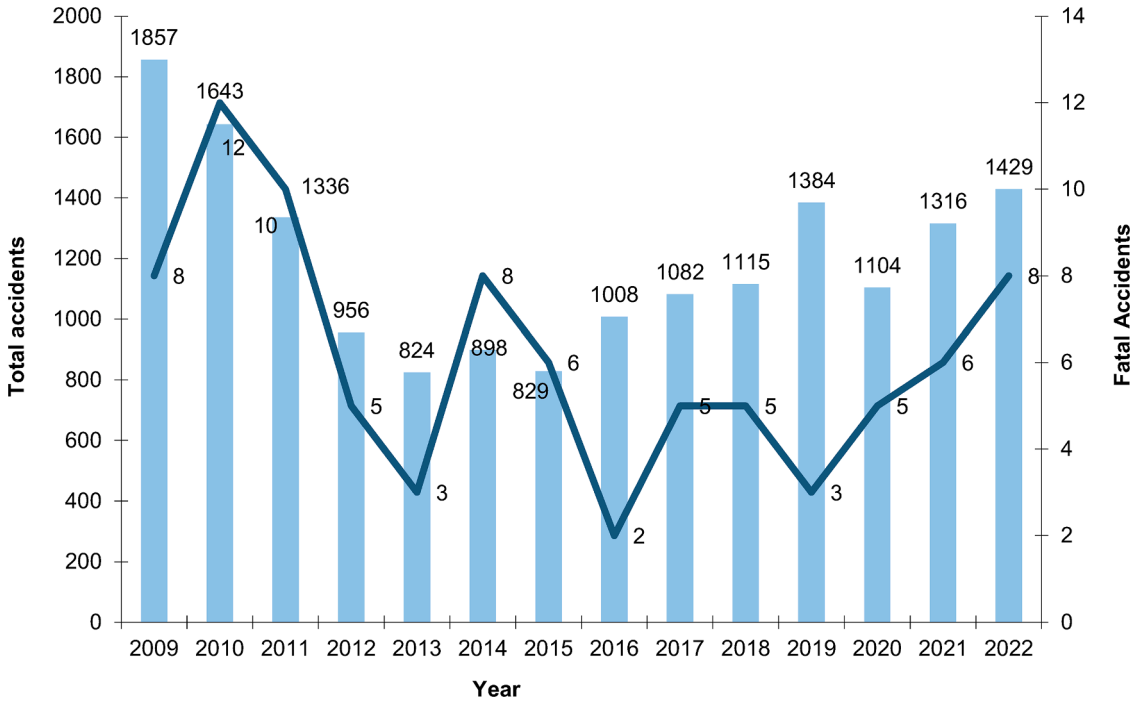

Fig. 2. Annual evolution of construction vehicle-related accidents and associated fatalities (2009–2022). Own elaboration.

the study variables and accident severity. The importance of each variable as a predictor of accident severity was quantified using the results of the contingency tables and Mutual Information values (MI) for each analysed variable. These results provide support for each of the variables that were selected as potential predictors of accident severity

3.2.1. Time variables

This subsection presents the analysis of the variables: weekday, hour block, and shift hour

Table S4 (Supplementary Information) shows the contingency table for the variable weekday and severity. Results indicate that there is no dependent relationship between them ($\chi 2 = 21.613$, df = 18 and $p = 0.25$). The effect size for the association tests was low (Cramér's V coefficient=0.02). Therefore, the day of the week cannot be considered a predictor of accident severity. There is no difference in the number of accidents, both total and fatal, depending on the day of the week. However, there is a considerable decrease during the weekend, which is typically a non-working day on most construction sites.

The analysis of the contingency table (Table S5) indicates a dependent relationship between the hour block and severity variables. This relationship was found to be statistically significant ($\chi 2 = 78.168$, df = 0 39, and $p < 0.001$; Someŕs D coefficient =−0001; Cramér's V coefficient=0.04). The hour block and severity are related; however, only 10 corrected standardised residuals (17.86 %) were significant at the 95 % confidence level ($p > 1.96$).

The highest number of related vehicle constructions is in the morning, from 9:00 to 13:59 hours (52.5 % Tn; 52.9 % Ln; 44.3 % Ls; 47.4 % VSn; 29.0 % Fn). In addition, the period between 18:00 and 18:59 shows the highest percentage of critical accidents (6.7 % Sn, 10.5 % VSn and 15.1 % Fn).

Finally, the shift hour variable was analysed. As Table S6 shows, there is no dependent relationship between shift hour and severity ($\chi 2 = 40.312$, df = 33, and $p = 0.178$; Cramér's V coefficient=0.03). Accident severity reveals significant fluctuations based on the specific time of the shift hour variable. The highest number of total, minor, and severe accidents occurred at the start of the working day (less than an hour and in the first hour; 31.7 % Tn; 31.8 % Ln; 30.6 % Sn), and the highest number of fatalities occurred in the fourth and first hour of the shift hour (15.1 % and 14.0 % respectively).

3.2.2. Worker variables

The contingency analysis reveals a correlation between accident severity and the age of the workforce ($\chi^2 = 136.230$, df =33, $p < 0.001$; Somers' D coefficient = 0.016) with 16 corrected standardised residuals (50.00 %) significant at the 95 % confidence level ($p > 1.96$). Cramér's V coefficient=0.05), as shown in Table S7. The age groups with the highest accident rates are between 30 and 39 (Tn 34.3 %) and 40 and 49 (Tn 27.8 %). Notably, the highest number of fatalities is observed among the age group between 40 and 49 (Fn 45.3 %) followed by the age group between 50 and 59 (Fn 27.9 %), and if the age ranges 30 to 39 and 50 to 59 are grouped with it, 86 % of fatal accidents are concentrated.

The job tenure variable refers to the duration of time a worker has been employed by a company, rather than their overall industry. The contingency table (Table S8) shows a correlation between accident severity and job tenure ($\chi^2 = 48.176$, df = 21, and $p < 0.001$; Someŕs D coefficient =0002; Cramér's V coefficient=0.03), with 8 corrected standardised residuals (25 %) reaching significance at the 95 % confidence level ($p > 1.96$). The distribution of job tenure and severity is non-uniform. The highest accident rates occur when workers have less seniority in the company. The group of workers with less than one year of experience account for 41.2 % of total accidents (Tn), 41.1 % of minor accidents (Ln), 45.8 % of serious accidents (Sn), 36.8 % of very serious accidents (VSn) and 45.4 % of fatalities (Fn). It is noteworthy that the rate of fatal accidents among workers who have been with the company for less than a month is 12 %, a figure which is considerably higher than the rate of minor accidents for this same group.

3.2.3. Company variables

This subsection includes the variables activity code (NACE) and firm size. The firm size variable was analysed. Table S9 presents the contingency analysis, showing no correlation between accident severity and firm size ($\chi 2 = 23.172$, df = 18 and $p = 0.184$; Cramér's V coefficient=0.02). The highest accident percentage has been recorded among companies employing fewer than five workers (31.7 % Tn; 31.8 % Ln; 30.6 % Sn; 21.1 % VSn; 23.3 % Fn), which are the most characteristic companies in the Spanish construction sector. Following this

