Array 30 (2026) 100743

Contents lists available at ScienceDirect

Array

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

Studying the impact of data preprocessing, hyperparameter tuning and machine learning algorithms in crash prediction explainability

Jon Díaz-Aparicio *, Erick Rodríguez-Esparza, Jenny Fajardo-Calderín, Enrique Onieva

DeustoTech, Faculty of Engineering, University of Deusto, Av. Universidades, 24, Bilbao, 48007, Biscay, Spain

ARTICLE INFO

Dataset link: https://www.openstreetmap.org, https://datos.madrid.es/portal/site/egob, https://www.vianova.io, https://soteriaproject.eu, https://www.nommon.es, https://soteriaproject.eu, https://github.com/Jondiii/crash_explainability.git

Keywords:
Crash prediction
Machine learning
Road safety
Imbalanced learning
Explainable AI

ABSTRACT

Road traffic crashes remain a major global concern, causing more than 1.3 million fatalities each year and underscoring the need for improved tools to understand and predict crash occurrence. This study presents an integrated retrospective crash-risk screening framework that merges four heterogeneous data sources (crash records, road infrastructure, connected vehicle data, and travel demand) to model road-segment crash risk in Madrid. Ten preprocessing configurations are created using oversampling (generate instances of the minority class), undersampling (removing instances of the dominant class), dataset expansion (new data generation), and SMOTE, each tested with and without normalization. Seven machine-learning algorithms (tree ensembles and SVMs) are evaluated under regression, multiclass classification, and binary classification formulations, resulting in a total of 210 experiments. Binary classification delivered the best performance, with Gradient boosting trained on normalized, undersampled data emerging as the strongest model. Subsequent Bayesian hyperparameter optimization further enhanced its predictive capability. Explainable AI analysis using SHAP values revealed that braking events are the most influential predictors of crash likelihood, followed by road length and traffic demand, emphasizing the relevance of driver-behavior indicators in safety modeling. Overall, the findings demonstrate the benefits of integrating traditional crash data with emerging connected vehicle and demand-based information. The study provides evidence that explainable machine learning approaches can effectively support data-driven decision-making for road-safety management and targeted intervention planning.

1. Introduction

The latest statistics released by the World Health Organization report that road traffic crashes cause approximately 1.3 million deaths worldwide every year [1], with more than 20 million people suffering non-fatal injuries. Beyond the direct human toll, these crashes impose substantial social, economic, and emotional costs on victims, their families, and society as a whole. This alarming situation underscores the importance of improving road safety, which requires a clear understanding of the main factors that contribute to crashes and the development of models capable of reliably predicting them.

A prominent area of research in road safety focuses on using machine learning (ML) algorithms to model traffic accidents in three key ways: real-time crash prediction, crash frequency estimation, and injury severity assessment [2]. According to the systematic review conducted by Ali et al. [3] on ML applied to crash prediction, decision tree–based models and neural networks are the main approaches for this task. That study also provides a comparative overview of the ML algorithms considered, to which the interested reader is referred.

Decision trees and their variants, such as random forests (RF) and extra trees, provide easy-to-understand internal structures that help explain their results [4]. These models have demonstrated success in various studies. For example, Iranmanesh et al. [5] showed that RF can accurately detect trends and predict crash frequency, and may even outperform traditional statistical models [6].

Boosting algorithms represent another category within ensemble learning, including Adaptive boosting, Gradient boosting, and Extreme gradient boosting (XGBoost). These algorithms can outperform Bayesian networks [7]. XGBoost, in particular, has shown strong performance in identifying key factors associated with injury severity at pedestrian crossings, as reported by Goswamy et al. [8].

Additionally, support vector machine (SVM) models have also been widely used in all aspects of crash modeling due to their ability to handle complex non-linear classification problems [9]. For example, Yu and Abdel-Aty [10] employed SVMs to model real-time crash occurrences, demonstrating their suitability for dynamic classification tasks. Neural Networks, in contrast, rely on more complex architectures than

* Corresponding author.
E-mail address: jon.diaz@deusto.es (J. Díaz-Aparicio).

https://doi.org/10.1016/j.array.2026.100743
Received 23 December 2025; Received in revised form 16 February 2026; Accepted 4 March 2026
Available online 28 March 2026

the previously discussed methods, and have been shown to outperform traditional approaches when modeling crash risk at intersections [11] or in identifying serial dependencies within the input data [12].

Regardless of the specific algorithm used, model performance often depends heavily on the choice of hyperparameters. Hyperparameter tuning is crucial for optimizing the performance and generalizability of supervised models [13], with grid search and random search being the most common choices for searching the configuration space [14,15], whereas techniques such as Bayesian optimization present a more informed search strategy [14].

The quality and variety of the input data also strongly influence model performance. Police-reported crash data remains the primary data source in most studies [3], although studies such as [16] or [17] indicate that this data is not free from biases due to underreporting.

Recent work also suggests that combining traditional crash data with emerging information sources such as connected vehicle data [18] and driver-behavior indicators [19] can improve crash prediction. At the same time, several authors highlight the need to explore how these heterogeneous data sources can be integrated effectively and how they contribute to predictive performance in road safety applications [20, 21]. Road infrastructure data is also included, as recent studies such as [22] show its applicability in safety estimation.

Crash datasets typically exhibit severe class imbalance (as crashes are rare events compared to normal traffic conditions), which makes model training challenging and often leads to biased predictions. This is also true when analyzing crashes where vulnerable road users (VRU, which includes pedestrians, motorcyclists, and cyclists) are involved. As noted by Ali et al. [3], this is a persistent issue in crash occurrence and injury severity modeling. Undersampling and oversampling techniques are commonly used to mitigate this problem, although each approach presents important limitations. The first may discard valuable information and reduce overall accuracy. In contrast, the second increases the number of minority-class samples but may generate unrealistic instances, depending on the technique used [23].

In addition to these drawbacks, oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) [24] create new minority-class instances that may not correspond to realistic real-world patterns [25]. While resampling methods can improve model performance, no single technique consistently yields the best results across different datasets. Consequently, the choice of resampling strategy should be guided by the specific characteristics of the data and the objectives of the study [26].

Recent advances in explainable artificial intelligence have improved the ability of ML models to provide meaningful insights into the relationships between crashes and their underlying causes [27]. Among available techniques, SHapley Additive exPlanations (SHAP) has become particularly prominent due to its strong theoretical foundation and clear visualization of feature contributions [28]. Crash modeling commonly incorporates explanatory variables related to road and environmental conditions, human behavior, and specific crash and vehicle characteristics, and studies such as [20] frequently identify traffic volume, road length, posted speed limits, and road curvature as some of the most influential factors.

1.1. Research gap

Despite these advances, important gaps remain. First, most existing studies rely primarily on police-reported data and do not systematically evaluate the added value of combining multiple heterogeneous data sources at the road-segment level, including infrastructure characteristics, driver behavior, and travel demand. Second, few studies examine how preprocessing strategies, such as different forms of resampling and normalization, interact with the choice of ML algorithm and with the problem formulation, whether regression, multiclass classification, or binary classification. Third, although explainable AI techniques such as SHAP are increasingly used, there is still limited work on how feature importance changes when key variables or entire data sources are removed, for example, when connected vehicle information is excluded.

This study is among the few that addresses these gaps simultaneously by developing an integrated crash screening framework for urban roads in Madrid, Spain, using four complementary data sources: crash records, road infrastructure information, connected vehicle events, and travel demand data. Ten preprocessing configurations are evaluated, and seven ML algorithms are tested under three problem formulations. The best-performing model is further optimized using Bayesian hyperparameter tuning, and SHAP analysis is employed to identify the most influential predictors and to assess the impact of removing key feature groups. It should be noted that the task is formulated as retrospective segment-level risk screening, in which multi-year crash occurrence (2019–2022) is modeled using infrastructure, behavioral, and demand features observed in 2022, rather than as same-year crash forecasting.

