Expert Systems With Applications 316 (2026) 131895

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Supervised clustering using SOM for severity-based pattern detection in urban traffic crashes

Lluís Bermúdez [a,b,*], Isabel Morillo [b], Anna Salazar [a]

[a] Riskcenter-IREA, Universitat de Barcelona, Barcelona, Spain
[b] Department of Economics, Financial and Actuarial Mathematics, University of Barcelona, Barcelona, Spain

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Shapley values
Kohonen network
Road safety planning

ABSTRACT

Urban traffic crashes remain a critical public health challenge, particularly for vulnerable road users. This study introduces a novel data-driven methodology to support urban road safety planning by identifying well-defined, interpretable crash typologies associated with fatal or serious injuries. The proposed framework relies on a *supervised clustering* strategy that integrates SHAP (SHapley Additive exPlanations) values with Self-Organizing Maps (SOM). Applied to urban crash data from Barcelona (2017-2019), the approach uncovers ten distinct and interpretable crash typologies, capturing high-risk scenarios such as speed-related nighttime collisions and pedestrian-heavy vehicle conflicts, as well as less explored patterns including two-wheeler falls and bicycle-motorcycle interactions. By combining SHAP-based explanations with topology-preserving neural mapping, the SOM framework reveals subtle gradations of risk, preserves neighborhood relationships among crash profiles, and enhances subgroup detection and interpretability beyond traditional unsupervised clustering methods and standard eXplainable Artificial Intelligence (xAI) summaries. These results underscore the potential of SOM-based supervised clustering to inform targeted, data-driven safety interventions. More broadly, the study advances methodological research on supervised clustering and offers a transferable tool for detecting high-dimensional risk patterns in urban safety analysis and other applied domains.

1. Introduction

According to the most recent estimates of the World Health Organization (WHO, 2023), road traffic crashes cause approximately 1.19 million deaths each year, along with 20 to 50 million non-fatal injuries, a significant proportion of which lead to long-term disabilities. These incidents place a significant financial burden on national economies, with associated costs estimated to reach 3% of a country's gross domestic product. Vulnerable road users, such as pedestrians, cyclists, and motorcyclists, account for more than half of global road traffic deaths.

In an urban traffic setting, Barcelona reported its lowest number of traffic fatalities in 15 years in 2024 (Ajuntament de Barcelona, 2025b). Despite a 3.5% decrease in overall injuries, serious injuries still increased by 6.1%. In particular, 93% of these severe cases involved vulnerable road users. According to the European Commission (EC, 2024), approximately 57% of serious injuries in the EU involve vulnerable road users.

These statistics highlight the critical need for sustained efforts to protect the most vulnerable road users and underscore the urgency of implementing enhanced road safety measures worldwide, such as the "Let's Protect Schools" initiative in Barcelona, evaluated in Lopez-Muley et al. (2025). Such actions align with the mandate of the World Health Organization to reduce traffic-related fatalities and serious injuries, which constitutes the central focus of this study.

Advancing meaningful progress in this field requires a comprehensive understanding of the mechanisms underlying crash occurrences and their contributing factors. This understanding can be facilitated by the development of predictive models that establish relationships between crash likelihood and severity and a set of variables, including roadway design, traffic flow conditions, vehicle characteristics, driver behavior, and environmental influences.

Crash prediction modeling has traditionally been grounded in statistical and econometric methods, which have played a central role in the field and have been extensively reviewed in the literature (Mannering et al., 2016). However, in recent decades, the use of machine learning (ML) techniques has seen substantial growth, largely due to their superior predictive performance. These models also surpass standard approaches by offering increased flexibility and eliminating the need for predefined data distribution assumptions. Building on recent developments in AI and evolving ML methodologies, Ali et al. (2024) present

* Corresponding author.
E-mail addresses: lbermudez@ub.edu (L. Bermúdez), imorillo@ub.edu (I. Morillo), asalazarb@ub.edu (A. Salazar).

https://doi.org/10.1016/j.eswa.2026.131895
Received 22 October 2025; Received in revised form 24 February 2026; Accepted 26 February 2026
Available online 1 March 2026

a comprehensive review of ML applications in crash prediction modeling, focusing on three key areas: crash occurrence, crash frequency, and injury severity prediction.

In this paper, with the aim of extracting data-driven insights to support the reduction of traffic-related fatalities and serious injuries by enhancing road safety measures, we focus on the development and analysis of injury severity prediction ML models.

Recent literature has employed a variety of supervised ML techniques (e.g., for binary or multiclass classification tasks) to address this purpose. For example, Iranitalab and Khattak (2017) compared statistical and ML models, including the multinomial logit model, the k-nearest neighbour (K-NN), the support vector machine (SVM), and random forests (RF). AlMamlook et al. (2019) applied the AdaBoost algorithm and evaluated its performance against several alternative models, such as naïve Bayes (NB) and RF. Arhin and Gatiba (2019) demonstrated the superior predictive capability of artificial neural networks (ANN) over several statistical and ML counterparts, while Zheng et al. (2019) reported comparable findings using convolutional neural networks (CNN). Similarly, Zhang et al. (2022) used NB, K-NN, logistic regression (LR), and extreme gradient boosting (XGB). Chen et al. (2025) provides key findings on applying ML models to imbalanced traffic crash data.

Most of the above studies adopted a binary classification, typically categorizing injury severity into two groups (e.g., fatal versus non-fatal or fatal/serious versus non-serious), and subsequently applied ML models to evaluate the influence of various risk factors on the severity of crash-related injuries. To address the inherent limitations of ML, particularly their lack of interpretability, several studies incorporate eXplainable Artificial Intelligence (xAI) techniques, such as SHAP (SHapley Additive exPlanation) values, to enhance transparency and support informed decision-making. See, for example, Wang et al. (2024) and Chen et al. (2025).

In this context, cluster analysis offers the potential to identify different subgroups of traffic crashes based on severity levels. By examining the shared characteristics within each cluster, targeted road safety interventions can be proposed to mitigate the occurrence of high-severity crashes. Despite this potential, the application of unsupervised ML techniques remains relatively limited in the literature (Suarez-del Fueyo et al., 2021). This may be attributed to the generally unsatisfactory reported results, often characterized by poorly defined clusters with significant overlap, which hinders reliable differentiation.

In particular, much of the existing research that involves unsupervised ML techniques adopts a hybrid framework that integrates clustering with classification algorithms. This hybrid approach, which first partitions crash data into clusters prior to classification, has been shown to improve the predictive accuracy of supervised ML models compared to those applied without prior clustering. See, for example, Taamneh et al. (2017), Hasheminejad et al. (2018), and Assi (2020).

Based on previous studies, we propose an inversion of this standard hybrid methodology by performing a supervised classification model before applying an unsupervised clustering technique. Lundberg et al. (2019) introduced the concept of *supervised clustering*, wherein clustering is performed on SHAP values produced by a supervised model, rather than being applied directly to the raw feature data in a fully unsupervised manner. The term "*supervised*" comes from the fact that SHAP values reflect the contribution of each feature to predictions made by the supervised model, distinguishing this approach from purely unsupervised clustering and from hybrid frameworks. Based on this, Gramegna and Giudici (2020) conducted a comparative analysis between clustering on input features and on SHAP values, finding that the SHAP-based approach led to more clearly defined and better differentiated clusters.

Extending this methodology further, Cooper et al. (2021) demonstrated that reducing SHAP values to two dimensions using Uniform Manifold Approximation and Projection (UMAP) prior to clustering improved both performance and interpretability, positioning UMAP as an effective preprocessing step.

As the main contribution of this paper, we propose extending the methodology introduced by Bermúdez et al. (2023) by incorporating a Kohonen neural network, also known as a Kohonen map or Self-Organizing Map (SOM), as an alternative preprocessing step to UMAP. The SOM is an unsupervised learning algorithm widely used for data visualization, clustering, and pattern recognition (Huysmans et al., 2006), which we argue is more suitable for the context of our analysis.

Specifically, this article pursues two main objectives: first, to improve the understanding of existing xAI techniques to perform a risk analysis related to traffic fatalities and serious injuries; and second, to introduce a new methodology (i.e., *supervised clustering* using SOM) to identify clusters of crashes with a high or low probability of resulting in fatal or serious outcomes. Finally, the study demonstrates the practical application of the approach using a real-world dataset and discusses how the resulting insights can inform and support the development of effective road safety interventions.

The structure of the paper is as follows. The Data section provides a description of the real-world dataset used in the study. The Methods section reviews the foundational concepts of the ML models and xAI techniques employed, and introduces the proposed *supervised clustering* methodology. The Results and Discussion sections present the empirical findings and examine their implications for road safety policy, respectively. Finally, the Conclusions section summarizes the main contributions of the study and outlines potential directions for future research.

2. Data

The available dataset on traffic crashes in Barcelona from 2017 to 2019 was obtained from a database maintained by the local police, accessible through the Open Data BCN service (Ajuntament de Barcelona, 2025a). Each year includes five interconnected datasets, linked by a common record code, that detail crashes, their immediate causes, the types of incidents, and information about the vehicles and victims involved.

After merging the datasets, the final data included 27,721 traffic crashes that involved victims during the study period. Traffic crashes without victims (approximately 10% of the dataset) were excluded for two main reasons. First, minor incidents often do not require police intervention, which could lead to a biased sample of such cases. Second, the study specifically focuses on crashes that resulted in fatalities (deaths), serious injuries (hospitalized for more than 24 h), or non-serious injuries (treated at the scene of the incident, in hospital emergency services, or hospitalized for less than 24 h).

The variables used in the analysis are presented in Table 1. As outlined in the Introduction, the primary focus of this study is injury severity, which can be defined through several indicators available in the dataset. In this analysis, a binary variable, *Severity*, is utilized to classify crashes into two distinct categories: those resulting in fatal or serious injuries (FS) and those involving only non-serious injuries (NS). According to standard practice, the severity of the crash is determined based on the highest level of injury sustained by any vehicle occupant involved in the incident (Chang & Mannering, 1999).

Many studies have been conducted to identify the determinants that influence the severity of crashes. For example, Vorko-Jović et al. (2006) examined risk factors associated with urban road traffic crashes, highlighting those with the greatest impact on individuals who sustained fatal, serious, or minor injuries. In our study, we have considered a set of potential influencing factors that reflect various aspects of the crash context. These include the temporal and spatial characteristics of the crash, the vehicle-related characteristics, the related causes of the crash, and the type of incident.