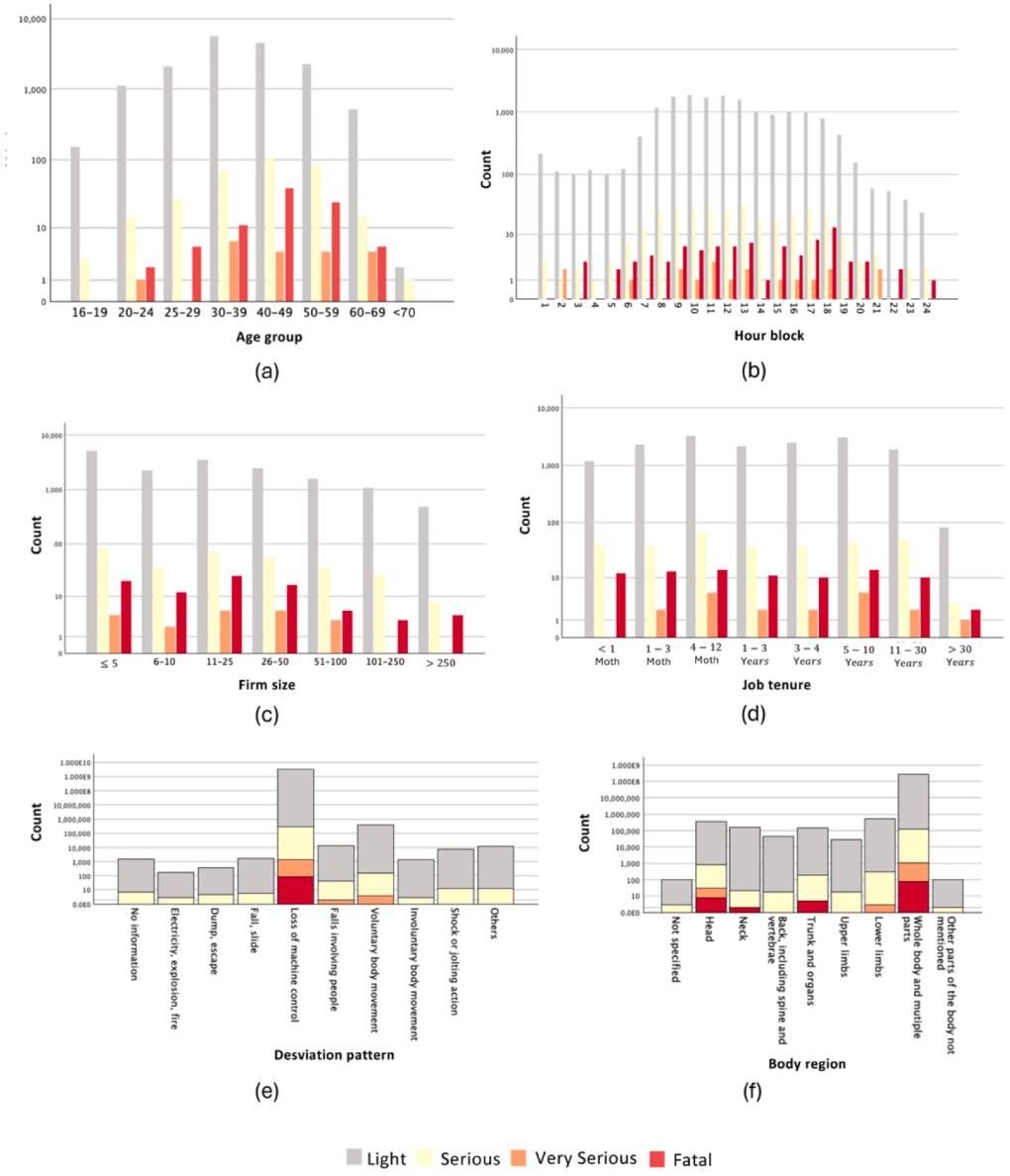

Fig. 3. Number of accidents by severity and age group, time slot, company size, length of service, deviation pattern, and body area variables (bar charts). Own elaboration.

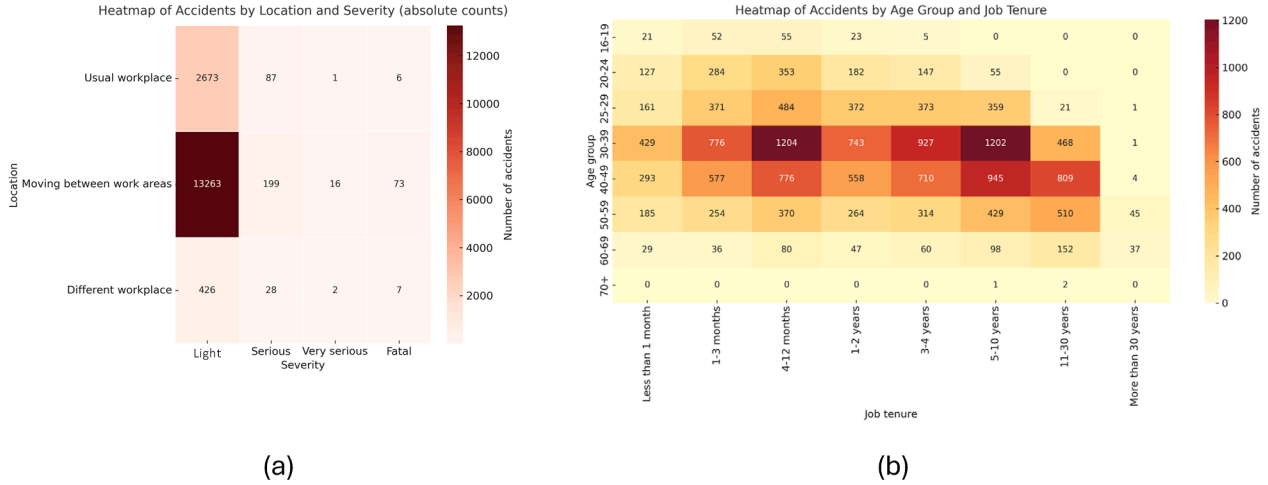

Fig. 4. Accidents by Site zone and severity (a); age group and Job tenure by accident severity (b) (heatmap diagrams). Own elaboration.

range, the range with the highest values is between 11 and 25 workers, and between 26 and 50 workers. These data are remarkable because these companies, in theory, have more resources for the organisation of occupational risk prevention.

3.2.4. Situation variables

This subsection presents the analysis of the variables: task, site zone, and deviation pattern. Firstly, the results of severity are presented concerning the task variable. The contingency table indicates no dependence between accident severity and site zone ($\chi2 = 0.382$, df = 6 and $p = 0.944$; Cramér's V coefficient=0.01). Most accidents occur at the usual work (90.8 % Tn; 90.8 % Ln; 91.7 % Sn; 89.5 % VSn; 90.7 % Fn).

Secondly, the relationship between site zone and severity variables is analysed. Both variables are related ($\chi2 = 100.278$, df = 6 and $p < 0.000$; Somer's D coefficient =−0001; Cramér's V coefficient=0.06) (Table S10), with only 3 corrected standardised residuals (25 %) being significant at the 95 % confidence level ($p > 1.96$).

Finally, in the group of situation variables, the relationship between severity and deviation pattern is analysed. Both variables are related but not strongly enough to be considered for predicting severity ($\chi2 = 89.289$, df = 27 and $p < 0.000$; Somer's D coefficient =−0009; Cramér's V coefficient=0.04) (Table S11), with 10 corrected standardised residuals (27.78 %) significant at the 95 % confidence level ($p > 1.96$). In most accidents, the deviation, i.e. the unusual event that has negatively impacted the normal work process and has resulted in the occurrence or cause of the accident, that generates them (Ln 67.0 %; Sn 67.8 %; VSn 78.9 %; Fn 100 %) is found under the heading 'Loss of machine control'. The second group corresponds to 'Voluntary body movement' (Ln 15.1 %; Sn 12.1 %; VSn 15.8 %).

3.2.5. Injury consequences variables

In this subsection, the severity of accidents is analysed, focusing on the variables of injury type and body region.

Injury type and severity variables present a dependent relationship between accident severity and injury ($\chi2 = 2412.624$, df = 21and $p < 0.001$; Somer's D coefficient =0.019; Cramér's V coefficient=0.22) (Table S10). A total of 20 corrected standardised residuals (62.5 %) are significant at the 95 % confidence level ($p > 1.96$). It can be concluded that injury type and severity are related. As shown in Table S12, there are fifteen categories in this variable, but there are no accidents reported in seven categories The highest number of minor injuries is found within the group 'Dislocations, sprains and strains' (Ln 55.1 %) followed by the group 'Wounds, superficial injuries' (Ln 22.8 %). Serious and very serious injuries fall into the groups 'Multiple lesions' (SAR 27.7 %; VSAR 57.9 %) and 'Crushed bones' (SAR 53.5 %; VSAR 21.1 %), while most occupational fatalities fall into the group 'Multiple lesions' (FAR 87.2 %, 75 deaths).

In Table S13, results indicate a correlation between accident severity and body region ($\chi2 = 876.139$ df = 24, and $p < 0.001$; Somer's D coefficient =0.033; Cramér's V coefficient=0.20). 23 corrected standardised residuals (63.89 %) are significant at the 95 % confidence level ($p > 1.96$). Among the fatal accidents (Fn) 74 % are in the group "Whole body and multiple parts". The second group, far behind with 7 fatalities (8.1 %) is the "Head". With regard to light accidents, the "Neck" group includes the highest number of minor accidents as a result of being on construction site vehicles (43.7 % Ln).

As described in the methodology section, the amount of information each variable contains about the severity variable was calculated using the mutual information (MI) value. The variables Hour Block, Age, Age, Site Zone, Deviation Pattern, Injury Type and Body Region were selected. These variables exhibit the highest MI values, along with a significant p-value ($p < 0.05$) (Table 4).

Table 4
Mutual information values (MI), G-statistic and p-value for the selection of influential variables. The selected variables are in bold.