The main contributions of this work are as follows:

- Development of an integrated crash screening framework that merges four heterogeneous datasets at the road-segment level.
- A systematic evaluation of ten preprocessing strategies and seven ML algorithms across three problem formulations, resulting in 210 experiments.
- Bayesian hyperparameter optimization of the best-performing model, followed by a SHAP-based analysis that examines both global feature importance and the effects of removing key features or entire data sources.
- Practical insights for road safety management, highlighting the value of connected vehicle information, especially braking-related events, in improving crash prediction and supporting data-driven intervention planning.

The proposed framework is broadly applicable to urban road networks where heterogeneous data sources are increasingly available. While the case study focuses on Madrid, the methodological approach is transferable to other cities that collect crash records, roadway characteristics, traffic demand data, and, where available, connected vehicle information. The framework is particularly relevant for transportation agencies and city planners seeking to identify high-risk road segments, prioritize safety interventions, and evaluate the added value of emerging data sources in safety analysis. Moreover, the evaluation of preprocessing strategies and modeling formulations makes the approach adaptable to different data availability scenarios, including contexts where connected vehicle data is sparse or unavailable.

1.2. Paper structure

The remainder of the article is structured as follows. Section 2 presents the methodology, including data aggregation, preprocessing, model training, and explainability analysis. Section 3 describes the study area, datasets, and experimental setup. Section 4 reports the main results and also contains a feature importance analysis using SHAP values. Lastly, Section 5 covers the discussion of the model input features based on the previous analysis, and Section 6 summarizes the conclusions and outlines directions for future research.

2. Methodology

The main goal of this study is to identify the most influential factors in crash modeling on urban roads. This section describes the methodological framework adopted to achieve this objective, which is also summarized in Fig. 1.

The process begins with the cleaning and aggregation of geolocated data at the road-segment level, integrating four heterogeneous sources: crash data, infrastructure data, connected vehicle data (CVD), and travel demand data (TDD). The resulting dataset is then preprocessed using multiple strategies, after which several ML algorithms

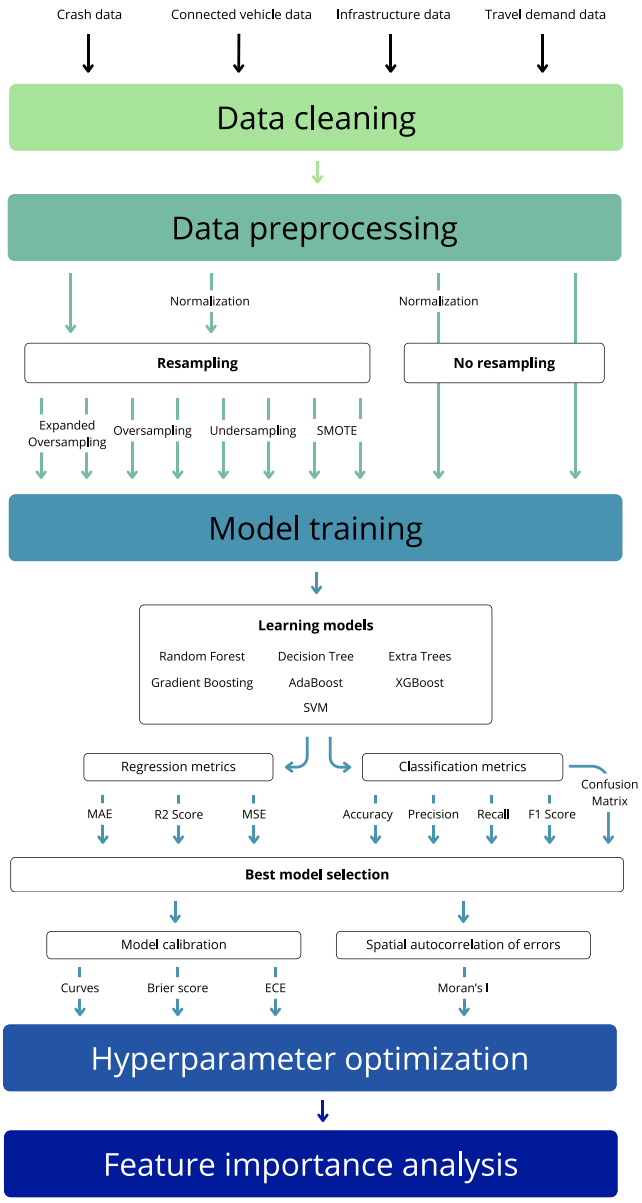

Fig. 1. Methodological framework.

commonly applied to crash prediction are trained and evaluated. The best-performing model is selected through a comparison of performance metrics and confusion matrices, followed by Bayesian hyperparameter optimization. Finally, explainability analysis is conducted to identify the most influential input features and the key factors associated with crash occurrence.

Despite crash screening being commonly formulated as a classification problem, the authors also wanted to test the models under a regression context. In a regression setting, the objective is to estimate the expected number of crashes for each road segment. In a classification setting, the goal is to assign each road to a crash-frequency category or to determine whether at least one crash is expected to occur. In this paper, both approaches are tested and compared.

2.1. Data sources

Four data sources are used to construct the experimentation dataset. Infrastructure data describes physical and regulatory characteristics of each road segment, including the presence of pedestrian crossings, give-way or stop signals, traffic lights, surface type, road length, and number of lanes. Dividing roads into subsegments was initially considered and eventually discarded, as previous studies report limited predictive benefits from segmentation [29,30].

Crash data contains records from official institutions detailing crash type (e.g., collision, side-impact, fall), date and time, and injury severity of the involved road users.

Vehicle events registered by sensors during driving form the core of CVD. These sensors provide real-time information on driver behavior and the actions taken by a particular vehicle during driving, and include events such as steering, braking, and acceleration. A static dataset containing said information, as well as the start and end times, and the intensity of the events, is used in this research.

Lastly, TDD contains the traffic volume per hour of the day within a particular area. The data is used to calculate the average annual daily traffic (AADT) per road, which gives an approximate number of the traffic volume the city experiences on a daily basis, following Eq. (1), where $AADT_{r,t}$ is the AADT on road segment r for vehicle type t, $V_{r,t,d}$ is the observed travel demand (traffic volume) on road segment r on day d for vehicle type t, and $D_{r,t}$ is number of days with available traffic data for road segment r and vehicle type t. Independent values are obtained for private and micromobility vehicles.

$$\mathrm{AADT}_{r,t} = \frac{1}{D_{r,t}} \sum_{d=1}^{D_{r,t}} V_{r,t,d} \tag{1}$$

All the datasets are georeferenced, and they are aggregated to the roads within the area of interest, resulting in one single, unique cleaned dataset that contains all the aforementioned information per road.

2.2. Preprocessing techniques

ML models often exhibit biased performance when trained on imbalanced datasets, where the number of observations belonging to one class significantly exceeds that of others. Crashes are relatively rare events, which makes retrospective predictions particularly challenging. To mitigate this issue, several resampling techniques can be applied before model training once the main training dataset is cleaned. In this work, we compare four different resampling approaches to balance the target variable distributions: basic resampling, resampling with feature expansion, undersampling, and SMOTE-based resampling. Although their end goal is the same, they have subtle differences in how it is achieved.