The temporal characteristics of the crash are captured by two variables: *Daytype*, which distinguishes between working days and weekends, and *Daytime*, which categorizes the shift of the day into morning, afternoon, and nighttime. The location of the crash is represented by the *Roadway* variable, indicating the type of road where the incident

Table 1
Description of variables, including definitions and categories. All data come from police reports, and there are no missing values.

Name	Description, levels and proportions
Crash severity	
Severity	Fatal/Serious victims (FS: 2.5%), Non-serious (NS: 97.5%)
Temporal characteristics	
Daytype	Type of day in which the crash occurred (Working: 81.2%, Weekend: 18.8%)
Daytime	Part of the day in which the crash occurred (Morning: 39.4%, Afternoon: 50.1%, Nighttime: 10.5%)
Spatial characteristics	
Roadway	Type of road on which the crash occurred (Street: 32.8%, Avenue[a]: 58.2%, Fastlane: 9%)
Vehicle-related characteristics	
Vehicles	Multi-vehicle crash (+1: 80.1%), Single-vehicle (1: 19.9%)
Bicycle	Presence of this vehicle (Y: 8%), No presence (N: 92%)
Twowheel	Presence of this vehicle (Y: 65.1%), No presence (N: 34.9%)
Heavy	Presence of this vehicle (Y: 7.4%), No presence (N: 92.6%)
Light	Presence of this vehicle (Y: 76.7%), No presence (N: 23.3%)
Related causes of the crash	
Pedestrian	Presence of this cause (Y: 6.8%), No presence (N: 93.2%)
Alcohol	Presence of this cause (Y: 3.1%), No presence (N: 96.9%)
Speed	Presence of this cause (Y: 0.5%), No presence (N: 99.5%)
Roadcond	Presence of this cause (Y: 0.8%), No presence (N: 99.2%)
Type of crash	
Rundown	Presence of this event (Y: 12.8%), No presence (N: 87.2%)
Rolloverfall	Presence of this event (Y: 14%), No presence (N: 86%)
Collision	Presence of this event (Y: 71.9%), No presence (N: 28.1%)
Shock	Presence of this event (Y: 3.1%), No presence (N: 96.9%)

[a] wide urban road with >4 lanes and >2 km.

occurred and categorized as street, avenue (wide urban road with over four lanes and longer than 2 km) or fast lane.

Vehicle-related variables include *Vehicles*, which differentiates between single-vehicle and multi-vehicle crashes, as well as indicators for the types of vehicles involved. These include *Bicycle*, *Twowheel*, *Heavy* and *Light*, each denoting the presence or absence of that vehicle type in a given crash.

The related causes of the crash are represented by variables such as *Pedestrian* (crash attributed to pedestrian behavior), *Alcohol* (presence of elevated blood alcohol concentration as reported by police), *Speed* (excessive speed cited as causal factor by police), and *Roadcond* (poor road conditions identified as the cause of the crash). It should be noted that speed is recorded only when it can be clearly established during accident investigation, which is often challenging in urban settings. Consequently, the contribution of speed to traffic accidents is likely to be underreported and its prevalence in the data should be interpreted with caution.

The type of crash is included using variables such as *Rundown* (a pedestrian struck and lying on the ground), *Rolloverfall* (involving a rollover or fall), *Collision* (impact between two vehicles), and *Shock* (impact with a stationary object). It is important to note that as binary variables that indicate the presence or absence of each condition, a single crash can involve multiple vehicle types, related causes, or types of crash. These explanatory factors have been widely employed in previous research (Liu et al., 2022; Wang et al., 2019).

For descriptive purposes, Table 1 also reports the proportion of observations corresponding to each category. Fortunately, fatal and serious injury incidents can be considered as “rare events", comprising only 2.5% of all crashes, compared to 97.5% involving non-serious injuries. This substantial disparity results in an imbalanced dataset.

In summary, crashes are observed more frequently on working days, particularly during the morning or afternoon, and tend to occur more often on avenues. Most of the crashes in the dataset involve more than one vehicle (80.1%). Light vehicles (76.7%) and two-wheelers (65.1%) are the most frequently involved, while bicycles (8%) and heavy vehicles (7.4%) appear less frequently. The contributing factors most frequently observed are pedestrian behavior (6.8%) and elevated blood alcohol concentration (3.1%), while poor road conditions (0.8%) and excessive speed (0.5%) are less commonly identified as causes of crashes. Crashes typically result from a collision (71.9%), followed by rollovers and falls (14%) and rundowns (12.8%), while shocks (3.1%) are rare in urban areas.

Finally, it should be noted that certain variables, such as alcohol involvement or pedestrian attribution, may be subject to reporting biases, as testing or attribution practices can depend on crash severity or enforcement priorities. Accordingly, patterns identified in the posterior analysis should be interpreted with caution, reflecting the reported data rather than evidence of causal relationships.

3. Methods

As stated in the Introduction, a variety of supervised ML techniques can be used to evaluate the influence of various risk factors on the severity of crash-related injuries. However, some model intricacies, such as imbalanced data, model validation, and interpretability, must be carefully considered. This section provides an overview of the approaches adopted to address these issues. In addition, we introduce a new *supervised clustering* methodology that enables the identification of clusters of crashes with a high or low probability of resulting in fatal or serious outcomes.

3.1. Supervised ML models intricacies

This study evaluates a variety of predictive models, including logistic regression (LR), random forests (RF), and extreme gradient boosting (XGB). LR represents a conventional approach, based on predefined assumptions, that provides easily interpretable explanations. In contrast, RF and XGB models apply more flexible, data-driven methodologies, which often enhance predictive performance, but typically reduce model interpretability, leading to being referred to as “black-box" ML models. However, xAI techniques can help mitigate these interpretability challenges by providing insight into the decision-making processes and outputs of these models.

In particular, we begin by presenting the most widely adopted xAI techniques, with a particular emphasis on model-agnostic methods, characterized by their flexibility, as they can be applied across a wide range of ML algorithms. Typically used post-hoc (i.e., after model predictions have been generated), model-agnostic xAI techniques provide either global explanations, which describe the overall behavior of the model, or local explanations, which focus on individual predictions (Adadi & Berrada, 2018).

For global interpretability, we highlight Feature Importance (FI) methods, which determine the variables that most significantly influence model predictions; and visualization tools such as Accumulated Local Effects (ALE), which illustrate how predictions respond to changes in individual features while controlling for others. For local interpretability, we focus on SHAP values, a cooperative game theory-based model that quantifies the contribution of each input feature to a specific prediction, providing a detailed explanation of the model’s decision for a given instance (Lundberg & Lee, 2017). While SHAP is widely used, recent studies have highlighted certain limitations (Huang & Marques-Silva, 2024; Slack et al., 2021). Complementary approaches, such as LIME or counterfactual explanations, may provide practical alternatives.

As shown in Table 1, the number of crashes involving fatal or serious injuries (FS) is substantially smaller than those leading to non-serious injuries (NS), resulting in an imbalanced dataset. This imbalance poses a challenge for binary classification tasks, as predictive models may become biased toward the majority class, leading to unreliable results (Chawla et al., 2002; Japkowicz & Stephen, 2002). To mitigate this issue, we consider different resampling techniques commonly used in the recent literature (Bazarnovi & Mohammadian, 2024; Chen et al., 2025). These techniques can be broadly categorized into three groups: undersampling methods, which reduce the size of the majority class to balance

the dataset; oversampling methods, which increase the representation of the minority class by duplicating or synthetically generating instances; and hybrid methods, such as the SMOTE algorithm, which combine both strategies.

To identify the most appropriate ML model and the optimal resampling technique to handle class imbalance in our dataset, we evaluated model performance using several metrics derived from the confusion matrix, including accuracy, sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV). In addition, we used the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) for further evaluation. For all models, 80% of the data was allocated for training (through 5 repeated 10-fold cross-validation), while the remaining 20% was reserved for testing.

ML models were implemented using the *caret* package in R, whereas the *iml* (interpretable machine learning) package was used to apply the xAI techniques.

3.2. Supervised clustering using SOM

Using any supervised ML model along with the xAI techniques mentioned above (e.g., FI and ALE), we can assess the impact of various risk factors on the severity of crash-related injuries. However, as noted in the Introduction, we may also want to identify clusters of crashes with similar predicted severity levels or comparable degrees of prediction confidence. Crashes within the same group are likely to share underlying characteristics, which may call for similar road safety interventions. This clustering approach supports more effective resource allocation by allowing us to prioritize the most critical clusters or those with greater predictive confidence.

To achieve this, we propose using a *supervised clustering* approach instead of applying a traditional unsupervised clustering method. *Supervised clustering* involves grouping data based on the SHAP values of each feature generated by a supervised model, rather than clustering directly on the raw feature data (Lundberg et al., 2019). According to Cooper et al. (2021), this approach offers several advantages. In our case, (i) it weights features according to their importance in predicting crash severity, emphasizing the most informative ones while reducing the influence of fluctuations on less relevant features; (ii) acts as a preprocessing step by transforming feature data into a common numerical scale aligned with the output of the supervised prediction model, particularly beneficial when working with categorical features; and (iii) serves as a valuable tool for interpretable clustering.

The main contribution of this paper, extending the work of Bermúdez et al. (2023), is a *supervised clustering* approach that combines SHAP values with a Kohonen neural network or SOM. Based on the findings of Gramegna and Giudici (2020), who showed that SHAP-based representations enhance the formation of distinct clusters, Cooper et al. (2021) further demonstrated that reducing SHAP values to two dimensions using Uniform Manifold Approximation and Projection (UMAP) prior to clustering improves both performance and interpretability. In our approach, we use Self-Organizing Map (SOM) as an alternative preprocessing step to UMAP, aiming to achieve similar improvements in performance and interpretability within our case study, while also addressing methodological considerations related to manifold stability and reproducibility. SOM, introduced by Kohonen (1982, 1995), has been widely used for clustering and pattern detection in high-dimensional data. However, its application to crash data analysis remains relatively limited (Karimi et al., 2023; Pal et al., 2018).