Variable	MI	G-stat	p-value
Weekday	0.0006	10.0619	0.1221
Hour Block	0.0011	20.7030	**0.0021**
Shift Hour	0.0010	17.3791	0.0664
Age	0.0040	73.0876	**< 0.001**
Company Staff	0.0006	11.3127	0.0792
Job Tenure	0.0008	13.6876	**0.0333**
Usual Work	0.0000	0.6970	0.4038
Site Zone	0.0018	32.2438	**< 0.001**
Deviation Pattern	0.0017	30.9322	**0.0001**
Injury Type	0.0464	840.7435	**< 0.001**
Body Region	0.0248	450.6054	**< 0.001**

3.3. Modelling: classifier model development

The classification models described in the previous section have been implemented using the KNIME Analytics Platform. Each model contains 38 nodes. Random Forests, XGBoost Tree and RProp MLP Learner nodes were used to develop the random forest, XGBoost and neuronal network models (see Supplementary Figure S1). The CVAC—O model incorporates the variables identified as significant in phase 2: Hour Block, Age, and Age. Conversely, the CVAC—C model incorporates the remaining significant variables identified in the preceding phase: Site Zone, Deviation Pattern, Injury Type, and Body Region. A total of 120 classification models were developed. The selection of the most appropriate model for the data was determined by utilising 70 % of the samples, as described in the methodology section. The model with the highest number of true positives in the critical severity class was selected. The F2 score for the critical class, and the GPS coefficient values were also considered. The model with the highest number of true positive samples of the Critical class was considered for models with the same GPS index value. The validation of the model was confirmed using the remaining 30 % of the samples, confirming the fulfilment of the condition mentioned above. Table 5 shows the optimal model parameters for each of the techniques used.

Table 6 presents the confusion matrix values of the optimal CVAC—O models for each of the three machine learning techniques, evaluated under three conditions: (i) the 70 % training sample without oversampling, (ii) k-fold cross-validation within the 70 % training data with oversampling of the minority class through synthetic instance generation, and (iii) the independent 30 % external test set. After thoroughly evaluating the available options, the neural network technique has been selected as the optimal approach for the CVAC—O model. This decision is supported by the independent 30 % external test set, which demonstrates the highest value for the true positives in the critical severity class as well as for critical class F2 scores and the GPS index.

RProp MLP Learner node implements the RPROP algorithm, using multilayer feed-forward networks (MLP) with a sigmoid activation function. It adjusts the size of the backpropagation steps independently by weight based only on the sign of the gradient. The final output of the

Table 5
Optimal parameters selected for each machine learning technique and classification model.

Machine Learning Technique	Parameter	CVAC—C Model	CVAC—O Model
Random Forest	Tree depth	7	9
XGBOOST	Eta	0.10	0.2
	Maximum depth	4	6
	Minimum depth	1	4
Multilayer Perceptron (MLP)	Hidden layer	1	1
	Number of hidden neuronal	8	10
	Activation function	Sigmoid	

Table 6
Confusion matrix, F2- score and GPS index values for the optimal CVAC—O models applying each of the three machine learning techniques.

Construction Vehicle Accident Classifier Organizational Factors (CVAC—O)								
Machine Learning Technique	**Random Forest**	**Sample**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	7242	1730	80.72 %	0.693 ± 0.337	0.242 ± 0.02
			Critical	162	89	35.46 %	0.143 ± 0.03	
			Overall	97.28 %	2.72 %	79.49 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	3004	841	78.13 %	0.814	0.224
			Critical	74	34	31.48 %	0.130	
			Overall	97.27 %	2.73 %	76.85 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	7119	1853	79.35 %	0.825	0.358
			Critical	114	137	54.58 %	0.229	
			Overall	97.28 %	2.72 %	78.67 %		
	HGBOOST	**Sampling**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	8896	76	99.15 %	0.988 ± 0.002	0.079± 0.02
			Critical	250	1	0.40 %	0.042± 0.01	
			Overall	97.28 %	2.72 %	96.47 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	3822	23	99.40 %	0.990	0.063
			Critical	105	3	2.78 %	0.033	
			Overall	97.27 %	2.73 %	96.76 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	8926	46	99.49 %	0.991	0.284
			Critical	215	36	14.34 %	0.166	
			Overall	97.28 %	2.72 %	97.17 %		
	Neural Networks	**Sample**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	5503	3469	61.34 %	0.663 ± 0.04	0.262 ± 0.02
			Critical	101	150	**59.76 %**	0.163± 0.02	
			Overall	97.28 %	2.72 %	61.29 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	2498	1347	64.97 %	0.663	0.278
			Critical	44	64	**59.26 %**	0.174	
			Overall	97.27 %	2.73 %	64.81 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	5712	3456	62.30 %	0.672	0.267
			Critical	97	154	61.35 %	0.167	
			Overall	97.34 %	2.66 %	62.28 %		

model with a single hidden layer is expressed by Eq. (8).

$$y = \sigma\left(\sum_{j=1}^{m} w_j^{(2)} * h_j + b^{(2)}\right) \tag{8}$$

where m is the number of neurons in the hidden layer $w^{(2)}{}_j$ is the weight connecting the hidden neuron h_j to the output neuron, $b^{(2)}$: bias in the output neuron, σ is the sigmoid activation function (Eq. (9)) and h_j is the value of the activation wing of the hidden neurons (Eq. (10)).

$$\sigma(z) = \frac{1}{1 + e^{(-z)}} \tag{9}$$

$$h_j = \sigma\left(\sum_{i=1}^{n} w_{ij}^{(1)} * x_i + b_j^{(1)}\right) \tag{10}$$

In Eq. (10), n denotes the number of input variables, $w^{(1)}{}_{ij}$ is the weight connecting the input x_i with the hidden neuron h_j, and $b_j^{(1)}$ is the bias of the neuron h_j.

The parameters (weights and biases) of the neural network of the CVAC—C classification model are documented in Tables S14 and S15. Tables S16 and S17 show weights and biases parameters for the CVAC—O neural network classification model.

The values of the confusion matrix for the optimal CVAC—C models of each technique are presented in Table 7. Considering the model selection criteria explained above, the multilayer perceptron model has also been selected as the best predictor model for severity.

For the neural network models, predicted class probabilities of the external validation sample (30 %) were extracted to analyse confidence in critical case predictions. Probability distributions were plotted to visualise the separation between critical and non-critical cases (Fig. 5). These analyses allow assessing discrimination of the neural networks beyond standard accuracy metrics. For classification purposes, a probability threshold of 0.50 was applied: cases with a predicted probability of at least 0.50 were assigned to the critical class, while those with a predicted probability of less than 0.50 were assigned to the non-critical class.

To calculate the normalised independent variable importance of each as predictors in the multilayer perceptron models (Fig. 6), the raw importance was first calculated as the sum of the product of the weights of each neuron along the paths linking each input to the output(s) and then rescaled so that the most influential predictor took the value 100 %. This metric allows variables to be ordered by relative influence within the network [79,80]. The categorical predictor variables with n levels have been expanded to n dummy variables. The normalised importance of these variables has been calculated by summing the values obtained for the importance of each of their dummy variables

Table 7
Confusion matrix, F2- Score and GPS index values for the optimal CVAC—C models applying each of the three machine learning techniques.

Construction Vehicle Accident Classifier Contextual Factors (CVAC—C)								
Machine Learning Technique	**Random Forest**	**Sample**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	7831	1141	87.28 %	0.913 ± 0.006	0.625 ± 0.001
			Critical	41	210	83.67 %	0.478 ± 0.011	
			Overall	97.28 %	2.72 %	87.18 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	3326	519	86.50 %	0.888	0.593
			Critical	15	93	86.11 %	0.445	
			Overall	97.27 %	2.73 %	86.49 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	7767	1205	86.57 %	0.889	0.608
			Critical	26	225	89.64 %	0.462	
			Overall	97.28 %	2.72 %	86.65 %		
	HGBOOST	**Sampling**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	8236	736	91.80 %	0.933 ± 0.003	0.627 ± 0.010
			Critical	66	185	73.71 %	0.481± 0.008	
			Overall	97.28 %	2.72 %	91.30 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	3535	310	91.94 %	0.932	0.630
			Critical	31	77	71.30 %	0.477	
			Overall	97.27 %	2.73 %	91.37 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	8271	701	92.19 %	0.932	0.638
			Critical	53	198	78.88 %	0.498	
			Overall	97.28 %	2.72 %	91.82 %		
	Neural Networks	**Sample**	**Observed**	**Predicted No Critical**	**Critical**	**Percent Correct**	**F2-Score (Mean ± SD)**	**GPS (Mean ± SD)**
		Training (70 %) Cross Validation	**No-Critical**	7986	986	89.01 %	0.912 ± 0.006	0.639 ± 0.007
			Critical	38	213	**84.86 %**	0.492± 0.010	
			Overall	97.28 %	2.72 %	88.90 %	**F2- Score**	**GPS**
		External Testing (30 %)	**No-Critical**	3432	413	89.26 %	0.909	**0.646**
			Critical	14	94	**87.04 %**	**0.501**	
			Overall	97.27 %	2.73 %	89.20 %	**F2- Score**	**GPS**
		Training (70 %) No Oversample	**No-Critical**	8004	968	89.21 %	0.911	0.664
			Critical	21	230	91.63 %	0.522	
			Overall	97.28 %	2.72 %	89.28 %		