Basic resampling relies on a bootstrap procedure to randomly duplicate samples from the minority class until both classes are approximately balanced. This method does not introduce new information, as it merely replicates existing samples, but effectively mitigates the dominance of the majority class during model training by preserving the original feature space and data variability, making it a simple yet widely used baseline for class imbalance correction.

The expanded resampling variant extends basic resampling by performing the bootstrap procedure not only on the minority class but also allowing for a mild expansion of the training data to increase the overall sample size. This method enhances model robustness by exposing the learning algorithm to more varied input instances (derived from repeated observations) while preserving the original class ratios after balancing and improving model generalization in small datasets.

In contrast, undersampling reduces the size of the majority class by randomly discarding samples until class frequencies are approximately equal. This method is computationally efficient and helps mitigate overfitting toward the dominant class. However, it can also lead to information loss, as potentially relevant majority-class samples are removed, making it more suitable when model interpretability is prioritized over absolute predictive power.

The SMOTE technique introduced by Chawla et al. [24] is an alternative strategy that generates synthetic samples for the minority

class rather than duplicating existing ones. Using the feature space of minority class instances, SMOTE interpolates between each sample and its nearest neighbors to produce synthetic points that lie along the line segments connecting them, increasing class diversity without simply replicating data.

In addition to resampling techniques, data normalization is also performed. Normalization is a feature scaling procedure introduced by Ioffe and Szegedy [31] that adjusts the range of numerical features to a predefined interval. In this research, given the heterogeneous nature of the input variables, normalization is used to map all numerical values to [0, 1], ensuring that features contribute proportionally during model training.

The four resampling approaches are applied independently, resulting in five datasets when including the original baseline. An additional five datasets are created by applying normalization prior to the aforementioned resampling techniques. In total, ten datasets are used for training: the original dataset, the resampled dataset, the resampled dataset with expanded data, the undersampled dataset, and the SMOTE-based dataset, each available in both normalized and non-normalized versions.

2.3. Model training

Recent studies indicate that ensemble learning methods are among the most reliable ML approaches for retrospective crash screening, although SVMs have also shown strong performance due to their ability to learn from complex spaces [3]. For these reasons, among the most popular choices in the literature, the following algorithms are selected for evaluation:

- **Decision trees** (DT) are non-parametric models that recursively partition the feature space into homogeneous regions based on a sequence of decision rules first introduced by Morgan and Sonquist [32]. They are easy to interpret and handle both numerical and categorical data, although they are prone to overfitting when used as standalone models.
- **Random forest** (RF) is an ensemble learning method introduced by Ho [33] that builds multiple DT during training and aggregates their outputs by taking the mode (for classification) or mean (for regression). Each tree is trained on a bootstrap sample of data and a random subset of features, which introduces diversity, reduces variance, and mitigates overfitting, thereby enhancing the model's robustness and generalization capability.
- **Extremely randomized trees** or **extra trees** (ET), is an ensemble method similar to RF but with greater randomization in the selection of split thresholds and features [34]. This increased randomness helps reduce variance and computational time, often at a small cost to bias, leading to faster and more stable performance.
- **Adaptive boosting**, or AdaBoost, is a boosting technique that combines multiple weak learners into a strong ensemble by iteratively adjusting sample weights [35]. Misclassified instances receive higher weights in subsequent iterations, allowing later learners to focus on the hardest cases.
- **Gradient boosting** builds an ensemble of weak learners in a stage-wise manner, where each new learner minimizes the residual errors of the previous ensemble through gradient-based optimization [36]. This approach allows for highly flexible and accurate models, albeit with increased computational cost and potential sensitivity to overfitting.
- XGBoost (**extreme gradient boosting**) is an optimized and regularized implementation of Gradient boosting introduced by Chen and Guestrin [37], designed to improve computational efficiency, scalability, and model generalization. It incorporates second-order gradient information, regularization, and advanced parallelization strategies, making it one of the most effective and widely used boosting algorithms in ML.
- SVM (**support vector machines**) are supervised learning algorithms that aim to identify the optimal hyperplane separating data points from different classes by maximizing the classification margin. Cortes and Vapnik [38] showed that maximizing this margin leads to strong generalization performance, particularly in high-dimensional feature spaces. Using kernel functions, SVMs can model complex nonlinear decision boundaries by implicitly mapping input data into higher-dimensional spaces.

The selected models are trained on the different datasets described in Section 2.2, and evaluation metrics are computed for each dataset-model combination. Depending on whether the problem is formulated as a regression or a classification problem, the corresponding metrics are used to identify the best-performing model.

To assess regression model performance, three commonly used metrics are used: the coefficient of determination (R2), the Mean Absolute Error (MAE), and the Mean Squared Error (MSE). The R2 score measures the proportion of variance in the dependent variable that is explained by the model, providing an indication of overall model fit. MAE quantifies the average absolute magnitude of prediction errors without considering their direction, while MSE computes the average squared difference between predicted and observed values.

Classification models employ different metrics to test their performance. Accuracy is the proportion of correctly classified instances among all predictions, providing a general indication of performance. However, it tends to overestimate performance when a class dominates the dataset. Precision measures the proportion of predicted positive cases that are truly positive, whereas recall quantifies the proportion of actual positive cases correctly identified by the model. High precision is more desirable when reducing false positives is important, while high recall is preferred when false negatives carry substantial cost, as in the present task. The harmonic mean of the two values is represented by the F1-score, which balances both metrics. In multiclass classification, precision, recall, and F1-score are computed using macro-averaging to assign equal weight to each class.

Beyond regression and classification metrics, we also evaluate the reliability and spatial consistency of model outputs. Calibration curves compare predicted crash probabilities with observed frequencies, allowing us to assess whether the model systematically over- or underestimates risk across probability ranges. This is summarized quantitatively using the Brier score [39], which measures the mean squared error of probabilistic predictions, and the Expected Calibration Error (ECE) [40], which captures deviations between predicted and observed probabilities across bins. In addition, Moran's I [41] is used to test for spatial autocorrelation in the model residuals, verifying whether prediction errors cluster geographically. A low Moran's I indicates that residuals are spatially unstructured, suggesting that the model is not leaving systematic spatial patterns unexplained and therefore does not suffer from strong spatial bias in its errors.

Lastly, confusion matrices are also examined to support model selection, as they provide a detailed breakdown of classification outcomes by showing the counts of true positives, true negatives, false positives, and false negatives. Therefore, they help interpret how different models and thresholds balance risk detection against over-prediction, providing practical insight into their suitability for deployment.

2.4. Hyperparameter optimization

Once the model training process is completed, the best-performing model is selected based on the evaluation metrics. This model, together with the dataset on which it is trained, proceeds to the next stage of the methodology, namely hyperparameter optimization. The objective of this phase is to identify the optimal set of hyperparameters that maximizes model performance. With this aim, various techniques were considered.

Grid search is a commonly used method that exhaustively explores the hyperparameter configuration space, although it tends to be computationally inefficient [14]. In contrast, random search offers a more efficient alternative by sampling hyperparameters randomly from a specified distribution [15]. Bayesian optimization provides a more informed search strategy by "balancing exploration and exploitation" [14], often resulting in reduced computational cost while achieving high-performing configurations, even with limited data.

Hyperparameters govern the learning process but cannot be inferred directly from the data. For this research, Bayesian optimization is employed, as it matches the computational and structural characteristics of our problem. The models tested involve a moderate-to-high-dimensional hyperparameter space across multiple learning algorithms, each requiring costly cross-validated training on a large spatial dataset. Under these conditions, exhaustive grid search becomes computationally prohibitive, while random search, although more efficient, ignores information from previously evaluated configurations and therefore wastes a substantial portion of the evaluation budget.