In summary, SOM performs dimensionality reduction using a discrete, topology-preserving grid that maintains the relative structure of the data. This property ensures a stable and reproducible representation, where neighboring nodes correspond to similar patterns, making SOM particularly well-suited for pattern recognition and, in our case, for clustering crash profiles and associating them with injury severity categories. Unlike UMAP, which produces a continuous, non-linear embedding through stochastic optimization and can vary across runs depending on hyperparameter settings, SOM provides a deterministic mapping once the grid size and learning parameters are fixed. The discrete grid structure supports consistent aggregation and labeling of clusters, enhancing interpretability and facilitating communication with non-technical audiences. Furthermore, the grid-based layout clearly illustrates how different combinations of factors correspond to varying severity outcomes, providing an intuitive and spatially organized representation that is both reproducible and easier to explain than non-linear projections.

In particular, our approach follows a four-step procedure: (1) extraction of SHAP values from a supervised ML model; (2) construction of a two-dimensional representation using the SOM methodology; (3) application of an unsupervised clustering algorithm to identify distinct groups of crashes; and (4) visualization and interpretation of the defining characteristics and decision rules associated with each cluster.

In the first step, following the methodologies outlined in Section 3.1, an ML model is trained on the data, after which SHAP values are computed for all crashes in the testing sample. Although these values offer local interpretability by quantifying the contribution of each feature to individual crash predictions, they can also be used as input features for the *supervised clustering* procedure described in the following steps.

In the second step, the SOM algorithm is applied to project the high-dimensional SHAP value data onto a two-dimensional grid. In this case, standardization of the input features is omitted to preserve the original scale and interpretability of the SHAP values. The SOM architecture is defined by selecting a hexagonal grid structure and determining an appropriate grid size (for illustrative purposes, one might consider a 10×10 grid), which specifies the number of neurons and, consequently, the granularity of the initial data partitioning. To determine the optimal grid size, various configurations are evaluated using established quality metrics: the quantization error is used to measure the average distance between each data point and the nearest neuron in the SOM grid, thereby assessing the fidelity of the data representation; the topographic error evaluates how well the SOM preserves the topological relationships inherent in the original feature space; and the Unified Distance Matrix (U-Matrix) offers a visual assessment of cluster boundaries and separability within the map. A suitable grid size is indicated by lower quantization and topographic errors, along with well-defined and interpretable cluster boundaries, as revealed by the U-Matrix (Forest et al., 2020).

In the third step, a clustering algorithm (e.g., k-means, k-medoids, or hierarchical clustering) is applied to further delineate well-separated clusters within the SOM space. Various algorithmic configurations can be evaluated to identify the clustering partition that most effectively distinguishes neurons associated with crashes predicted to result in fatal or serious injuries from those linked to non-serious outcomes.

Finally, we introduce two complementary visualization tools to facilitate a deeper understanding of the resulting clusters. First, for each cluster, we present a set of figures that display the average SHAP values across all features, the mean values of the original features, and a boxplot of the predicted probabilities for fatal or serious injury outcomes. Second, a decision tree is constructed using the cluster labels (as assigned in the third step) as the response variable and the original features as predictors. The tree is grown without prior pruning constraints to reveal the underlying decision rules that define each cluster. This approach provides interpretable insights into the typical SHAP value patterns and crash characteristics associated with each group of crashes.

For this approach, we used the *kohonen* and *aweSOM* R packages.

4. Results

4.1. Supervised ML model

Following the methodologies outlined in Section 3.1, the optimal supervised ML model is selected based on standard performance

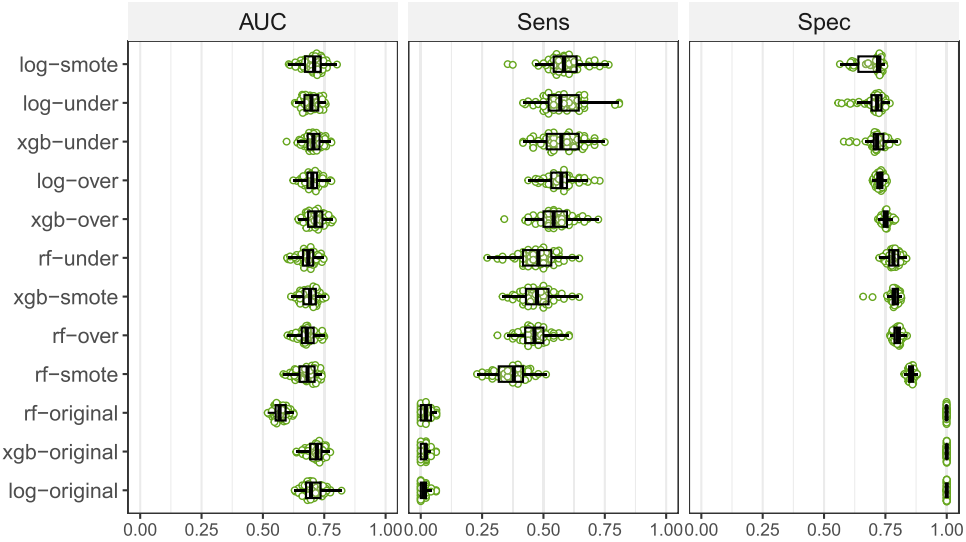

Fig. 1. Performance metrics (AUC, sensitivity, and specificity) of logistic regression (log), random forest (rf), and extreme gradient boosting (xgb) models under original, undersampling (under), and oversampling (over) settings. Boxplots (median and IQR) are overlaid with beeswarm points showing individual observations.

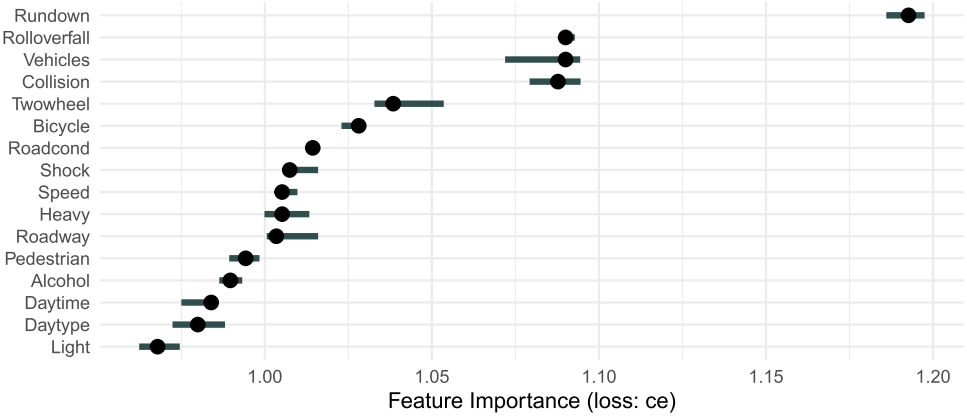

Fig. 2. Feature importance (FI) of the RF model.

metrics. As shown in Fig. 1, although the AUC values are comparable across models, the most favorable balance between specificity and sensitivity is achieved by the RF model with undersampling or the XGB model with oversampling. This figure clearly shows that models without any balancing technique perform noticeably worse, underscoring the importance of applying such adjustments.

In the remainder of this section, we emphasize the usability of global xAI techniques to interpret, understand, and build trust in the fitted RF model with the undersampling technique (which was selected as one of two equally suitable models for the following steps, without a specific methodological preference over XGB model with oversampling). Using the testing dataset, the model achieves an accuracy of 77.38%, a sensitivity (FS cases) of 49.51%, and a specificity (NS cases) of 78.08%. In addition, throughout this analysis, model predictions refer to the estimated probability of a crash resulting in a FS outcome. For this model, the overall mean predicted probability is 0.3, which contributes to a notable number of crashes being incorrectly classified as FS crashes.

As illustrated in Fig. 2, the FI scores identify the key variables that influence the prediction of the severity of the crash in our model. The x-axis shows the increase in prediction error after permuting each feature, normalized relative to the original error; a value of 1 indicates no change in error (no importance). In particular, features such as *Rundown*, *Rolloverfall*, *Vehicles*, and *Collision* emerge as critical factors.

However, FI scores do not provide information on the direction or nature of the influence that key variables exert on the model predictions. ALE plots overcome this limitation by depicting the average effect of each feature on the model predictions. Fig. 3 presents the ALE plots for all the features of the RF model. In our case, the ALE plots show how the model predictions (probability of FS crash) change, on average, with different values of a categorical feature. For example, as illustrated by the ALE plot for the *Rundown* feature, the presence of a rundown event in a crash is associated with an increase of approximately 0.3 in the average predicted probability, compared to the global mean prediction.

Focusing on the most influential features, the presence of excessive speed, a rundown event, a shock impact, or an elevated blood alcohol concentration is associated with an increased probability of a crash resulting in a fatal or serious outcome. To a lesser extent, crashes that occur during weekends, at nighttime, in the fast lane, or involving a single vehicle also exhibit a higher probability of leading to an FS crash.