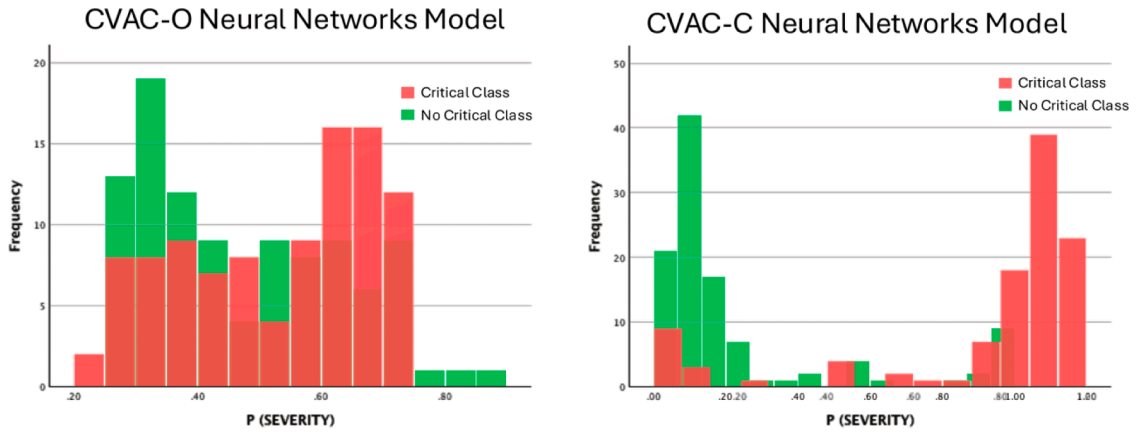

Fig. 5. Predicted probability distributions of the 30 % external validation sample for neural network models: CVAC—O (left) and CVAC—C (right): own elaboration.

4. Discussion

Hit and run over by construction site vehicles in the construction sector are mostly light accidents but are one of the most important causes of fatal accidents in the construction sector in Spain, as is the case in other countries [6,31,32]. Examining the evolution of the data during the study period reveals a decrease in the accident rate during the economic crisis years (2012–2014) and during the confinement resulting from the COVID-19 pandemic. This trend underscores the significance of the construction sector to the economy and its impact on

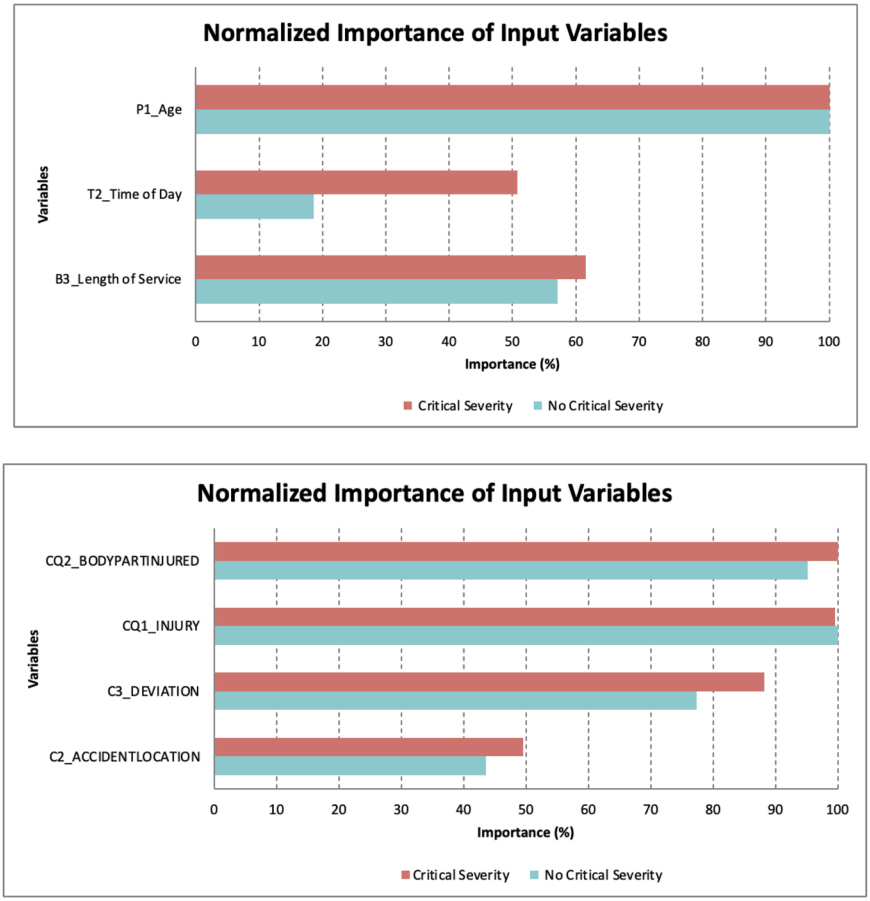

Fig. 6. Normalised importance values of the variables in the model CVAC-0 (a) and CVAC—C (b): own elaboration.

occupational accident rates.

No correlation has been made between the number of employees in the construction sector and the number of accidents per year, and by the autonomous community. Still, in global terms, it can be affirmed that there is a relationship between the level of activity in the sector and the number of employees [81,82] and the accidents, i.e., the greater number of employees at the construction sector, the more hit and run over by vehicles tend to occur.

Hour block, age, job tenure, site zone, deviation pattern, injury type, and body region are significant in the severity of accidents at the 95 % confidence level for the accidents by construction site vehicles in the construction sector in the analysis of the consistency tables. These variables, together with the variable shift hour, have the highest values of the Mutual Inflation (MI) report. All of them, except the variable shift hour, were selected for the development of the predictive models. The shift hour variable was not selected as the G-stat was not significant.

Regarding the results concerning the time variables, there is no influence of the workday on the severity of the accidents, so it is not possible to speak about the "weekday effect" or "Monday effect" described by other authors in Spain [83] and other countries [84,85]. Nor is there a relationship between the severity and the shift hour variable. The number of total accidents decreases as the hours of the workday pass, contrary to expectations and associated with workers' accumulated fatigue, results like those obtained in other studies [86,87].

Regarding the hour block variable, accident occurrence was most dominant around 9:00 and 13:59, consistent with the injuries/fatalities data of most construction found in other studies [16,33]. On the other hand, in the predictive CVC—O model, this variable has the least influence. An interesting point is that in the predictive model, the influence of this variable is greater for critical accidents than for non-critical accidents, which is in line with the results obtained by Hinze.

Construction site vehicle accidents arise mostly in men and, in the case of fatalities, almost entirely in men. These results are consistent with the distribution of the working population by gender in the Spanish construction sector. Of the 16,781 accident cases analysed, 94.4 % correspond to men and 5.6 % to women. This distribution mirrors the composition of the Spanish construction workforce, where women represent only around 11 % of total employment in the sector [88]. The predominance of male accident cases, therefore, reflects differential exposure, not necessarily a higher individual probability of injury by sex

Non-fatal accidents were most common among workers aged 30–49, whereas fatal cases were concentrated between ages 40 and 59. These results align with the findings of other studies in the construction sector in the USA [16] and other productive sectors such as the metal sector in Spain [64]. The number of accidents among workers under 30 years of age is lower, contrary to the results of other studies specific to the construction sector [42,65]. This could be explained by several factors: the better physical conditions due to the age of younger workers [89, 90], the difficulty of older workers to adjust to new roles, changing job circumstances [91], or overconfidence stemming from age and experience related to the position [92,93]. The higher share of severe accidents among older workers could reflect factors such as task assignment, physical condition, or risk perception; however, these explanations cannot be confirmed with the present dataset. Age is the most influential variable in the CVC—O model.

Within the construction sector, the sub-sector Building construction and Electrical, plumbing and other installations on construction sites present the highest number of total and light accidents. Regarding the fatal accidents, in addition to the subsector Building construction, the Demolition and site preparation. This data could be used to enforce

training requirements for workers in companies operating in these specific sub-sectors [16,94].

There is no influence of the company staff regarding the severity of the accident, but the highest number of accidents is observed in companies with fewer than five employees, with the next highest number of accidents occurring in companies with 11 to 25 employees. These results are in line with other studies in Spain's construction sector [39,63] and demonstrate that the lack of resources in small enterprises makes it more difficult to adhere to occupational safety and health regulations [14,16], leading to less safe working environments that heighten the likelihood of accidents occurring [95].