The Bayesian optimization approach constructs a surrogate probabilistic model of the objective function and selects new hyperparameter configurations to evaluate based on past results, thereby balancing exploration and exploitation [42].

2.5. Feature importance analysis

The final step of the methodology is the feature importance analysis performed on the hypertuned model. This analysis is based on Shapley values, originally introduced by Shapley et al. [43] in cooperative game theory as a mechanism for fairly distributing credit (or blame) among contributing players.

When applied to ML, each input feature is treated as a player contributing to the model's prediction, and these contributions are quantified through the SHAP (SHapley Additive exPlanations) framework [44]. SHAP efficiently computes Shapley-based feature attributions for complex models by evaluating how predictions change across all possible coalitions of features, yielding a consistent estimate of each feature's marginal contribution. These values provide both local interpretability, by explaining individual predictions, and global interpretability, by summarizing feature influence across the entire dataset. In this work, SHAP values are computed for the best-performing hypertuned model to quantify the relative contribution of each input feature to the predicted crash outcomes.

The methodological framework described above defines the complete pipeline used in this study, from data aggregation and preprocessing to model training, hyperparameter optimization, and feature-importance analysis. The next section describes how this pipeline is applied to the Madrid case study, detailing the datasets used, the construction of the experimental scenarios, and the configuration of the machine-learning models.

3. Experimentation

3.1. Data description

Due to data availability, the study area selected for the experimentation corresponds to the city center of Madrid, Spain, where private companies provided CVD and TDD for the year 2022. Table 1 presents a summary of all the data available. All infrastructure data, consisting of urban-road characteristics and network topology, is obtained from Open Street Maps[1] (OSM). As OSM provides georeferenced road-segment geometries, it is used as the spatial foundation onto which all additional datasets are aggregated.

The following driving road infrastructure data are extracted from OSM: road length (m), maximum speed (km/h), number of lanes, road and surface type, traffic lights, stop and give-way signs, pedestrian crossings, cycling paths, public bus stops, street lighting, and side parking. The first three variables are used as-is, while the road and surface types are represented by one-hot encodings. All remaining attributes are encoded as binary indicators to denote whether the element is present on the road segment, with the value 0 indicating that it is not present and 1 indicating that at least one element is present. All in all, almost 2000 edges are used in the experimentation.

Regarding crash data, the dataset used is available through Madrid's Open Data Portal,[2] the official open-data platform of the City Council of Madrid, which provides crash records from 2019 to 2025. Since crashes are rare events and meaningful analysis requires multiple years of data to distinguish isolated occurrences from persistent trends, and because CVD and TDD are only available for 2022, we considered only crashes from 2019 to 2022 within the study area. Per road segment, the number of recorded crashes is aggregated, which is the variable to be predicted.

The CVD used within the study is provided by a private company and, therefore, cannot be disclosed. The dataset consists of timestamped driving events from 2022 captured by vehicle sensors from Madrid. Each entry includes the geospatial location of the event, the date, start and end times, the event type (right and left over-steering, harsh braking, or harsh acceleration), as well as the maximum and average acceleration magnitudes (m/s) associated with the event.

Lastly, TDD is incorporated into the road dataset, and it is also provided by a private company. The data consists of hourly counts of private and micromobility vehicles for each road segment in 2022. From these values, AADT is computed for both private and micromobility users by dividing the total observed demand by the number of days with available records.

Crash data plays a fundamentally different role than CVD and TDD in our framework. While CVD and TDD are used to characterize contemporary driving behavior and exposure patterns at specific locations, crash data are used to capture historical safety outcomes and longer-term risk patterns. In this study, CVD and TDD provide a snapshot of how drivers interact with the road network in 2022, whereas crash records are employed to identify segments that have exhibited elevated crash risk over a multi-year period.

The use of crash data from 2019–2022 is motivated by both statistical and practical considerations. Crash occurrence is a relatively sparse and noisy value at the segment level, and relying on a single year would lead to unstable and highly variable labels. Aggregating over multiple years improves the robustness of the risk signal. At the same time, the COVID-19 pandemic substantially disrupted traffic volumes and mobility patterns during 2020 and parts of 2021, resulting in anomalously low crash counts. Restricting the analysis to 2020–2022 would therefore bias the historical crash distribution downward. Including 2019 allows the crash dataset to better reflect typical pre-pandemic conditions and provides a more representative baseline of underlying safety risk.

Although CVD and TDD are only available for 2022, their purpose in the model is to explain spatial variation in crash risk rather than to reproduce the exact number of crashes in a single year. Using multi-year crash outcomes together with one year of behavioral and demand data, therefore, allows the model to learn stable associations between infrastructure, traffic behavior, and historically observed safety patterns, while mitigating the distortions introduced by the pandemic period.

As mentioned, the infrastructure data serves as the basis for the experimentation dataset and contains a list of all road segments within the area of interest (1956 segments). Intersections are not modeled explicitly, since each segment in the dataset already begins and ends

[1] openstreetmaps.org

[2] datos.madrid.es

Table 1
Summary of input data used during model training, grouped by category and variable type.

Category	Variable type	Variables
Infrastructure Data	Numeric	Traffic lights, give-way and stop signs, pedestrian crossings, cycling paths, bicycle and car parking spaces on the left and right sides, and street lighting.
	Categorical	Road and road surface type, number of traffic lanes, zonal speed, and maximum speed limit.
	Binary	Presence of a roundabout within 25 meters, and public transport stops.
Travel Demand Data	Numeric	AADT for micromobility and private vehicle users.
Connected Vehicle Data	Numeric	Mean, average, and maximum average acceleration values and number of right-turn cornering, left-turn cornering, braking, and speed-up events.

at an intersection. When a physical road crosses multiple intersections, OSM naturally represents it as several shorter segments, which aligns with the segment-level resolution required for this study.

For each road segment, the total number of CVD events by event type, along with their average and maximum acceleration magnitudes, is aggregated. AADT for private and micromobility vehicles is also assigned to each segment, resulting in 14 additional input features. Finally, each segment is assigned the total number of crashes and the number of crashes involving VRU, which serve as the target variables for the ML models. In total, 1724 crashes are assigned (1071 of them with VRU involvement) among 707 road segments. The remaining 1249 roads have no crashes recorded.

In the case of crashes and CVD, they are assigned to the closest segment, as long as the distance between the event and the closest point of the segment is less than or equal to 20 m.

3.2. Model training

Before model training, additional preprocessing is required to address class imbalance and support the learning process of the ML algorithms. Using the resampling and normalization techniques described in Section 2.2, ten distinct versions of the dataset are generated. The prediction task is then formulated in three different ways: regression (estimate the number of crashes per road segment), multi-class classification (assign each segment to a crash-frequency category), and binary classification (determine whether one crash is expected to occur or not in a given road). For multiclass classification, class weights are balanced for models that support this functionality (primarily tree-based methods).

The seven algorithms described in Section 2.3 are trained on each of the ten datasets under the three formulations (regression, multiclass classification, and binary classification), resulting in a total of 210 experiments. The models are implemented in Python using the scikit-learn library [45] for DT, RF, AdaBoost, Gradient boosting, ET, and SVM. The XGBoost algorithm is obtained from the Python library with the same name by Chen and Guestrin [37]. The experimentation dataset is split into training, validation, and testing subsets, with a proportion of $80-10-10$, respectively. Model selection and hyperparameter tuning are performed using the validation set, and the final evaluation of the best hyperparameter configuration is performed on the test set.