It is important to note that the FI scores presented in Fig. 2 reflect both the main effects (the direct influence of individual features) and the interaction effects (the combined influence of a feature with others) on model performance. To assess the strength of these interaction effects, we can employ the H-statistic proposed by Friedman and Popescu (2008). For this case, the analysis does not reveal any substantial interaction effects that warrant further investigation. However, if relevant interactions were identified, second-order ALE plots, which isolate

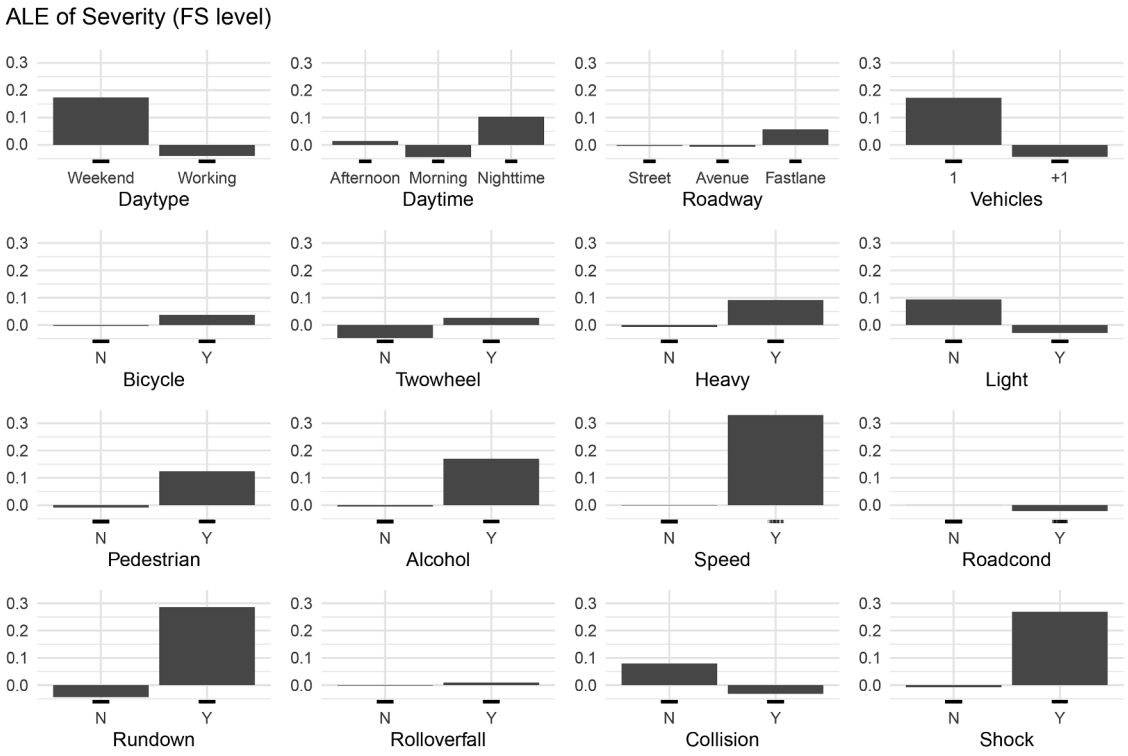

Fig. 3. Accumulated local effects (ALE) for all features of the RF model.

interaction effects by removing the main effects, would be used to estimate and interpret the joint impact of feature pairs on the model predictions.

After introducing the most common global xAI techniques, SHAP values are a commonly used local xAI method to understand the importance of characteristics in individual predictions. Each SHAP value quantifies the contribution of a feature to the disparity between the prediction of the model for a single instance and the overall average prediction (30% in our analysis).

At this point, rather than using SHAP values to explain individual predictions, we take an initial step toward their application in global interpretation (see the next section). Specifically, Fig. 4 presents the average SHAP values across all observations in the testing dataset, grouped according to the four categories of the confusion matrix. In this context, each panel can be interpreted as representing the SHAP values of a prototypical crash within each confusion matrix category.

Fig. 4 shows that, as expected, crashes predicted by the model as FS exhibit similar average SHAP values, which are predominantly positive. Likewise, crashes predicted as NS display comparable patterns, with mostly negative SHAP values. The boxplots of predicted FS probabilities further indicate that misclassifications tend to occur when predicted probabilities are closer to the 0.5 threshold. However, no clear patterns emerge that would help explain which features most influence the model's misclassifications or overall accuracy. Although a more granular analysis of SHAP value distributions for misclassified cases could offer further insights, this level of detail is beyond the scope of the current study. An example of such an approach can be found in Fig. 6 of Bermúdez et al. (2023).

4.2. Supervised clustering using SOM

Following the methodology described in Section 3.2, the SHAP values of the features for all crashes in the testing dataset were used to train the SOM model. Several hexagonal grid configurations were evaluated using standard quality metrics, including Quantization Error (QE), Topographic Error (TE), Kaski-Lagus Error (KLE), and the standard deviation of the U-Matrix values (UMatrixSD). As these metrics capture different and sometimes competing aspects of SOM performance, no single configuration can be considered universally "optimal". QE and TE assess quantization and topology preservation, respectively, while KLE combines both criteria (lower values preferred), whereas UMatrixSD reflects structural contrast and separation between neighboring neurons (higher values preferred), which is particularly relevant for cluster interpretability. To compare metrics with opposite directions and different scales, KLE and UMatrixSD were min-max normalized, and a composite indicator (CombinedScore) was computed as normalized UMatrixSD minus normalized KLE, assigning equal weight to both. Based on the results reported in Table 2, a 7×4 grid (28 neurons) with a neighborhood radius of 3.25 was selected. Although some configurations had lower QE or TE, the 7×4 grid offered stronger U-Matrix differentiation with competitive topology preservation, reflecting a trade-off favoring interpretability over strict error minimization.

Table 2
Top 8 alternative grid configurations using different standard metrics (QE: Quantization Error, TE: Topographic Error, KLE: Kaski-Lagus Error -combining aspects of QE and TE-, UMatrixSD: standard deviation of the U-Matrix values, CombinedScore: Min-Max Scale of UMatrixSD minus Min-Max Scale of KLE).

Grid size	QE	TE	KLE	UMatrixSD	CombinedScore
7x4	2.91	0.26	5.78	4.34	0.73
4x6	2.69	0.25	5.50	4.05	0.70
7x5	2.29	0.24	5.54	4.05	0.68
6x5	2.57	0.34	6.02	3.95	0.44
6x6	2.32	0.23	5.39	3.43	0.43
4x5	3.26	0.27	7.04	4.54	0.31
4x7	2.59	0.29	6.27	3.87	0.29
5x5	2.73	0.28	6.70	4.07	0.21

After training the SOM model using SHAP values, we applied three clustering algorithms (k-means, k-medoids, and hierarchical clustering) to identify subtypes of crash scenarios that exhibit similar patterns of feature contributions to model predictions. To determine the

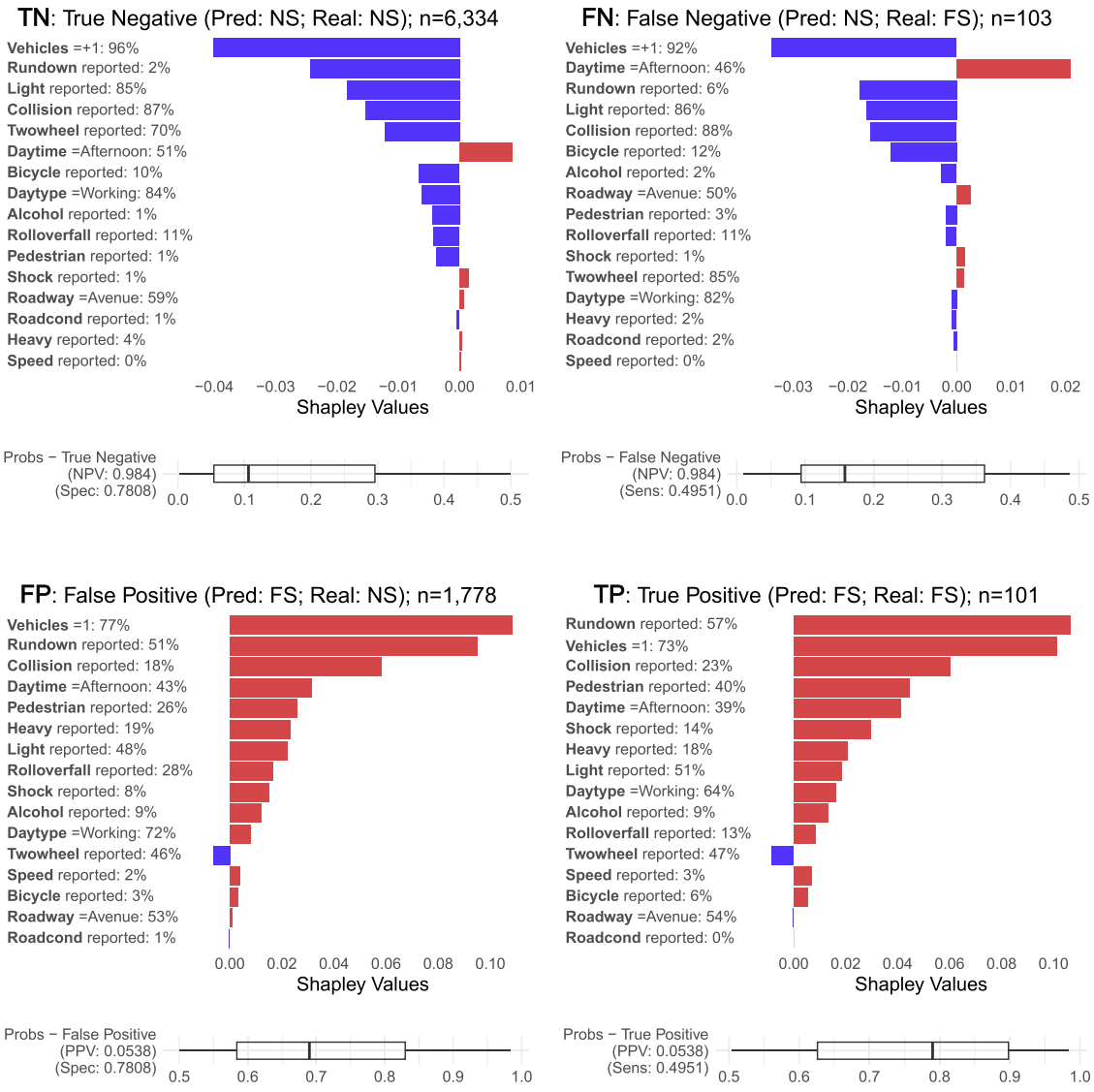

Fig. 4. Average SHAP values across all observations in the testing dataset, stratified by confusion matrix category. Boxplots show the distribution of predicted FS probabilities and the left columns display feature percentages: for binary (Y/N) variables, the presence (Y) percentage; for all others, the most frequent level.

optimal number of clusters for hierarchical clustering, we used the elbow method. The resulting SOM map, shown in Fig. 5, shows 10 distinct clusters that represent different types of crashes. To evaluate the effectiveness of the SOM model in distinguishing between crashes predicted as FS and NS, Fig. 5 groups the clusters into two regions based on their observed FS rate relative to the overall FS rate in the testing dataset (2.5%). Clusters with an observed FS rate below the dataset average are shown in blue and labeled as NS clusters, whereas clusters with an observed FS rate above the dataset average are shown in red and labeled as FS clusters. Notably, the red clusters (1, 4-7, 10) exhibit an average predicted FS probability of 63.25%, while the blue clusters (2-3, 8-9) show a substantially lower average predicted FS probability of 23.98%.

To further interpret the characteristics of each cluster, Figs. 6 and 7 present the average SHAP values for all observations within each cluster, as was done in Fig. 4 for each cell of the confusion matrix.