The results obtained for fatalities in the construction sector concerning length of service show that approximately between 40–45 % of the total and fatal accidents occur to workers with less than one year of experience in the company. These results are consistent with other specific works in Spain, such as the one on the metal sector in the autonomous community of Andalusia by Carrillo-Castrillo et al. (2016) [96] or the one on fatal accidents in Spain by Fuentes-Bargues et al. (2023) [97]. Carrillo-Castrillo et al. (2016) [96] suggested these situations are caused by unsuitable work methodologies or an absence of sufficient training. To address this issue, the introduction of a more proactive information policy, as well as training on occupational safety, would be a valuable addition, even before the start of activities within the company [98]. This variable has significant importance in the predictive CVC—O model.

The analysis by accident location and whether the workers were carrying out their usual work shows that the great majority of accidents involved workers carrying out their usual work and moving between work areas. The continuous changes on construction sites, and if on top of that changes in the work areas entail both internal and external movements on the construction sites, generate a lack of appreciation of dangerous situations on the part of the workers [16]. These situations could be solved by greater involvement of the supervisors, site managers, site managers, preventive resources, etc [14] and by the introduction of technology that allows remote location and monitoring for simple proximity alerts between teams and workers on construction sites. Examples of these systems are the one proposed by Riaz et al [17], which proposes combining global positioning systems (GPS) with intelligent sensors and wireless networks, or the one proposed by Zhu et al., combining radio frequency techniques (RFT), ultra-wideband (UWB) and global positioning systems. This variable has the least influence on the CVAC—C model. The influence of this variable is greater in critical accidents. If we examine the contribution of each dummy variable in the normalised significance, it becomes apparent that the variable with the greatest contribution corresponds to the usual work (accident reported in the usual workplace), which also coincides with the scientific work of Hinze and Romero cited above.

The main deviation that leads to an accident is "Loss of machine control", with 67.7 % of the total accidents and 86.0 % of the fatal accidents. This result is perhaps predictable due to the type of accident related to vehicles and work equipment, but it is like other results about fatal accidents, such as the work of Santos et al. in Portugal [99] and Fuentes-Bargues et al. in Spain [64]. This variable is important in the CVAC—C predictive model. It has a greater influence on determining whether an accident is critical. The dummy variable that contributes most to its normalised importance is "Loss of machine control". This result aligns the predictive model with the studies mentioned above.

When the consequence of the accident is the death of the worker, the injuries are found within the group "Multiple injuries", and the injured body part is identified within the group "Whole body and multiple parts" (74.0 %). In the case of light accidents, injuries corresponding to "Dislocations, sprains and strains" and "Wounds, superficial injuries" account for 76.1 %. These results are like the results obtained globally in the construction sector in Spain [14,62] and in other sectors, such as the mechanical sector [100–103].

In light accidents, the body part injured with the highest percentage is the "Neck" group (44.7 %), which is associated with bumps and jolts because of running over or driving vehicles. The group of variables associated with the consequences of accidents show the highest normalised significance values of the CVAC—C model. The contribution to the normalised significance of the dummy variables of these variables to the determination of the accident as critical or non-critical is also in line with the results of this research and the studies described above.

The dataset does not include exposure data (e.g., hours worked, machine operation time) or detailed worker characteristics (e.g., health status, training), which prevents causal inference. Results should therefore be interpreted as associations rather than causal effects

The limitations of the machine learning models developed are linked to the data itself and to the modelling methodology of each of the techniques chosen. The starting data are very unbalanced (ratio of critical samples to non-critical ones is low) and are limited., The primary limitations and challenges of the research include the absence of individual data in accident reports for certain variables that could provide statistically significant explanatory power, such as completed education level, workers' educational background, training and experience on the job, and health and safety training. Additionally, a further limitation is that specific incidence rates cannot be calculated for each variable examined due to the unavailability of data on hours worked. To address the overfitting issues, the number of samples in the critical case has been increased, and non-linear techniques, such as neural networks, have been incorporated. Several scientific studies utilise SMOTE as a valid technique to circumvent the aforementioned limitations [104–106]. In this instance, the data has been divided into two samples (70 %−30 %) to train the model with a balanced sample and then validate its behaviour with a sample of 30 % of the data (unbalanced sample). Increasing the number of cases in the minority sample (training sample), all three techniques used have yielded good results. The two-step approach to mitigate risks of variance increase, or overfitting due to SMOTE, oversampling, applying SMOTE exclusively to the training set (70 %), and validating on a separate imbalanced test set (30 %), together with to compare across three algorithms, and applying to the training sample (70 %) without applying SMOTE, confirms that results are consistent across balanced and unbalanced conditions. obtaining consistent predictive patterns (see Tables 6 and 7). This approach could reduce the likelihood that the results are artefacts of oversampling. However, in general, the number of true positive cases has decreased with the validation sample. This decrease was significant for the HGBOOST technique, and only slight for the Random Forest technique, with the neural network models demonstrating the best performance. The outcome achieved with the trained model using the validation sample is attributable to how SMOTE balances the initial sample [107]. As Moreno-Torres (2012) [54] explained, the model trained with the balanced sample and tried to predict the unbalanced sample cases tends to think that there is more data for one class than the other. Seto et al. (2022) [108] also state that applying gradient boosting makes the unbalance of the data worse. This is less critical when applying Random Forest [109] or a neural network [56]. The hyperparameter ranges explored in this study were defined based on preliminary tests and prior applications in construction safety research, to ensure computational feasibility and avoid unstable configurations under highly imbalanced data. Nevertheless, we acknowledge that broader search spaces (e.g., deeper trees, wider learning rates, or larger neural architectures) could provide alternative optimal configurations. These hyperparameters were optimised using the full 70 % balanced training set and subsequently validated through k-fold cross-validation, rather than by means of a nested cross-validation scheme. While this strategy preserved an untouched 30 % test set and allowed for efficient optimisation, nested cross-validation would provide a less biased estimate of generalisation error. Future work could therefore explore extending the robustness of the proposed models.

From a practical perspective, our findings suggest several areas where safety management in the construction industry should be

strengthened. Workers with low tenure require targeted induction and mentoring programs to mitigate their higher vulnerability, while older workers (40–59 years) would benefit from ergonomic adaptation and closer supervision, given their disproportionate share of severe and fatal accidents. Mobility across construction sites, particularly movements between work areas, should be re-designed through segregated routes, traffic control plans, and the systematic use of signallers, as this scenario accounts for the majority of critical accidents. Micro-enterprises and subcontractors, which concentrate a high proportion of fatal cases, demand tailored training and simplified auditing mechanisms to ensure compliance with safety standards. Finally, technological aids such as 360° cameras, proximity sensors, and reverse alarms, as well as digital predictive tools for real-time monitoring, should be mainstreamed to reduce risks associated with vehicle manoeuvring and loss of control. These recommendations are consistent with EU-OSHA and NIOSH guidelines and provide a practical pathway to translate predictive insights into preventive action

5. Conclusions

Vehicle run-over construction site incidents make up a modest share of minor injuries yet remain one of the main sources of fatal harm on Spanish worksites.

The variables hour block, age, job tenure, site zone, deviation pattern, injury type, and body region, significant in the severity of related construction vehicle accidents and chosen for the generation of the classification models, show within the models a similar behaviour to the conclusions found in previous scientific studies.

The CVC—O model allows for the classification of accident severity considering variables related to the significant characteristics of the company found in the official DELT@ database. The variable with the greatest impact is the worker's age. Construction site vehicle accidents occur mostly in men and, in the case of fatalities, almost entirely in men. Regarding age, the highest accident rates are between 30 and 49 years, and in the case of fatal accidents, between 40 and 59 years. These are experienced middle-aged workers, which may suggest overconfidence in their work as a trigger for accidents. There is no "Weekday effect" or "Monday effect" in the construction vehicle accidents. The highest number of accidents occurs in the early hours of the working day, and the time of day with the highest number of accidents is concentrated between 9:00 and 13:59 and between 18:00 and 18:59, so checks by supervisors, safety technicians, and project managers should be reinforced in these time slots. Government and/or sectoral information and awareness-raising policies to prevent hit-and-run accidents in the construction sector should focus mainly on micro-enterprises (less than 5 workers) and small enterprises (up to 25 workers), as they have the worst accident rates. At the construction company level, the training of new workers must be reinforced, as the worst accident figures are for workers with less than one year's experience. This training could even be carried out before joining the company by online means. On specific issues, it would also be necessary to remember the use of PPE (helmets, safety belts, etc.) and to take refresher courses on driving machinery and vehicles.