The experimental setup outlined above defines the full set of scenarios evaluated in this study. The next section presents the results obtained under each problem formulation and preprocessing configuration.

4. Results & discussion

4.1. Model training

This section presents the results of the model training process over the training set. Given the large number of experiments conducted, the outcomes are summarized into three tables, each corresponding to one of the prediction tasks: regression (Table 2), multiclass classification (Table 3), and binary classification (Table 4). For clarity, only the top five configurations for each task are shown, along with the best-performing model trained on the original dataset without any preprocessing. Models are ranked according to their primary evaluation metric: R^2 for regression and recall for classification tasks.

Table 2 shows that RF and SVM achieve the highest R^2 values, and both benefit from oversampling and normalization. However, overall regression performance remains limited across all configurations. This is related to the statistical characteristics of segment-level crash data, which are sparse, highly skewed, and dominated by zero values, making precise count prediction inherently challenging.

In practical safety applications, agencies are often more interested in identifying segments with elevated crash risk rather than accurately predicting the exact number of crashes. Therefore, given both the data properties and the operational focus on risk screening, the remainder of the analysis focuses on the classification-based formulations.

Table 3 shows that RF consistently achieves the highest recall in the multiclass setting, with oversampling and its expanded variant providing slight but steady improvements. However, overall performance remains moderate, suggesting that discretizing crash counts into categories may obscure relevant variation across road segments.

In contrast, the binary classification results in Table 4 clearly outperform the other two problem formulations. Gradient boosting achieves the highest recall (0.831) when trained with undersampling or normalized-undersampling preprocessing, making it the most suitable model for identifying crash-prone segments.

4.2. Best model selection

As the best results are obtained by the binary classification models, only they are considered when selecting the best model. To avoid overfitting and ensure robust comparison, a second evaluation stage is conducted using the validation dataset (Table 5), which contains 115 negative instances (no crash) and 81 positive instances.

Confusion matrices are employed to support model selection, providing a clear visualization of correct and incorrect predictions for each class. In the current case, aside from maximizing true positives and true negatives, lowering the number of false negatives (incorrectly predicted "no crash") is the highest priority, given their greater human and economic consequences compared to false positives.

Based on these criteria, although the first configuration achieves very high recall, its extremely low precision indicates an excessive number of false positives, making it unsuitable for practical deployment. The second configuration (Gradient boosting with normalization and undersampling) is selected instead as the best-performing model, obtaining a Brier score of 0.15 and an ECE of 0.06. Moran's I has also been calculated for the model, which scores $I = 0.022$.

Table 2
Top training results for the regression task.

Model	Preprocessing	R^2	MAE	MSE
RF	Oversampling	0.400	0.430	0.642
SVM	Normalization & E. Oversampling	0.398	0.422	0.644
RF	Normalization & Oversampling	0.396	0.436	0.646
RF	Normalization	0.396	0.428	0.646
SVM	Normalization & Oversampling	0.381	0.437	0.662
RF	No preprocessing	0.346	0.634	1.219

Table 3
Top training results for the multiclass classification task.

Model	Preprocessing	Accuracy	Precision	Recall
RF	Oversampling	0.796	0.685	0.796
RF	E. Oversampling	0.791	0.673	0.791
RF	Normalization & E. Oversampling	0.791	0.675	0.791
RF	No preprocessing	0.791	0.710	0.791
RF	Normalization & Oversampling	0.786	0.683	0.786

Table 4
Top training results for the binary classification task.

Model	Preprocessing	Accuracy	Precision	Recall
GradientBoosting	Undersampling	0.801	0.659	0.831
GradientBoosting	Normalization & Undersampling	0.801	0.659	0.831
AdaBoost	Normalization & Undersampling	0.786	0.485	0.825
SVM	No preprocessing	0.837	0.569	0.825
RF	Undersampling	0.796	0.500	0.825

Table 5
Binary classification model's performance on the validation dataset.

Model	Preprocessing	Accuracy	Precision	Recall
GradientBoosting	Undersampling	0.679	0.508	0.969
GradientBoosting	Normalization & Undersampling	0.842	0.774	0.738
AdaBoost	Normalization & Undersampling	0.811	0.706	0.738
SVM	No preprocessing	0.663	0.496	0.969
RF	Undersampling	0.668	–	–

4.3. Hyperparameter selection

Hyperparameter optimization is a key step in improving the performance and generalization of ML models. Traditional grid or random search approaches can be computationally expensive and often explore the hyperparameter space inefficiently. Tree-structured Parzen Estimators (TPE), introduced by Bergstra et al. [46], offer a Bayesian optimization strategy that models promising regions of the search space more effectively. Building on this idea, Optuna [47] provides a flexible framework that leverages TPE to automatically identify high-performing hyperparameter configurations.

In this study, Optuna is used to optimize the hyperparameters of the Gradient boosting classifier. The search space includes the number of estimators, learning rate, maximum tree depth, minimum samples required for splitting and for leaf nodes, subsample fraction, maximum number of features considered at each split, and the loss function. The optimization is performed using stratified 5-fold cross-validation with 50 trials, optimizing for the F1-score, and results in the hyperparameters available in Table 6.

The hyperparameter optimization process resulted in a Gradient boosting configuration characterized by a moderate tree depth and a relatively large ensemble size. The selected model uses a learning rate of 0.006, 447 estimators, a maximum depth of 8, and a minimum of 11 samples per leaf. A subsample ratio of approximately 0.58 is chosen to introduce stochasticity and reduce overfitting, while the exponential loss function is selected, reflecting the model's increased sensitivity to misclassified instances.

A new Gradient boosting classifier with these hyperparameters is trained on the training dataset and subsequently evaluated on the test set, which contains 115 negative and 81 positive classes, to assess its generalization performance. The model achieves a precision score of 0.747, a recall score of 0.802, and an accuracy score of 0.806, for a combined F1 score of 0.774, showing higher prediction for both classes. These results indicate that the hypertuned model improves balanced predictive performance and maintains strong crash-detection capability.

Table 6
Selected hyperparameters for the Gradient boosting classifier.

Hyperparameter	Value
Learning rate	0.006
Maximum depth	8
Number of estimators	447
Minimum samples to split	43
Minimum samples per leaf	11
Subsample fraction	0.578
Loss function	Exponential
Maximum features	None

4.4. Model calibration and risk thresholding

Beyond discrimination performance, it is essential to assess whether predicted crash probabilities are well calibrated, i.e., whether they correspond to observed crash frequencies. A well-calibrated model ensures that a segment assigned a probability of 0.8 experiences crashes approximately 80% of the time in the long run, which is critical for risk-based prioritization and policy decisions.

Fig. 2 presents the reliability (calibration) curve of the final, hypertuned Gradient boosting model. The curve departs from the diagonal,

Table 7
Performance trade-offs for recall-targeted probability thresholds.