Finally, we build a decision tree to visually represent the classification rules that characterize each cluster. As described in Section 3.2, the tree is trained using the original crash-related variables as input features and the SOM assigned cluster labels as the target variable. Although the primary objective is interpretability rather than predictive accuracy, we applied post-pruning by removing terminal nodes with fewer than n = 30 observations to avoid overinterpretation of splits supported by very small samples. The resulting decision tree, shown in Fig. 8, facilitates the identification of distinguishing rules associated with each cluster.

It should be emphasized that Figs. 6 (or 7) and 8 together provide a more comprehensive understanding of the decision-making process of the RF model, revealing distinct, yet complementary views of its predictive logic.

For illustrative purposes, we focus on the cluster with the highest number of observations (Cluster 2). As shown in Fig. 6, the model predicted 5,074 out of 5,209 crashes as NS. According to SHAP values, the most influential feature for this cluster is *Vehicles*, followed by *Rundown*, *Light*, *Collision*, and *Rolloverfall*, all of which have a negative impact on prediction, thus reducing the predicted probability of FS. With a predicted median probability of FS below 0.1, the cluster achieves very high levels (0.98) of NPV and specificity. Based on the feature values, this cluster consists exclusively of crashes involving multiple vehicles that did not result in a rundown. Most of these incidents involve light or two-wheel vehicles and are characterized by collisions without rollover or fall. Neither excessive speed nor elevated blood alcohol concentration is reported.

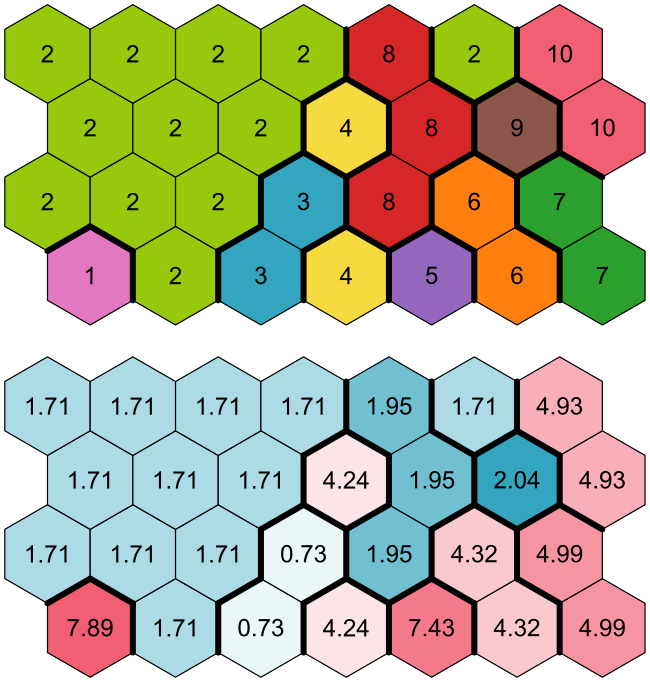

Fig. 5. Top: SOM map with clusters for all observations in the testing dataset. Bottom: The same SOM map, with clusters grouped by their observed FS rate relative to the testing dataset average (2.5%): blue indicates below-average (NS) clusters and red indicates above-average (FS) clusters. The observed FS percentage in each cluster is reported.

Fig. 8 reveals a primary terminal node that covers most of the cases in this cluster, along with three additional terminal nodes that contain relatively few observations. The main node accounts for 4,842 crashes, with a misclassification rate within the decision tree of only 0.2% (1-4,842/4,852). The decision rule associated with this node closely aligns with the feature-based description provided earlier: crashes resulting from a collision; without involvement of bicycles; no reported elevated blood alcohol concentration; no rollover, fall or shock; and no heavy vehicles. The remaining three terminal nodes, which show somewhat different characteristics, may correspond to outliers in the predicted probability distribution shown in the boxplot in Fig. 6, and merit further analysis.

5. Model and cluster analysis

In this section, we present a summary of the key insights derived from the RF model discussed previously, with an emphasis on their implications for improving road safety strategies aimed at reducing traffic-related fatalities and serious injuries.

Firstly, in relation to the confidence that can be placed in the RF model, Figs. 1 and 4 suggest that the model has difficulty identifying enough patterns to reliably distinguish between the two crash states. Although global accuracy is around 77%, there exists a pronounced disparity between the rate of correct predictions for FS crashes (Sens: 49.51%) and the same rate for NS crashes (Spec: 78.08%). Similarly, the proportion of correctly classified NS cases among predicted NS cases (NPV: 98.4%) is very far from the corresponding proportion for FS cases (PPV: 5.38%). These findings highlight the intrinsic complexity of injury severity prediction and the limitations of using supervised ML models for such classification tasks.

False positive cases deserve closer examination, as they contribute significantly to the limited confidence in the predictions of the RF model when a crash is classified as FS. These false positives arise because the RF model assigns an overall mean probability of 0.3 to FS crashes, substantially higher than their actual proportion in the dataset (0.025). However, as illustrated in Fig. 4, both true and false positive cases exhibit similar average SHAP values, making them difficult to distinguish. A conservative interpretation might consider these false positives as potential at-risk crashes that warrant attention similar to true positive cases. In contrast, the predictions of the RF model for NS crashes can be regarded with high confidence, with a small error rate of 1.6% likely attributable to unobserved factors, such as subtle crash details, that the model fails to capture and which may influence the severity of the injuries.

In general, insights from global xAI techniques (see Figs. 2 and 3) indicate that the most influential factors driving the model predictions are primarily related to the type of crash (such as *Rundown*, *Rolloverfall* and *Collision*), as well as vehicle characteristics (including *Vehicles*, *Bicycle*, *Twowheel* and *Heavy*). To a lesser extent, several categorical variables linked to spatial and causal features (such as *Roadcond*, *Roadway*, *Speed*, *Pedestrian* and *Alcohol*) also contribute to the model's decisions, aligning with findings reported in the existing literature (Amini et al., 2022; Santos et al., 2022).

To overcome the limited accuracy of the supervised ML models to predict the severity of injuries, as well as the restricted analytical precision of the current global xAI methods to interpret these models, we propose a *supervised clustering* approach using SOM, as described in Section 3.2. Our approach aims to identify groups of crashes that share both similar predicted severity levels and comparable individual characteristics, potentially indicating the need for similar road safety interventions. As shown in Figs. 5–7, the clusters exhibit markedly different empirical proportions of FS crashes, supporting confidence in their characterization despite the overall performance of the model. In the following analysis, we summarise the defining features of each cluster based on Figs. 6-8 and propose targeted interventions for each profile, drawing on recommendations from the existing literature.

We begin with the four clusters (2-3 and 8-9) grouped as NS crashes in Fig. 5. All of these clusters primarily contain crashes predicted as NS, although they differ in terms of prediction probability and confidence, consistent with the percentage of observed FS crashes detailed in Fig. 5. According to Fig. 6, clusters 2 and 3 exhibit a median predicted probability of FS below 0.2 and a specificity rate that exceeds 97%. In contrast, clusters 8 and 9 show higher median predicted probabilities, closer to the 0.5 threshold, along with lower specificity rates, particularly in cluster 8, where the specificity drops to 78%.

Cluster 2, previously described in the final paragraphs of Section 4.2, only differs from **Cluster 3** in the presence of bicycles in crashes (see the second left node in Fig. 8): the latter is characterized by the presence of this type of vehicle. Specifically, Cluster 3 consists of two neurons (as shown in Fig. 5): one representing collisions between bicycles and light vehicles, and the other between bicycles and primarily two-wheeled vehicles. The observed proportion of FS crashes involving light vehicles is three times lower than in the latter. This suggests that bicycle-light vehicle collisions rarely result in fatal or serious injuries. However, with a mean predicted probability of 39.23% according to the RF model, collisions between bicycles and two-wheeled vehicles warrant the attention of competent authorities. Planning and implementing protected bike lanes that physically separate bicycles from motorized two-wheelers could help reduce these conflicts and improve safety (Ling et al., 2020; Wegman & Schepers, 2024).

Cluster 8 consists of rollovers or falls involving mainly two-wheeled vehicles and, to a lesser extent, light vehicles. However, these incidents do not coincide with other typical contributing factors to FS crashes, such as the presence of rundowns, shocks, heavy vehicles, excessive speed, or blood alcohol concentration, which are completely absent in this cluster. Although the feature *Rolloverfall* has a positive SHAP value, it is not sufficient alone to classify a crash as FS, highlighting the importance of excluding these additional risk factors from the scene. In addition, a closer examination of the three neurons comprising this cluster (see Fig. 5) reveals that falls involving two-wheeled vehicles, particularly when occurring without the involvement of other vehicles and