The CVAV-C model considers the variables specific to the circumstances and consequences of the accidents reported in the database. The variables Body Part Injured, Injury and Deviation have the greatest influence on the model, especially in the case of critical accidents. The results of this study reinforce the need for a review of the protocols for the use of PPE (helmets, safety belts, etc.) by managers and supervisors, as well as the requirement for specific training for workers who are responsible for driving vehicles and machinery on a construction site.

These findings demonstrate significant associations between accident severity and factors such as age, tenure, and company size, but causal mechanisms cannot be established with the present data

As seen in the literature review, there are not many studies on hit and run over by vehicles at the country level, so developing comparisons with accident data from other countries would be an interesting future line of work. The potential repercussions of accidents involving construction vehicles on women employed within the construction sector will also be a focal point in forthcoming research.

Beyond identifying key predictors of accident severity, this study also highlights concrete opportunities for improving safety management in construction, particularly regarding vulnerable worker groups, internal mobility, and the adoption of preventive technologies, as discussed above.

Funding

This research in this paper is supported in part by Consejería de Educación de la Junta de Castilla y Léon, funded by the Subvenciones Destinadas al Apoyo a Grupos de Investigación Reconocidos (GIR) de las Universidades Públicas de Castilla y León (2024), grant number 777,441

Data availability

Accident records were provided upon request to the Spanish Ministry of Labour. An anonymous version and the KNIME workflow can be shared with researchers upon request.

CRediT authorship contribution statement

A Sánchez-Lite: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **J.L. Fuentes-Bargues:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **J.M. Geijo-Barrientos:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **C González-Gaya:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **A.Z. Sampaio:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Acknowledgments

This paper is based on the ongoing research of the Working Group on Risk Engineering in Manufacturing (REM) of the Manufacturing Engineering Society and the Recognised Research Group—Rodrigo Zamorano of History of Science and Technology (University of Valladolid). The authors therefore wish to express their gratitude for the support from both institutions.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rineng.2025.107762.

References

[1] R.A. Haslam, S.A. Hide, A.G.F. Gibb, D.E. Gyi, T. Pavitt, S. Atkinson, A.R. Duff, Contributing factors in construction accidents, Appl. Ergon. Invit. Pap. Spec. Ed. Ergon. Build. Constr. 36 (4) (2005) 401–416.

[2] J. Takala, P. Hämäläinen, N. Nenonen, K. Takahashi, O. Chimed-Ochir, J. Rantanen, Comparative analysis of the burden of injury and illness at work in

selected countries and regions. cent, Eur. J. Occup. Environ. Med. 23 (2017) 6–31.
[3] A. Suraji, R. Duff, S.H. Peckitt, Development of causal model of construction accident causation, J. Constr. Eng. Manag. (2001) 337–344.
[4] J.C. Cawley, B.C. Brenner, Occupational electrical injuries in the US 2003–2009. IEEE Paper no. ESW-2012-24, 2012.
[5] B. Hola, M. Szóstak, Analysis of the state of the accident rate in the construction industry in European Union Countries, Arch. Civ. Eng. (4) (2015) 19–34, https://doi.org/10.1515/ace-2015-0033. LXI.
[6] Z. Zhu, M.W. Park, C. Koch, M. Soltani, A. Hammad, K. Davari, Predicting movements of onsite workers and mobile equipment for enhancing construction site safety, Autom. Constr. 68 (2016) 95–101, https://doi.org/10.1016/j.autcon.2016.04.009.
[7] M. Paguay, J.D. Febres, E. Valarezo, Occupational Accidents in ecuador: an approach from the construction and manufacturing industries, Sustainability 15 (2023) 12661, https://doi.org/10.3390/su151612661.
[8] E.S. Pollack, R.T. Chowdhury, Trends in Work-Related Death and Injury Rates Among U.S. Construction workers, 1992–1998, Center to Protect Workers' Rights, Silver Spring, MD, 2001.
[9] C.M. Tam, S.X. Zeng, Z.M. Deng, Identifying elements of poor construction safety management in China, Saf. Sci. 42 (7) (2004) 569–586.
[10] M.A. Mariscal, E.M. López-Perea, J.R. López-García, S. Herrera, S. García-Herrero, The influence of employee training and information on the probability of accident rates, Int. J. Ind. Erg. 72 (2019) 311–319, https://doi.org/10.1016/j.ergon.2019.06.002.
[11] I. Rodríguez-Garzón, V. Lucas-Ruiz, M. Martínez-Fiestas, A. Delgado-Padial, Association between perceived risk and training in the construction industry, J. Constr. Eng. Manag. 141 (5) (2014) 04014095, https://doi.org/10.1061/(ASCE) CO.1943-7862.0000960.
[12] Z. Ismail, S. Doostdar, Z. Harun, Factors influencing the implementation of a safety management system for construction sites, Saf. Sci. 50 (3) (2012) 418e423, https://doi.org/10.1016/j.ssci.2011.10.001.
[13] M. Suárez-Cebador, J.C. Rubio-Romero, A. López-Arquillos, Severity of electrical accidents in the construction industry in Spain, J. Saf. Res. 48 (2014) 63–70, https://doi.org/10.1016/j.jsr.2013.12.002.
[14] A. Romero Barriuso, B.M. Villena Escribano, A. Rodríguez Sáiz, The importance of preventive training actions for the reduction of workplace accidents within the Spanish construction sector, Saf. Sci. 134 (2021), https://doi.org/10.1016/j.ssci.2020.105090.
[15] D. Almaskati, S. Kermanshachi, A. Pamidimukkala, K. Loganathan, Z. Yin, A review on construction safety: hazards, mitigation strategies, and impacted sectors, Buildings 14 (2024) 526, https://doi.org/10.3390/buildings14020526.
[16] J. Hinze, X.Y. Huang, L. Terry, The nature of struck-by accidents, J. Constr. Eng. Manag. 131 (2) (2005) 262–268.
[17] Z. Riaz, D.J. Edwards, A. Thorpe, SightSafety: a hybrid information and communication technology system for reducing vehicle/pedestrian collisions, Autom. Constr. 15 (2006) 719–728, https://doi.org/10.1016/j.autcon.2005.09.004.
[18] P. Kines, Case studies of occupational fall from height. Cognition and behaviour in context, J. Saf. Res. 34 (2003) 263–271.
[19] C. Chia-Fen, C. Tin-Chang, T. Hsin-I, Accident patterns and prevention measures for fatal occupational falls in the construction industry, Appl. Erg. 36 (4) (2005) 391–400.
[20] Y. Kang, S. Siddiqui, S.J. Suk, S. Chi, C. Kim, Trends of fall accidents in the US construction industry, J. Constr. Eng. Manag. 143 (8) (2017), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001332.
[21] A. Zermane, M. Zahirasri, M.R. Baharundin, H.M. Yusoff, Analysis of the contributing factors for fatal accidents due to falls from heights in Malaysia and the USA, Sci. Technol. 28 (S1) (2020) 15–36.