Target recall	Threshold τ	Precision (CI-95%)	Recall (CI-95%)	F1 (CI-95%)
0.90	0.276	0.589 [0.526, 0.651]	0.904 [0.852, 0.949]	0.714 [0.661, 0.763]
0.95	0.174	0.535 [0.477, 0.595]	0.952 [0.914, 0.985]	0.685 [0.633, 0.734]
0.99	0.133	0.512 [0.456, 0.571]	0.993 [0.978, 1.000]	0.676 [0.625, 0.726]

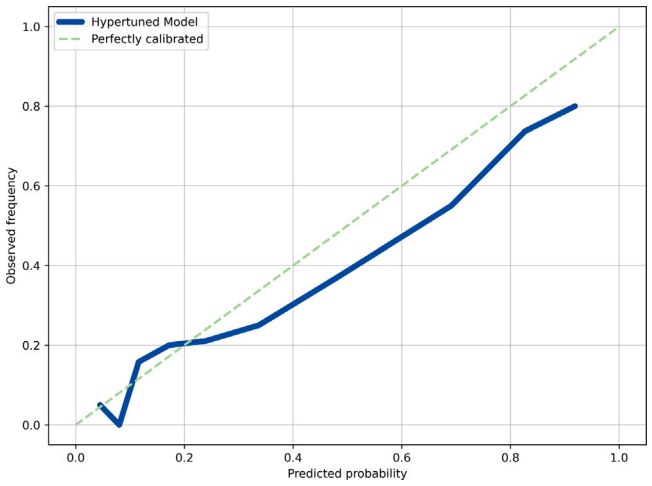

Fig. 2. Calibration curve of the Gradient boosting model with normalization and undersampling.

particularly at higher predicted probabilities, indicating systematic miscalibration. Specifically, the model tends to overestimate crash risk for high-risk segments: predicted probabilities in the upper bins correspond to substantially lower empirical crash rates. This pattern is consistent with the scarcity of positive crash examples and the strong influence of highly informative but spatially sparse connected-vehicle signals, which can inflate predicted probabilities on segments where such data are present.

This behavior is summarized by two complementary calibration metrics. The Brier score of 0.142 reflects moderate overall probabilistic accuracy, capturing both discrimination and calibration errors. Meanwhile, the ECE of 0.093 indicates that, on average, predicted probabilities deviate from observed frequencies by about 9 percentage points across probability bins. Together, these values confirm that although the model ranks risky segments well, its predicted probabilities should be interpreted primarily as relative risk indicators rather than as perfectly calibrated absolute estimates.

Given this miscalibration, threshold selection becomes a relevant operational decision. Rather than using a fixed probability threshold, we adopt a recall-targeted thresholding strategy, which is more appropriate for safety screening applications where missing high-risk locations is more costly than over-flagging low-risk ones. Table 7 reports the performance of three operating points chosen to achieve target recall levels of 0.90, 0.95, and 0.99. As the recall target increases, the threshold is progressively lowered, capturing more true crash-prone segments at the cost of reduced precision and higher false-positive rates. This trade-off is expected in imbalanced safety problems and reflects the practical reality that more conservative screening policies require accepting more false alarms.

From a policy perspective, this framework allows decision-makers to select thresholds based on their tolerance for false positives versus missed high-risk segments. For example, a high-recall operating point may be suitable for network-wide safety audits or proactive intervention planning, while a higher-precision threshold may be preferable when resources for detailed inspections or engineering treatments are limited. Importantly, this approach decouples risk ranking from probability calibration, enabling robust prioritization even when predicted probabilities are imperfectly calibrated.

4.5. Spatial autocorrelation of errors

A key concern in spatial risk modeling is whether prediction errors are spatially clustered, which could indicate unmodeled spatial processes or inflated performance due to spatial dependence. If nearby road segments tend to share similar residuals, standard validation procedures may overestimate generalization performance, and the model may systematically fail in specific areas.

To assess this, we computed Moran's I on the binary prediction errors (false positives and false negatives) using a spatial weights matrix defined over the road network. The resulting value, $I = -0.022$, is close to zero, indicating no meaningful global spatial autocorrelation. This suggests that the model's errors are not spatially clustered but instead are approximately randomly distributed across the network. In practical terms, the model is not systematically over- or under-predicting crashes in specific neighborhoods or corridors.

Fig. 3 provides a spatial visualization of the classification errors, where false positives are shown in green, false negatives in blue, and correctly classified segments in gray. The map reveals that both types of errors are scattered throughout the study area rather than forming large contiguous clusters. This visual pattern is consistent with Moran's I result and supports the conclusion that the model does not suffer from strong spatial bias in its misclassifications.

4.6. Feature importance impact analysis

The final step is to conduct a feature importance analysis to identify the most influential factors driving the model's predictions. This analysis is carried out using the SHAP [44] Python library. SHAP's TreeExplainer [48] is employed to compute Shapley-based feature attributions for the hypertuned Gradient boosting model, with the resulting importance rankings shown in Fig. 4.

Braking-related events emerge as the dominant predictors of crash likelihood. Higher counts of braking and harsh-braking events consistently increase the predicted risk, highlighting the strong link between abrupt deceleration behavior and crash-prone conditions. Other CVD indicators also contribute to the model's decisions, particularly the frequency of left-steering events and the average magnitude of right-steering events, although their influence is comparatively smaller.

Regarding infrastructure features, road length is the only variable with a clear impact on the model. Other geometric and regulatory attributes show weaker effects: higher posted speed limits and multi-directional street layouts tend to increase crash risk, whereas roads with a greater number of lanes appear less likely to experience crashes. Finally, traffic-demand variables show that segments with lower AADT levels are less prone to crash occurrence, consistent with reduced exposure on low-volume roads.

To assess the contribution of different feature groups, additional Gradient boosting models are trained using modified versions of the input dataset. Three variants are evaluated: one using only the top seven most influential features, one excluding those top features, and one removing all CVD data. The corresponding classification metrics on the test set, along with the 95% confidence intervals calculated by bootstrapping, are reported in Table 8.

First, the impact of the less important features is examined by keeping only the most influential variables identified through the SHAP analysis and removing the rest. The features removed from the training dataset included the following: the number and average magnitude of braking events, the number of left cornering events, AADT of private

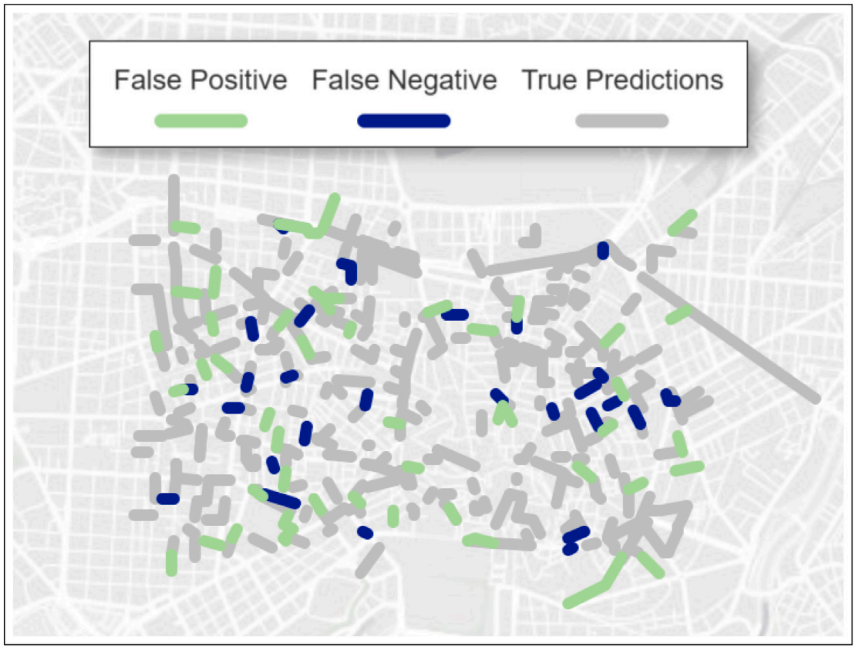

Fig. 3. Map showing the distribution of false positives (green), false negatives (blue), and correctly classified segments (gray).

Table 8
Performance comparison of the hypertuned Gradient boosting model with different input features.