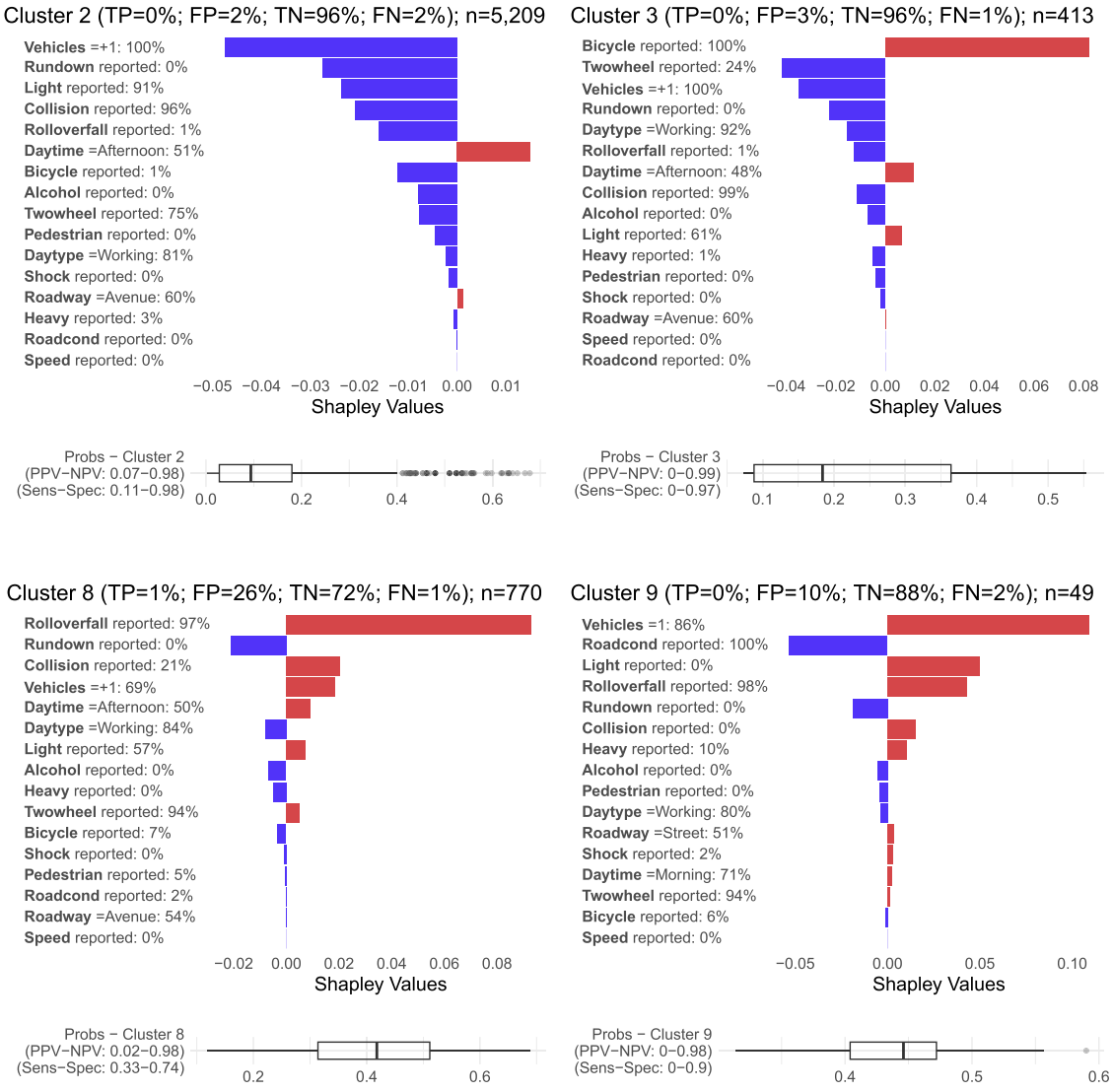

Fig. 6. Average SHAP values for all observations within each NS cluster, accompanied by boxplots showing the distribution of predicted FS probabilities. The left column displays feature percentages for each NS cluster: for binary (Y/N) variables, the presence (Y) percentage; for all others, the most frequent level.

outside of fast lanes, are associated with a higher probability of resulting in a fatal or serious (FS) crash (51% vs. 31%). In this sense, designing motorcycle-friendly road furniture can help reduce this type of crash (Comi et al., 2022; Dadashova et al., 2022).

Cluster 9 represents a distinct but less frequent subset of rollover and fall crashes. Unlike Cluster 8, it does not involve light vehicles and is characterized by poor road conditions identified as contributing to falls involving two-wheeled vehicles, suggesting the need for regular maintenance. Although such road conditions are rare in the overall dataset, their concentration within this cluster reflects the ability of the proposed SOM-based supervised clustering approach to identify coherent risk profiles even when they involve low-frequency events. However, given the limited absolute number of cases, this cluster should be interpreted with caution. Moreover, due to the similarities between Clusters 8 and 9 in terms of observed FS rates and average predicted FS probability, they may alternatively be viewed as a single broader rollover/fall risk profile. This case also highlights the value of the decision tree in Fig. 8, which helps to define the decision rules associated with this particular group of crashes.

Next, let us examine the rest of the clusters, grouped as FS crashes in Fig. 5. These clusters primarily contain crashes predicted as FS, except for Cluster 6, which has more than 75% of cases with a probability of FS below the threshold of 0.5. However, we have included here because the proportion of observed FS crashes (4.32%) is higher than the overall FS rate in the testing dataset (2.5%), as seen in Fig. 5.

In general, excluding Cluster 6, the RF model exhibited limited accuracy in predicting FS crashes, with only 5-12% (PPV) of the predicted FS crashes actually resulting in a FS crash. However, the model successfully identified 90-100% (Sens) of actual FS crashes, allowing us to consider the high number of false positive cases as at-risk cases that deserve appropriate road safety strategies to prevent future FS crashes. By leveraging the insights derived from the clustering analysis presented below, competent authorities can design and implement more targeted and efficient strategies.

Cluster 1 represents a small subset of FS cases characterized by excessive speed, identified by police reports as a contributing factor. This highlights the critical role of targeted interventions to reduce speeding, such as traffic calming measures. Additional distinguishing features of this cluster, as indicated in Fig. 7, refer to the temporal and spatial context of the crashes. Specifically, a relatively high proportion of incidents occurred at night and in fast lanes, suggesting the need for specific (time

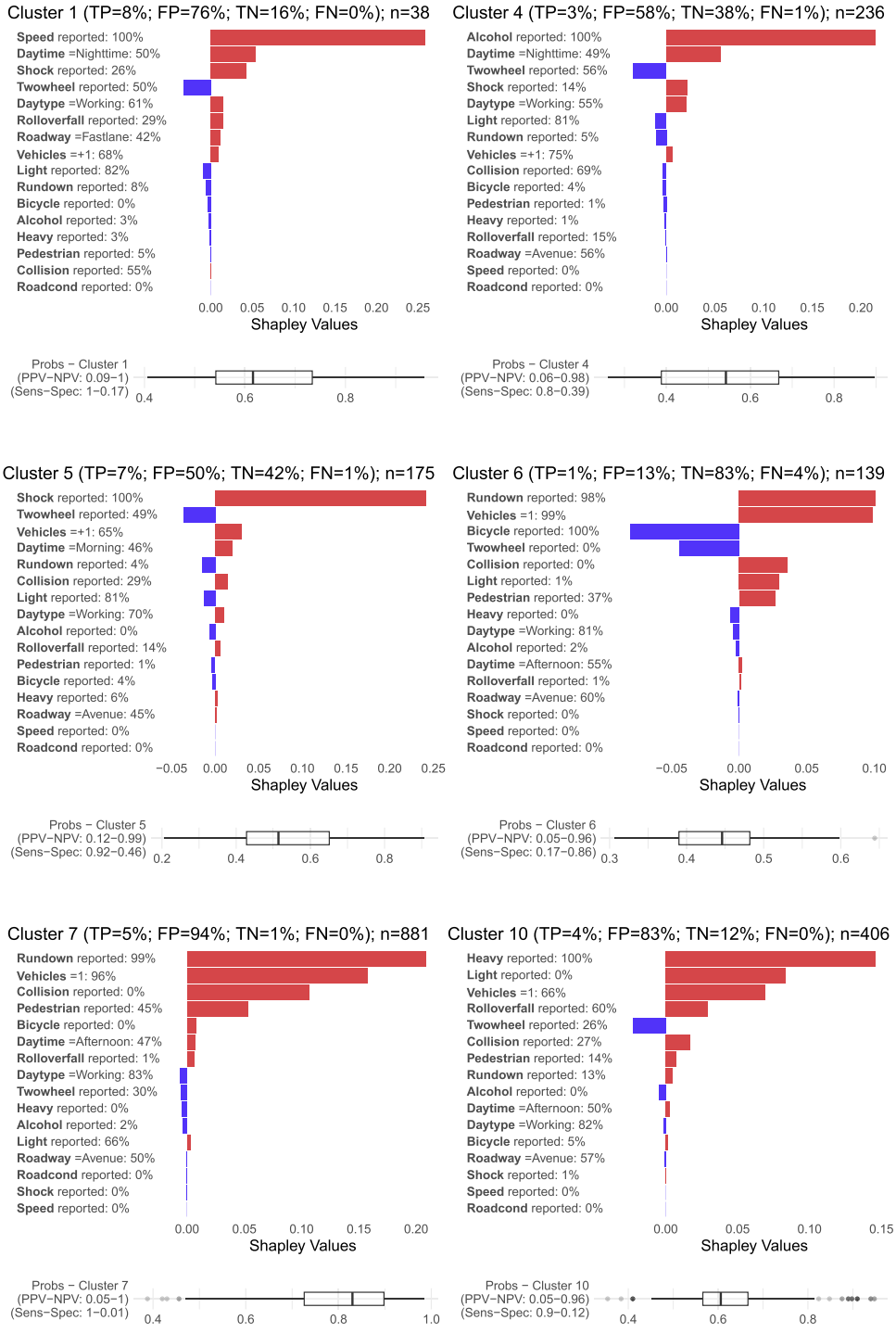

Fig. 7. Average SHAP values for all observations within each FS cluster, accompanied by boxplots showing the distribution of predicted FS probabilities. The left column displays feature percentages for each FS cluster: for binary (Y/N) variables, the presence (Y) percentage; for all others, the most frequent level.

and location) speed control strategies (Ding et al., 2019; Doecke et al., 2020).

Cluster 4, considerably larger than Cluster 1, corresponds to FS cases characterized by elevated blood alcohol concentration as reported by the police. According to Fig. 7, these incidents are more likely to occur on weekend nights, raising the question of whether this pattern reflects the timing and location of police-administered alcohol breath tests. However, as with the previously discussed speed-related issue, it remains essential to strengthen interventions aimed at reducing FS crashes caused by alcohol consumption in urban areas (EC, 2024; Houwing & Stipdonk, 2014).