[22] L.G. Kang, Statistical analysis and case investigation of fatal fall-from-height accidents in the Chinese construction industry, Int. J. Ind. Eng. Theory Appl. Pract. 29 (3) (2022) 413–431.
[23] C.F. Chi, C.C. Yang, Z.L. Chen, In-depth accident analysis of electrical fatalities in the construction industry, Int. J. Ind. Erg. 39 (4) (2009) 635–644.
[24] C.A. Janicak, Occupational fatalities due to electrocutions in the construction industry, J. Saf. Res. 39 (6) (2008) 617–621.
[25] M. Suárez-Cebador, J.C. Rubio-Romero, J.A. Carrillo-Castrillo, A. López-Arquillos, A decade of occupational accidents in Andalusian (Spain) public universities, Saf. Sci. 80 (2015) 23–32, https://doi.org/10.1016/j.ssci.2015.07.008.
[26] K. Hakkinen, Crane accidents and their prevention, J. Occup. Accid. 1 (1978) 353–361.
[27] M.F. Milazzo, G. Ancione, V. Spasojevic Brkic, Safety in crane operations: an overview on crane-related accidents, in: Proceedings of the 6th International Symposium on Industrial Engineering, SIE, Tokyo, Japan, 2015, pp. 36–39.
[28] V. Herrera-Pérez, F. Salguero-Caparrós, M.C. Pardo-Ferreira, J.C. Rubio-Romero, Key factors in crane-related occupational accidents in the spanish construction industry (2012–2021), Int. J. Environ. Res. Public Health 20 (2023) 7080, https://doi.org/10.3390/ijerph20227080.
[29] F.G. Benavides, M.T. Giráldez, E. Castejón, N. Catot, M. Zaplana, J. Delclós, J. Benach, D. Gimeno, Análisis de los mecanismos de producción de las lesiones leves por accidentes de trabajo en la construcción en España, Gac. Sanit. 17 (5) (2003) 353–359.
[30] W. Wu, H. Yang, D. Chew D., S. Yang, A. Gibb, Q. Li, Towards an autonomous real-time tracking system of near-miss accidents on construction sites, Autom. Constr. 19 (2) (2010) 134–141.
[31] O. Golovina, J. Teizer, N. Pradhananga, Heat map generation for predictive safety planning: preventing struck-by and near miss interactions between workers-on-foot and construction equipment, Autom. Constr. 71 (2016) 99, https://doi.org/10.1016/j.autcon.2016.03.008. -15.
[32] N.H. Abas, The analysis of struck-by accidents at construction sites in Johor, Int. J. Integr. Eng. 12 (2020) 262–275.
[33] J. Hinze, Construction Safety, Prentice-Hall, Upper Saddle River, N.J, 1997.
[34] J. Teizer, B.S. Allread, U. Mantripragada, Automating the blind spot measurement of construction equipment, Autom. Constr. 19 (4) (2010) 491–501 (2010).
[35] W. Wu, H. Yang, Q. Li, D. Chew, An integrated information management model for proactive prevention of struck-by-falling-object accidents on construction sites, Autom. Constr. 34 (2013) 67e74, https://doi.org/10.1016/j.autcon.2012.10.010.
[36] S. Fass, R. Yousef, D. Liginlal, P. Vyas, Understanding causes of fall and struck-by incidents: what differentiates construction safety in the Arabian Gulf region? Appl. Erg. 58 (2017) 515–526, https://doi.org/10.1016/j.apergo.2016.05.002.
[37] B. Thomson, MS thesis, Univ. of Washington, 1996.
[38] Z. Woźniak, H. Bożena, The structure of near misses and occupational accidents in the polish construction industry, Heliyon 10 (2024) e26410, https://doi.org/10.1016/j.heliyon.2024.e26410.
[39] M.A. Camino-López, D.O. Ritzel, I. Fontaneda, O.J. González-Alcantara, Construction industry accidents in Spain, J. Saf. Res. 39 (2008) 497–507, https://doi.org/10.1016/j.jsr.2008.07.006.
[40] T.L. Bunn, C.A. Northcutt, R. Honaker, P. Maloney, Quantitative and narrative analysis of dump truck-related injuries and fatalities in the United States, Safety 11 (1) (2025) 17.
[41] Romano, N.T., & Cassini, V.J. (2014). Preventing worker injuries and deaths from backing construction vehicles and equipment at roadway construction worksites.
[42] W. Fan, C. Carroll, L. Radley, E. Hostetler, S. Choe, F. Leite, C. Caldas, Prevention of Backing Fatalities in Construction Work Zones (No. TxDOT Report 0-6703-1), Texas Department of Transportation. Research and Technology Implementation Office, 2019.
[43] S. Sarkar, A. Pramanik, J. Maiti, G. Reniers, Predicting and analyzing injury severity: a machine learning-based approach using class-imbalanced proactive and reactive data, Saf. Sci. 125 (2020) 104616. Ok.
[44] M. Alkaissy, M. Arashpour, E.M. Golafshani, M.R. Hosseini, S. Khanmohammadi, Y. Bai, H. Feng, Enhancing construction safety: machine learning-based classification of injury types, Saf. Sci. 162 (2023) 106102.
[45] X. Luo, X. Li, Y.M. Goh, X. Song, Q. Liu, Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents, Saf. Sci. 163 (2023) 106138.
[46] S. Yoon, T. Chang, S. Chi, Developing an integrated construction safety management system for accident prevention, J. Manag. Eng. 40 (6) (2024) 04024051.
[47] B. Zhong, X. Pan, P.E. Love, L. Ding, W. Fang, Deep learning and network analysis: classifying and visualizing accident narratives in construction, Autom. Constr. 113 (2020) 103089.
[48] M. Shayboun, D. Kifokeris, C. Koch, A review of machine learning for analysing accident reports in the construction industry, J. Inf. Technol. Constr. 30 (2025) 439–460.
[49] M. Cavalcanti, L. Lessa, B.M. Vasconcelos, Construction accident prevention: a systematic review of machine learning approaches, Work 76 (2) (2023) 507–519.
[50] A.J. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Application of machine learning to construction injury prediction, Autom. Constr. 69 (2016) 102–114.
[51] M. Alkaissy, I. Elbeltagi, et al., Enhancing construction safety: machine learning-based classification of injury types, Saf. Sci. 162 (2023) 106102.
[52] E.D. Oguz Erkal, M.R. Hallowell, A. Ghriss, S. Bhandari, Predicting serious injury and fatality exposure using machine learning in construction projects, J. Constr. Eng. Manage 150 (3) (2024) 04023169.
[53] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[54] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern. Recognit. 45 (1) (2012) 521–530.
[55] T. Dankowski, A. Ziegler, Calibrating random forests for probability estimation, Stat. Med. 35 (22) (2016) 3949–3960.
[56] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.
[57] Occupational Safety and Health Administration (OSHA). Preventing Backovers (2023). U.S. Department of Labor. Available online: https://www.osha.gov/preventing-backovers/solutions. (accessed on 20 September 2025).
[58] National Institute for Occupational Safety and Health. Preventing Dump Truck-Related Injuries and Deaths During Construction—Guidance for Employers. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, DHHS (NIOSH) Publication No. 2023-137. Available online: https://www.cdc.gov/niosh/docs/2023-137/default.html. (Accessed on 20 September 2025).
[59] European Agency for Safety and Health at work (EU-OSHA), Vehicle Safety E-Guide – Workplace transport, 2023. Available on: https://eguides.osha.europa.eu/vehicle-safety/what-is-your-area-of-work/workplace-transport (Accessed on 20 September 2025).
[60] Ministry of Labour and Social Affairs of Spain (MLSAS), Orden TAS/2926/2002, de 19 de noviembre, por la que se establecen nuevos modelos para la notificación de los accidentes de trabajo y se posibilita su transmisión por procedimiento