Case	Accuracy (CI-95%)	Precision (CI-95%)	Recall (CI-95%)
Hypertuned model	0.806 [0.768, 0.844]	0.708 [0.643, 0.774]	0.816 [0.752, 0.876]
Top 7 features only	0.796 [0.755, 0.834]	0.689 [0.622, 0.754]	0.822 [0.762, 0.882]
Without top 7 features	0.778 [0.737, 0.819]	0.674 [0.606, 0.746]	0.782 [0.713, 0.847]
Without CVD features	0.698 [0.651, 0.745]	0.579 [0.503, 0.654]	0.698 [0.623, 0.770]

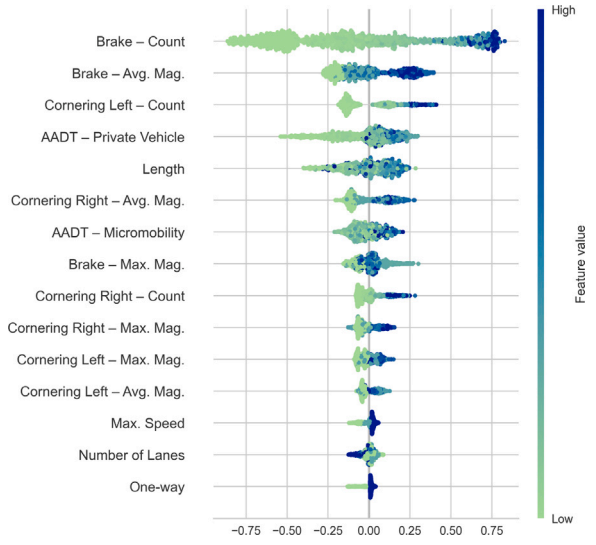

Fig. 4. Feature importance based on Shapley values.

vehicles, road length, the average magnitude of right cornering events, and AADT of micromobility users. The number of features to remove is based on the SHAP analysis.

The new trimmed model obtained accuracy, precision, and recall values of 0.796, 0.689, and 0.822, respectively, which is only a slight decrease when compared to the model trained over all the features. This model exhibits a strongly overlapping confidence interval across all metrics with the base hypertuned model, indicating that most of the predictive power is concentrated in a relatively small subset of variables, whereas the remaining features contribute only marginally to overall performance.

Next, the relevance of the seven most important features is tested by excluding them entirely from the training dataset. The resulting model obtained accuracy, precision, and recall values of 0.778, 0.674, and 0.782, respectively. The corresponding SHAP feature analysis (Fig. 5(a)) shows that, once the top predictors are removed, the maximum magnitude of braking events becomes the most influential factor. Other CVD features also gain relative importance, while infrastructure characteristics, such as the absence of traffic lights or lower speed limits, tend to indicate lower crash risk. Asphalted roads are more prone to crashes, although this can mean that those roads are the ones that see higher traffic demand.

To interpret these ablation results correctly, it is important to note that SHAP attributions can be affected by correlations and interactions among input variables. This is particularly relevant in the experiment where the seven most influential features are removed, because several of these variables are highly correlated with other connected-vehicle signals. In such cases, the model can partially compensate for their removal by relying on closely related CVD features, which may obscure the true impact of eliminating individual predictors.

For this reason, we also performed a more controlled ablation in which all CVD features were removed simultaneously. Since CVD constitutes the largest cluster of correlated variables and dominates the SHAP rankings, removing it as a group provides a clearer assessment of how model behavior changes when this entire information source is unavailable.

Without CVD, the model's performance dropped more substantially, yielding accuracy, precision, and recall values of 0.698, 0.579, and 0.698, respectively, which is a more significant decrease than in previous experiments. The confidence intervals are also clearly shifted

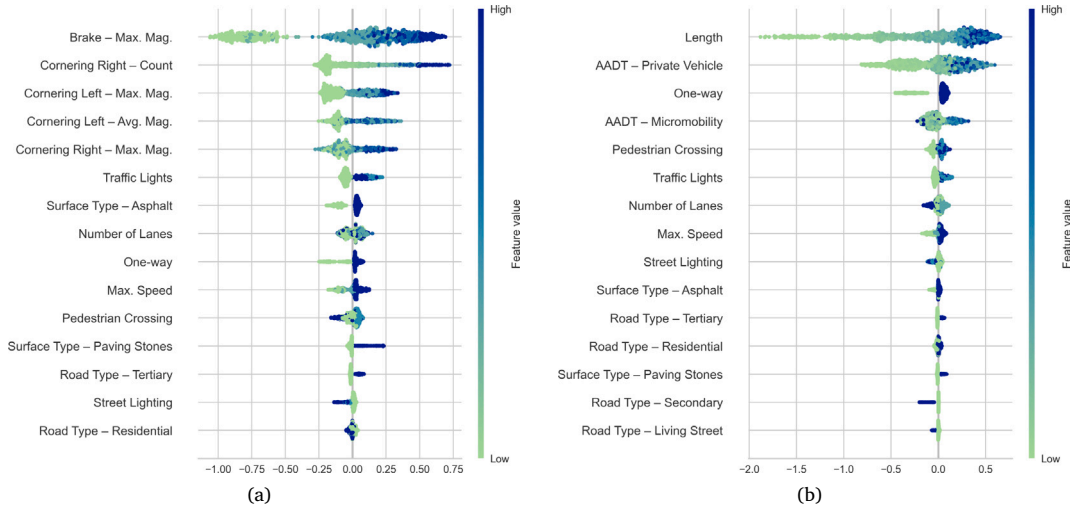

Fig. 5. SHAP feature importance after removing (a) the 7 most relevant features or (b) CVD.

downward across all metrics, indicating that the contribution of CVD data is not only large in magnitude but also statistically robust. The associated SHAP analysis (Fig. 5(b)) indicates that, when CVD data are removed, segment length and private-vehicle AADT become the most influential predictors.

Overall, the seven most relevant features allow the model to correctly identify a larger proportion of crash-prone segments, with only a modest decrease in overall predictive power. In contrast, removing all CVD produces a substantial deterioration in performance, highlighting the critical role that behavioral driving events play in crash prediction.

4.7. Degree of curvature

Silva et al. [20] reported that roadway curvature can strongly influence risk screening. However, the datasets available for this study do not provide a direct measure of curvature. To address this limitation, we implemented a geometry-based curvature measure directly from the OSM polylines.

For each road segment, curvature was computed between consecutive coordinate pairs along the geometry and then aggregated over the entire edge. This approach, however, revealed a methodological issue: OSM edges often represent long road sections that contain both straight and curved sub-segments. As a result, aggregating curvature at the segment level leads to misleading values; for example, a mostly straight segment containing a short bend yields a non-zero curvature, while a strongly curved sub-segment embedded in a longer straight segment is heavily diluted. The inverse situation also occurs, producing unstable and noisy curvature estimates.

When this geometry-based curvature variable was incorporated into the models, predictive performance decreased, and confidence intervals widened (ranging between 0.5 and 0.9 for the 95% intervals), indicating reduced stability. This result highlights the sensitivity of curvature measures to the spatial resolution at which geometry is represented. At the resolution of OSM segments, curvature may be affected by aggregation artifacts that limit its interpretability.

Roadway curvature is likely better suited to a sub-segment-level representation in which each unit corresponds to a homogeneous geometric element. In contrast, when working at the segment level, curvature can introduce geometric noise and reduce model reliability. For this reason, and to preserve methodological consistency, it was excluded from the final dataset.

5. Feature importance discussion

This section discusses the results of the SHAP feature analysis presented in Section 4.6. The importance of CVD in retrospective crash screening is evident across all experiments. The hypertuned model relies mainly on braking-related events, both how often they occur and how strong they are, to estimate crash likelihood. When these features are removed, the maximum braking magnitude becomes more influential, probably acting as a proxy for the missing event counts.