Cluster 5 groups crashes involving impacts with stationary objects (*Shock*), primarily by light vehicles, and not associated with the previously discussed risk factors (*Speed* and *Alcohol*). Although these crashes can occur under various conditions, they are more likely to occur during the morning or night hours and on avenues or fast lanes. In such contexts, removing or relocating fixed roadside objects where feasible, or enhancing nighttime visibility by upgrading street lighting near these

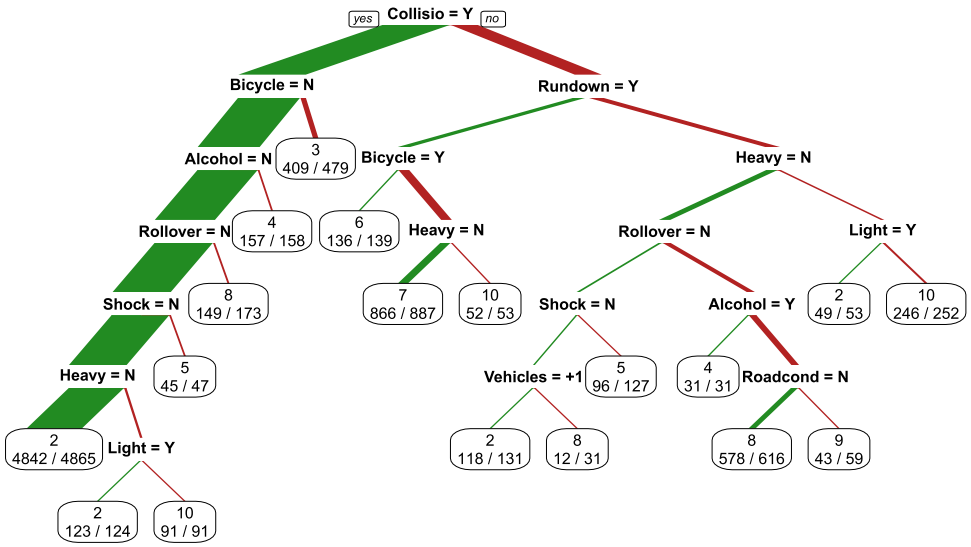

Fig. 8. Decision tree illustrating the decision rules for each cluster. The final nodes include the classification rate at the node, expressed as the number of correct classifications and the number of observations in the node.

objects, may help reduce FS crashes within this group, which exhibits one of the highest observed proportions of FS outcomes (see Fig. 5).

Cluster 7 consists almost entirely of rundown incidents involving light or two-wheeled vehicles. It has the highest median of the predicted probability of FS, although, as shown in Fig. 5, it does not exhibit the highest observed proportion of FS crashes. The cluster consists of two neurons of almost equal size, which are exactly divided according to the value of the *Pedestrian* feature, that is, whether the police attributed the incident to pedestrian behavior or not. The neuron associated with pedestrian-related causes has an observed FS crash rate 2.6 times higher than the other. This difference is also captured by the RF model, with mean predicted probabilities of 90.82% and 75.11%, respectively. In such circumstances, promoting pedestrian education campaigns in high-risk neighborhoods and implementing traffic calming measures in areas with high pedestrian activity may help reduce this type of crash (Batouli et al., 2020; EC, 2023b).

Cluster 6 captures a less common subset of rundown incidents, those involving bicycles as the striking vehicle. Compared to Cluster 7, this type of incident exhibits a lower observed proportion of FS crashes, as illustrated in Fig. 5, resulting in a mean predicted probability of FS below the 0.5 threshold (approximately 45%). However, as previously discussed, we grouped with the FS-predominant clusters due to its elevated observed severity. A closer examination of the two constituent neurons further supports this classification. As in Cluster 7, these neurons are differentiated by the value of the *Pedestrian* feature. In particular, when the incident is attributed to pedestrian behaviour, the observed FS crash proportion rises to 6%. Consequently, to mitigate FS crashes in this context, interventions analogous to those proposed for Cluster 7 may be effective along with the development of protected bike lanes that physically separate bicycles and pedestrians (Teschke et al., 2012).

Cluster 10 comprises exclusively crashes involving heavy vehicles, with no participation of light vehicles. As shown in the corresponding boxplot of predicted FS probabilities in Fig. 7, the data exhibit substantial variability, characterized by a wide range and outliers on both ends. This pattern suggests a further examination of the two constituent neurons within the cluster. The first neuron, which refers to rundown incidents where heavy vehicles are the striking vehicle and the crash is attributed to pedestrian behavior, presents the highest observed proportion of FS crashes in the testing dataset (15.38%). It corresponds to the upper end of the boxplot, with a mean predicted FS probability of 82.21%. In contrast, the second neuron, representing collisions between heavy vehicles and bicycles or two-wheeled vehicles, as well as rollovers involving heavy vehicles, exhibits a significantly lower proportion of FS crashes (2.93%) and aligns with the lower end of the boxplot. In particular, to address the riskiest neuron, targeted infrastructure and road design measures can be implemented, such as restricting heavy vehicle access or establishing low-speed zones in areas with high pedestrian density, and improving pedestrian crossings (EC, 2023a; Yang et al., 2021).

6. Conclusions

This study proposes and applies a novel methodology for severity-based pattern detection in urban traffic crashes, combining supervised ML models, xAI techniques, and a SOM-based *supervised clustering* approach. The main objective is to identify interpretable, data-driven crash typologies associated with fatal or serious injury outcomes, thus supporting the design of targeted and effective road safety interventions.

The application of the SOM-based *supervised clustering* approach offered useful insights beyond what traditional supervised ML models or global xAI techniques could reveal on their own, providing a more nuanced understanding of crash typologies. This approach allowed us to group crashes into well-defined clusters based on common severity-related patterns. The resulting 10-cluster typology revealed distinct subgroups of crashes that not only shared common predictive characteristics but also corresponded to real-world traffic scenarios. Some of them align with known high-risk scenarios (e.g., rundowns involving heavy vehicles or excessive speed), and others highlight less obvious but critical patterns, such as falls involving two-wheeled vehicles or collisions between bicycles and motorcycles.

Consequently, high-severity SOM clusters provide a concrete basis for suggesting targeted road safety measures. Crashes related to excessive speed in fast lanes at night call for focused speed enforcement and traffic calming. Those involving alcohol, particularly on weekend nights, highlight the need for strengthened sobriety controls. Incidents with impacts on fixed objects suggest improving roadside infrastructure and lighting. Rundown crashes involving pedestrians, bicycles, or two-wheelers point to the importance of protected pedestrian and cycling infrastructure. Solo falls of two-wheeled vehicles, especially in poor road conditions, underscore the need for dedicated lanes and regular surface maintenance. Meanwhile, patterns involving heavy vehicles in pedestrian areas support restricting their access and improving crossings.

In conclusion, the proposed methodology offers a promising way to improve the interpretability of injury severity prediction models by

identifying high-severity crash profiles to inform more precise and effective urban traffic safety planning. An important advantage of the SOM-based approach is its topology-preserving mapping, which maintains neighborhood relationships among similar crash profiles. This enables the formation of spatially organized and interpretable clusters, revealing subtle gradations of risk that may not be apparent with simpler methods. While the overall policy recommendations (e.g., targeted interventions for high-severity clusters) remain consistent, the SOM-based approach provides clearer and more actionable delineations of subgroups, supporting more nuanced and informed decision-making.

It should be noted that the dataset covers the period up to 2019, intentionally excluding the COVID-19 years to avoid structural distortions in mobility patterns. Since then, changes such as altered commuting behavior, increased micromobility adoption, and Vision Zero-oriented policies may have affected exposure and risk distributions. Although the specific cluster composition could vary under post-pandemic conditions, the proposed methodological framework is independent of mobility trends and can be readily updated with more recent data. Future validation using post-2020 datasets would therefore help assess temporal stability and transferability.

Based on the proposed methodology, future studies could explore a range of enhancements to broaden its applicability. First, testing alternative clustering algorithms and comparing their performance would help assess the robustness of the proposed approach. In particular, evaluating cluster stability under resampling or bootstrap procedures would provide additional insights, especially for rare event subsets. Second, incorporating additional data sources, such as traffic volume or detailed road geometry, could enhance the explanatory power of the clusters. Third, validating the clusters using more recent data or datasets from other cities would provide evidence of generalizability. Fourth, developing a cost-effectiveness framework for prioritizing both interventions and clusters would support practical decision-making under resource constraints. Finally, exploring broader contextual factors, such as temporal changes and resource limitations, could further improve the applicability of the methodology in real-world settings.

Finally, although developed for urban traffic crash severity data, the proposed SOM-based supervised clustering method is fundamentally a topology-preserving, supervised pattern recognition approach. As such, it could potentially be applied to other high-dimensional domains, for example EEG signal analysis (Yu et al., 2020, 2024, 2018, 2025). Domain-specific adaptations and validation would, of course, be required to ensure robust performance.

CRediT authorship contribution statement

Lluís Bermúdez: Conceptualization, Methodology, Formal analysis, Writing – original draft, Supervision; **Isabel Morillo:** Software, Data curation, Validation, Writing – review & editing; **Anna Salazar:** Software, Data curation, Validation, Writing – review & editing.

Code Availability

The code used to process the data is publicly available at: https://github.com/AnnaSalazar/Supervised-Clustering

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work has been partially supported by AGAUR (grants 2021SGR00299 and 2023CLIMA00012).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ajuntament de Barcelona (2025a). Barcelona's open data service. https://opendata-ajuntament.barcelona.cat/en.

Ajuntament de Barcelona (2025b). Reduction in the number of people injured in traffic accidents in the city. https://ajuntament.barcelona.cat/seguretatiprevencio/en/noticies/es-redueix-el-nombre-de-persones-ferides-i-de-sinistres-de-transit-a-la-ciutat-1473802.

Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, *194*, 107378. https://doi.org/10.1016/j.aap.2023.107378

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE jordan international joint conference on electrical engineering and information technology (JEEIT)* (pp. 272–276). https://doi.org/10.1109/JEEIT.2019.8717393

Amini, M., Bagheri, A., & Delen, D. (2022). Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support. *Reliability Engineering & System Safety*, *226*, 108720. https://doi.org/10.1016/j.ress.2022.108720

Arhin, S. A., & Gatiba, A. (2019). Predicting injury severity of angle crashes involving two vehicles at unsignalized intersections using artificial neural networks. *Engineering, Technology & Applied Science Research*, *9*(2), 3871–3880. https://doi.org/10.48084/etasr.2551

Assi, K. (2020). Traffic crash severity prediction - A synergy by hybrid principal component analysis and machine learning models. *International Journal of Environmental Research and Public Health*, *17*(20). https://doi.org/10.3390/ijerph17207598

Batouli, G., Guo, M., Janson, B., & Marshall, W. (2020). Analysis of pedestrian-vehicle crash injury severity factors in Colorado 2006-2016. *Accident Analysis & Prevention*, *148*, 105782. https://doi.org/10.1016/j.aap.2020.105782

Bazarnovi, S., & Mohammadian, A. K. (2024). Addressing imbalanced data in predicting injury severity after traffic crashes: A comparative analysis of machine learning models. *Procedia Computer Science*, *238*, 24–31. https://doi.org/10.1016/j.procs.2024.05.192

Bermúdez, L., Anaya, D., & Belles-Sampera, J. (2023). Explainable AI for paid-up risk management in life insurance products. *Finance Research Letters*, *57*, 104242. https://doi.org/10.1016/j.frl.2023.104242

Chang, L.-Y., & Mannering, F. (1999). Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. *Accident Analysis & Prevention*, *31*(5), 579–592. https://doi.org/10.1016/S0001-4575(99)00014-7

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, J., Liu, P., Wang, S., Zheng, N., & Guo, X. (2025). Prediction and interpretation of crash severity using machine learning based on imbalanced traffic crash data. *Journal of Safety Research*, *93*, 185–199. https://doi.org/10.1016/j.jsr.2025.02.018

Comi, A., Polimeni, A., & Balsamo, C. (2022). Road accident analysis with data mining approach: Evidence from Rome. *Transportation Research Procedia*, *62*, 798–805. https://doi.org/10.1016/j.trpro.2022.02.099

Cooper, A., Doyle, O., & Bourke, A. (2021). Supervised clustering for subgroup discovery: An application to COVID-19 symptomatology. In *Machine learning and principles and practice of knowledge discovery in databases* (pp. 408–422). Cham: Springer. https://doi.org/10.1007/978-3-030-93733-1_29

Dadashova, B., Silvestri-Dobrovolny, C., Chauhan, J., Perez, M., & Bligh, R. (2022). Hotspot analysis of motorcyclist crashes involving fixed objects using multinomial logit and data mining tools. *Journal of Transportation Safety & Security*, *14*(7), 1201–1219. https://doi.org/10.1080/19439962.2021.1898070

Ding, C., Rizzi, M., Strandroth, J., Sander, U., & Lubbe, N. (2019). Motorcyclist injury risk as a function of real-life crash speed and other contributing factors. *Accident Analysis & Prevention*, *123*, 374–386. https://doi.org/10.1016/j.aap.2018.12.010

Doecke, S. D., Baldock, M. R. J., Kloeden, C. N., & Dutschke, J. K. (2020). Impact speed and the risk of serious injury in vehicle crashes. *Accident Analysis & Prevention*, *144*, 105629. https://doi.org/10.1016/j.aap.2020.105629

EC (2023a). Road Safety Thematic Report - Professional drivers of trucks and buses. Technical Report European Road Safety Observatory. Brussels, Directorate General for Transport. https://road-safety.transport.ec.europa.eu/document/download/e19cf119-eed4-4cb3-b1fd-1fd0b4554992_en?filename=Road_Safety_Thematic_Report_Professional_drivers_trucks_and_buses_2023.pdf.

EC (2023b). Thematic report - Pedestrians. Technical Report European Road Safety Observatory. Brussels, Directorate General for Transport. https://road-safety.transport.ec.europa.eu/document/download/af69c2f8-124e-42d7-9e2c-9c02e01c0538_en?filename=ERSO-TR-Pedestrians-20240528.pdf.

EC (2024). Annual statistical report on road safety in the EU, 2024. https://road-safety-charter.ec.europa.eu/content/annual-statistical-report-road-safety-eu-2024-0.

Forest, F., Lebbah, M., Azzag, H., & Lacaille, J. (2020). A survey and implementation of performance metrics for self-organized maps. https://doi.org/10.48550/arXiv.2011.05847

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954. https://doi.org/10.1214/07-AOAS148

Gramegna, A., & Giudici, P. (2020). Why to buy insurance? An explainable artificial intelligence approach. *Risks*, *8*(4), 137. https://doi.org/10.3390/risks8040137

Hasheminejad, S. H.-A., Zahedi, M., & Hasheminejad, S. M. H. (2018). A hybrid clustering and classification approach for predicting crash injury severity on rural roads. *International Journal of Injury Control and Safety Promotion*, *25*(1), 85–101. https://doi.org/10.1080/17457300.2017.1341933

Houwing, S., & Stipdonk, H. (2014). Driving under the influence of alcohol in the Netherlands by time of day and day of the week. *Accident Analysis & Prevention*, *72*, 17–22. https://doi.org/10.1016/j.aap.2014.06.004

Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*, *171*, 109112. https://doi.org/10.1016/j.ijar.2023.109112

Huysmans, J., Baesens, B., Vanthienen, J., & Van Gestel, T. (2006). Failure prediction with self organizing maps. *Expert Systems with Applications*, *30*(3), 479–487. https://doi.org/10.1016/j.eswa.2005.10.005

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis: An International Journal*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Karimi, E., Haghighi, F., Sheykhfard, A., Azmoodeh, M., & Shaaban, K. (2023). Self-organized neural network method to identify crash hotspots. *Future Transportation*, *3*(1), 286–295. https://doi.org/10.3390/futuretransp3010017

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps (vol. 43). Biological Cybernetics. https://doi.org/10.1007/BF00337288

Kohonen, T. (1995). Self-organizing maps. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-97610-0

Ling, R., Rothman, L., Cloutier, M.-S., Macarthur, C., & Howard, A. (2020). Cyclist-motor vehicle collisions before and after implementation of cycle tracks in Toronto, Canada. *Accident Analysis & Prevention*, *135*, 105360. https://doi.org/10.1016/j.aap.2019.105360

Liu, S., Li, Y., & Wei, F. (2022). Mixed logit model based diagnostic analysis of bicycle-vehicle crashes at daytime and nighttime. *International Journal of Transportation Science and Technology*, *11*(4), 738–751. https://doi.org/10.1016/j.ijtst.2021.10.001

Lopez-Muley, C., Continente, X., Ferrer Fons, M., Guxens, M., Cortès, E., Pérez, K., & López, M. J. (2025). An urban intervention to transform the school surroundings in Barcelona: A mixed-methods evaluation of "let's protect schools" effects on road safety, street liveliness, and wellbeing. *Cities*, *165*, 106164. https://doi.org/10.1016/j.cities.2025.106164

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/arXiv.1705.07874

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, *11*, 1–16. https://doi.org/10.1016/j.amar.2016.04.001

Pal, C., Hirayama, S., Narahari, S., Jeyabharath, M., Prakash, G., & Kulothungan, V. (2018). An insight of World Health Organization (WHO) accident database by cluster analysis with Self-Organizing Map (SOM). *Traffic Injury Prevention*, *19*(sup1), S15–S20. https://doi.org/10.1080/15389588.2017.1370089

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Feature attributions and counterfactual explanations can be manipulated. https://arxiv.org/abs/2106.12563.

Suarez-del Fueyo, R., Junge, M., Lopez-Valdes, F., Gabler, H. C., Woerner, L., & Hiermaier, S. (2021). Cluster analysis of seriously injured occupants in motor vehicle crashes. *Accident Analysis & Prevention*, *151*, 105787. https://doi.org/10.1016/j.aap.2020.105787

Taamneh, M., Taamneh, S., & Alkheder, S. (2017). Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. *International Journal of Injury Control and Safety Promotion*, *24*(3), 388–395. https://doi.org/10.1080/17457300.2016.1224902

Teschke, K., Harris, M. A., Reynolds, C. C. O., Winters, M., Babul, S., Chipman, M., Cusimano, M. D., Brubacher, J. R., Hunte, G., Friedman, S. M., Monro, M., Shen, H., Vernich, L., & Cripton, P. A. (2012). Route infrastructure and the risk of injuries to bicyclists: A case-crossover study. *American Journal of Public Health*, *102*(12), 2336–2343. https://doi.org/10.2105/AJPH.2012.300762

Vorko-Jović, A., Kern, J., & Biloglav, Z. (2006). Risk factors in urban road traffic accidents. *Journal of Safety Research*, *37*(1), 93–98. https://doi.org/10.1016/j.jsr.2005.08.009

Wang, C., Abdel-Aty, M., Han, L., & Easa, S. M. (2024). Analyzing speed-difference impact on freeway joint injury severities of Leading-Following vehicles using statistical and data-driven models. *Accident Analysis & Prevention*, *206*, 107695. https://doi.org/10.1016/j.aap.2024.107695

Wang, D., Liu, Q., Ma, L., Zhang, Y., & Cong, H. (2019). Road traffic accident severity analysis: A census-based study in China. *Journal of Safety Research*, *70*, 135–147. https://doi.org/10.1016/j.jsr.2019.06.002

Wegman, F., & Schepers, P. (2024). Safe system approach for cyclists in the Netherlands: Towards zero fatalities and serious injuries? *Accident Analysis & Prevention*, *195*, 107396. https://doi.org/10.1016/j.aap.2023.107396

WHO (2023). Global status report on road safety 2023. https://www.who.int/publications/b/68866.

Yang, C., Chen, M., & Yuan, Q. (2021). The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis & Prevention*, *158*, 106153. https://doi.org/10.1016/j.aap.2021.106153

Yu, H., Lei, X., Song, Z., Liu, C., & Wang, J. (2020). Supervised network-based fuzzy learning of EEG signals for alzheimer's disease identification. *IEEE Transactions on Fuzzy Systems*, *28*(1), 60–71. https://doi.org/10.1109/TFUZZ.2019.2903753

Yu, H., Li, F., Liu, J., Liu, D., Guo, H., Wang, J., & Li, G. (2024). Evaluation of acupuncture efficacy in modulating brain activity with periodic-aperiodic EEG measurements. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 2450–2459. https://doi.org/10.1109/TNSRE.2024.3421648

Yu, H., Wu, X., Cai, L., Deng, B., & Wang, J. (2018). Modulation of spectral power and functional connectivity in human brain by acupuncture stimulation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(5), 977–986. https://doi.org/10.1109/TNSRE.2018.2828143

Yu, H., Zeng, F., Liu, D., Wang, J., & Liu, J. (2025). Neural manifold decoder for acupuncture stimulations with representation learning: An acupuncture-brain interface. *IEEE Journal of Biomedical and Health Informatics*, *29*(6), 4147–4160. https://doi.org/10.1109/JBHI.2025.3530922

Zhang, S., Khattak, A., Matara, C. M., Hussain, A., & Farooq, A. (2022). Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. *PLOS ONE*, *17*(2), 1–19. https://doi.org/10.1371/journal.pone.0262941

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., & Wang, Z. (2019). Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access*, *7*, 39897–39910. https://doi.org/10.1109/ACCESS.2019.2903319