electrónico, Boletín Oficial del Estado 279 (2002) 40988–41013, núm21st November.
[61] CNAE 09, National classification of economic activities in Spain, Available on, https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614, 2009. Accessed May 24, 2023.
[62] A. López-Arquillos, J.C. Rubio-Romero, A. Gibb, Analysis of construction accidents in Spain, 2003-2008, J. Saf. Res. 43 (2012) 381–388, https://doi.org/10.1016/j.jsr.2012.07.005.
[63] F.J. Forteza, J.M. Carretero-Gomez, A. Sese, Occupational risks, accidents on sites and economic performance of construction firms, Saf. Sci. 94 (2017) 61–76, https://doi.org/10.1016/j.ssci.2017.01.003.
[64] J.L. Fuentes-Bargues, A. Sánchez-Lite, C. González-Gaya, V.F. Rosales-Prieto, G. Reniers, A study of situational circumstances related to Spain's occupational accident rates in the metal sector from 2009 to 2019, Saf. Sci. 150 (2022) 105700, https://doi.org/10.1016/j.ssci.2022.105700.
[65] M. Suárez-Cebador, J.C. Rubio-Romero, A. López-Arquillos, Severity of electrical accidents in the construction industry in Spain, J. Saf. Res. 48 (2014) 63–70, https://doi.org/10.1016/j.jsr.2013.12.002.
[66] L. Sanmiquel, J.M. Rossell, C. Vintró, Study of Spanish mining accidents data mining techniques, Saf. Sci. 75 (2015) 49–55, https://doi.org/10.1016/j.ssci.21015.01.016.
[67] R.A.A. Ince, B.L. Giordano, C. Kayser, G.A. Rousselet, J. Gross, P.G. Schyns, A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula, Hum. Brain Mapp. 38 (2017) 1541–1573, https://doi.org/10.1002/hbm.23471.
[68] IBM Corp, IBM SPSS Statistics for Windows, Version 30.0, IBM Corp, Armonk, NY, 2024.
[69] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed., O'Reilly Media, 2022.
[70] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[71] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, KNIME - the Konstanz information miner, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 26–31, https://doi.org/10.1145/1656274.1656280.
[72] A.G. KNIME, KNIME Analytics Platform, Version 5.0.0, Talacker 50 8001, 2025. Zürich Switzerland.
[73] S. O'Hagan, D.B. Kell, Software review: the KNIME workflow environment and its applications in genetic programming and machine learning, Genet. Program. Evolvable Mach. 16 (2015) 387–391, https://doi.org/10.1007/s10710-015-9247-3.
[74] L. Sasse, O. van den Bosch, V. Fortuin, J. Gawlikowski, S. Auer, O. Kramer, Overview of leakage scenarios in supervised machine learning, J. Big. Data 12 (2025) 92, https://doi.org/10.1186/s40537-025-01193-8.
[75] S. Chowdhury, Y. Lin, B. Liaw, L. Kerby, Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance, J. Big. Data 8 (1) (2021) 123, https://doi.org/10.1007/s40537-021-00568-3.
[76] S. Wang, Y. Bao, S. Yang, HS-SMOTE: oversampling method for multiple dynamic interpolations based on regular hexagon scoring mechanism, Expert. Syst. Appl. 265 (2025) 125855, https://doi.org/10.1016/j.eswa.2024.125855.
[77] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79, https://doi.org/10.1214/09-SS054.
[78] J.L. Fuentes-Bargues, A. Sánchez-Lite, I. Iglesias, C. González-Gaya, Proposal of a workplace classification model for heart attack accidents from the field of occupational safety and health engineering, Heliyon. 10 (2024) e37647, https://doi.org/10.1016/j.heliyon.2024.e37647.
[79] G.D. Garson, Interpreting neural-network connection weights, AI Expert 6 (4) (1991) 46–51.
[80] J.D. Olden, D.A. Jackson, Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks, Ecol. Modell. 154 (1) (2002) 135–150, https://doi.org/10.1016/S0304-3800(02)00064-9.
[81] National Statistical Institute (NSE), Economía /Cuentas económicas / Contabilidad nacional anual de España: principales agregados. Resultados / PIB a Precios De Mercado, 2023. Available on: https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177057&menu=resultados&idp=1254735576581. Access 30 July 2024.
[82] National Statistical Institute (NSE), Actividad, ocupación y paro /Mercado laboral / Encuesta Población Activa / Resultados nacionales / Ocupados por sexo y rama de actividad. https://ine.es/jaxiT3/Datos.htm?t=4128#!tabs-tabla, 2023. Access 30 July 2024.
[83] I. Fontaneda, M.A. Camino López, O.J. González Alcántara, B.A. Greiner, The "weekday effect": a decrease in occupational accidents from Monday to Friday—an extension of the "Monday effect, Biomed. Res. Int. (2024) 1–12, https://doi.org/10.1155/2024/4792081.
[84] J.P. Leigh, T.R. Miller, Occupational illnesses within two national data sets, Int. J. Occup. Environ. Health 4 (2) (1998) 99–113.
[85] E. Ebrahimihoor, M. Karpman, J. Grover, N. Muganlinskaya, Day of the week variation in emergency department arrivals, chest pain, and acute myocardial infarction throughout 2016-2019, J. Community Hosp. Intern. Med. Perspect. 13 (5) (2023), https://doi.org/10.55729/2000-9666.1237.
[86] L.A. Backkonsnala, Occupational injuries by hour of day and day of week, Aust. N. Z. J. Public 31 (1) (2007) 88–89, https://doi.org/10.1111/j.1467-842X.2007.tb00896.x.
[87] K. Hänecke, S. Tiedemann, F. Nachreiner, H. Grzech-Sukalo, Accident risk as a function of hour at work and time of day as determined from accident data and exposure models for the German working population, Scand. J. Work Environ. Health 24 (Suppl 3) (1998) 43–48.
[88] Observatorio Industrial de la Construcción, La mujer en el sector de la construcción: Informe 2025, Fundación Laboral de la Construcción (2025). Available on: https://www.observatoriodelaconstruccion.com (Accessed on 20 September 2025).
[89] T.N. Hanvold, P. Kines, M. Nykänen, S. Thomée, K.A. Holte, J. Vuori, M. Wærsted, K.B. Veiersted, Occupational safety and health among young workers in the nordic countries: a systematic literature review, Saf. Health Work 10 (2019) 3–20, https://doi.org/10.1016/j.shaw.2018.12.003.
[90] J. Berecki-Gisolf, F.J. Clay, A. Collie, R.J. McClure, The impact of aging on work disability and return to work: insights from workers' compensation claim records, J. Occup. Environ. Med. 54 (3) (2012) 318–327, https://doi.org/10.1097/JOM.0b013e31823fdf9d.
[91] B. Stoesz, K. Chimney, C. Deng, H. Grogan, V. Menec, C. Piotrowski, S. Shooshtari, N. Turner, Incidence, risk factors, and outcomes of non-fatal work-related injuries among older workers: a review of research from 2010 to 2019, Saf. Sci. 126 (2020) 104668, https://doi.org/10.1016/j.ssci.2020.104668.
[92] M.C. Barth, An aging workforce in an increasingly global world, J. Aging Soc. Policy. 11 (2–3) (2000) 83–88, https://doi.org/10.1300/J031v11n02_09.
[93] L. Peng, A.H.S. Chan, A meta-analysis of the relationship between ageing and occupational safety and health, Saf. Sci. 112 (October 2018) (2019) 162–172, https://doi.org/10.1016/j.ssci.2018.10.030.
[94] G. Bravo, C.A. Viviani, M. Lavalliere, P.M. Arezes, M. Martínez, I. Dianat, H. I. Castellucci, Do older workers suffer more workplace injuries? A systematic review, Int. J. Occupat. Saf. Erg. (2020) 1–56, https://doi.org/10.1080/10803548.2020.1763609.
[95] M. Loosemore, N. Malouf, Safety training and positive safety attitude formation in the Australian construction industry, Saf. Sci. 113 (2019) 233–243, https://doi.org/10.1016/j.ssci.2018.11.029.
[96] M. Picchio, J.C. Van Ours, Temporary jobs and the severity of workplace accidents, J. Saf. Res. 61 (2017) 41–51.
[97] J.A. Carrillo-Castrillo, J.C. Rubio-Romero, L. Onieva, A. López-Arquillos, The causes of severe accidents in the Andalusian manufacturing sector: the role of human factors in official accident investigations, Hum. Factors Erg. Manuf. Ind. 26 (1) (2016) 68–83, https://doi.org/10.1002/hfm.20614.
[98] J.L. Fuentes-Bargues, A. Sánchez-Lite, C. González-Gaya, M.A. Artacho-Ramírez, Descriptive analysis and a proposal for a predictive model of fatal occupational accidents in Spain, Heliyon. 9 (11) (2023) e22219, https://doi.org/10.1016/j.heliyon.2023.e22219.
[99] M. Szóstak, Analysis of occupational accidents in the construction industry with regards to selected time parameters, Open Eng. 9 (2019) 312–320, https://doi.org/10.1515/eng-2019-0027.
[100] A.J.R. Santos, E.L. Rebelo, J.C. Mendes, Towards better prevention of fatal occupational accidents in Portugal, Int. Labour. Rev. 157 (3) (2018) 409–433, https://doi.org/10.1111/ilr.12114.
[101] B. Gulhan, M.N. Ilhan, E. Fusun Civil, Occupational accidents and affecting factors of metal industry in a factory in Ankara, Turk. J. Public Health 10 (2) (2012) 76–85.
[102] S.B. Batti Gonçalves, T. Mamoru Sakae, F. Liberali Magajewski, Prevalence and factors associated with work accidents in a metal-mechanic company, Rev. Bras. Med. Trab. 16 (1) (2018) 26–35, https://doi.org/10.5327/Z1679443520180086.
[103] S. Xu, M. Zhang, L. Hou, Formulating a learner model for evaluation construction worker's learning ability during safety training, Saf. Sci. 116 (2019) 97–107, https://doi.org/10.1016/j.ssci.2019.03.002.
[104] D.J. Provan, A. Rae J., S.W.A. Dekker, An ethnography of the safety professional's dilemma: safety work or the safety of work? Saf. Sci. 117 (2019) 276–289, https://doi.org/10.1016/j.ssci.2019.04.024.
[105] K. Tong, Z. Han, Y. Shen, Y. Long, Y. Wei, An Integrated Machine Learning and Deep Learning Framework for Credit Card Approval Prediction, in: 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2024, pp. 853–858, https://doi.org/10.1109/ICPICS62053.2024.10795883.
[106] R. Vatambeti, et al., Enhancing sparse data recommendations with self-inspected adaptive smote and hybrid neural networks, Sci. Rep. 15 (2025) 17229, https://doi.org/10.1038/s41598-025-02593-9.
[107] M. Carvalho, A.J. Pinho, S. Brás, Resampling approaches to handle class imbalance: a review from a data perspective, J. Big. Data 12 (2025) 71, https://doi.org/10.1186/s40537-025-01119-4.
[108] H. Seto, A. Oyama, S. Kitora, H. Toki, R. Yamamoto, J. Kotoku, T. Moriyama, Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data, Sci. Rep. 12 (2022) 15889, https://doi.org/10.1038/s41598-022-20149-z.
[109] T. Dankowski, A. Ziegler, Calibrating random forests for probability estimation, Stat. Med. 35 (22) (2016) 3949–3960, https://doi.org/10.1002/sim.6959.