CVD exhibits strong coverage bias in the study area, and, as a result, CVD variables are extremely informative where they exist (explaining their dominance in SHAP rankings) but are absent across much of the network. This creates a spatial and socio-economic skew, since connected vehicles are more prevalent in newer vehicles, commercial fleets, and high-traffic corridors, and less common on residential streets or in lower-income areas [49,50]. Consequently, the model may systematically under-represent crash risk on segments without CVD coverage, raising concerns about equity, generalizability, and fair deployment. The performance degradation observed in the "without CVD" ablation indicates that a substantial portion of predictive skill currently derives from this uneven sensing rather than from universally available infrastructure features.

Low maximum braking acceleration is associated with a reduced probability of crashes. This is consistent with the idea that lower speeds provide drivers with more time to react, thereby reducing risk. In contrast, posted speed limits contribute relatively little to model predictions, particularly when CVD is unavailable, suggesting that speed-limit information may be more relevant at a micro-scale rather than at an aggregated road-segment scale.

Cornering events also play a notable role in model output. With the full feature set, frequent left cornering events and the average magnitude of right cornering events are among the most influential inputs. When the most important features are removed, the model shifts to relying on the count of right cornering events and the maximum or average magnitudes of remaining CVD signals. Although the maximum magnitude of left cornering events has limited influence in the hypertuned model, it becomes more relevant in reduced-feature scenarios, suggesting compensatory behavior in the model's decision process. Harsh acceleration events, on the other hand, appear rarely in the dataset and therefore show minimal contribution in all SHAP analyses.

After removing the seven most relevant features, CVD variables still dominate the model's decisions, but several infrastructure-related

attributes gain relative importance. Road length remains the most informative infrastructure feature, while maximum speed becomes moderately influential. When CVD features are further reduced or entirely removed, other infrastructure variables begin to play a more prominent role, particularly the presence of traffic lights, which is associated with higher crash probability, and the number of lanes, which the model uses as an additional indicator of exposure and risk.

Regarding traffic demand, high private vehicle AADT is consistently associated with increased crash likelihood, more so than micromobility AADT. The influence of these variables remains relatively stable across the different model configurations.

Overall, CVD features represent the strongest predictors of crash occurrence, reinforcing findings from prior studies [18,51] that emphasize the value of behavioral driving signals in safety analysis. In the absence of CVD, features such as road length and AADT still provide useful information, but none of the reduced-feature models achieve the predictive performance of the hypertuned model. These results highlight the benefits of integrating multiple heterogeneous data sources in crash prediction and underscore the central role of CVD in capturing real-time driving behavior associated with crash risk.

Additionally, CVD variables do not merely act as proxies for traffic exposure, which is already partially captured by AADT measures. Instead, they provide fine-grained behavioral signals that reflect drivers' immediate responses to local road conditions, such as abrupt braking or steering corrections. These micro-level driving dynamics likely capture short-term conflict patterns that are not observable through static infrastructure attributes or aggregated demand indicators alone.

6. Conclusions & future work

This article aimed to study the influence of multiple traffic factors on road crash prediction in urban environments. To this end, an integrated framework is proposed that combines infrastructure characteristics, historical crash records, travel demand information, and CVD to model crash occurrence at the road-segment level.

After data aggregation and cleaning, several common preprocessing techniques are applied to address class imbalance and support model learning, resulting in ten different dataset configurations. A set of widely used supervised ML models for crash prediction is then evaluated under regression, multiclass classification, and binary classification formulations. The experimental results showed that binary classification provides the most reliable performance, with a Gradient boosting model trained on normalized and undersampled data achieving the best overall results. Bayesian hyperparameter optimization further improved the predictive performance of this model, which was subsequently evaluated on an independent test set.

Once the hyperparameters are selected, a feature importance analysis is conducted using SHAP. The results indicate that CVD, and more specifically braking-related events, are the most influential factors in crash prediction. Other variables, such as AADT and road length, also contribute to the model's decision-making process, although to a lesser extent. In contrast, other CVD events and most road infrastructure attributes show a comparatively lower impact. To further examine the role of individual features, additional analyses are performed.

Specifically, three additional Gradient boosting models are trained: one using only the seven most relevant input features identified by SHAP, another excluding these features, and a third model trained without any CVD. All modified models experienced a decrease in overall performance. However, the model trained with only the seven most relevant features achieved higher recall, indicating an improved ability to identify crash-prone segments. When CVD features are removed entirely, SHAP analysis highlighted road length and AADT as the most important predictors, underscoring their role when behavioral information is unavailable.

Overall, these findings are consistent with previous studies in the state-of-the-art and confirm that road crashes are complex phenomena driven by the interaction of multiple factors.

The results highlight the importance of integrating heterogeneous data sources and demonstrate the added value of connected vehicle information for capturing driving behavior directly associated with crash risk. No single variable is sufficient to accurately predict crashes in isolation, underscoring the need for comprehensive, data-rich modeling approaches in road safety analysis.

Several sources of bias affect the outcome and input data used in this study and should be acknowledged as limitations. Police-reported crash records are known to suffer from underreporting, particularly for minor incidents and crashes involving VRU, which can distort the true spatial distribution of risk. CVD is also unevenly distributed across the city, with higher availability on major corridors and in areas with newer or fleet-based vehicles, introducing potential socioeconomic and spatial bias in behavioral measurements. Similarly, travel demand estimates depend on the presence and quality of sensing and aggregation infrastructure, which is not uniformly deployed across the network. Together, these factors imply that observed crash risk and model performance may be partly influenced by data availability rather than by underlying safety conditions alone, limiting generalizability and equity in real-world deployment.

Future work should explore alternative modeling techniques, such as deep learning architectures, as well as complementary statistical approaches based on count models (e.g., Poisson or Negative Binomial regression) to better capture crash frequency and overdispersion. Further research should also investigate how influential factors change when focusing on specific road-user groups or different crash contexts, such as intersections. Expanding the temporal and spatial coverage of connected-vehicle and travel-demand data is expected to improve model robustness and equity. Finally, the impact of road curvature, in conjunction with the other data sources, could be examined in more detail, as several studies identify it as a key determinant of crash risk.

As a final note, the authors want to mention that the whole process was also followed to model crashes for VRU. However, this line of research was eventually discarded due to poor performance results, primarily driven by data limitations, as the CVD used in this study is generated exclusively by passenger cars. Consequently, CVD features capture the driving behavior of motorized vehicles but provide little direct information about pedestrian or cyclist movements, interactions, or exposure. Because CVD constitutes the most informative feature group in the full model, its limited relevance to non-vehicular users substantially weakens the model's ability to explain VRU crash patterns.

Additionally, VRU crashes are comparatively rare at the edge level, and their distribution is highly imbalanced. Within the VRU category, a large proportion of recorded incidents involve motorcycles, while pedestrian and bicycle crashes are both less frequent and more spatially heterogeneous. This sparsity further reduces the statistical power available for learning stable associations, especially when disaggregating by VRU type.

CRediT authorship contribution statement

Jon Díaz-Aparicio: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Erick Rodríguez-Esparza:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Jenny Fajardo-Calderín:** Writing – review & editing, Validation, Supervision, Software, Investigation, Data curation. **Enrique Onieva:** Writing – review & editing, Visualization, Resources, Project administration, Funding acquisition, Conceptualization.

Ethics & permissions statement

The proprietary data employed in this research was provided by Nommon and Vianova to the consortium of the SOTERIA project.[3] In

[3] https://soteriaproject.eu

the SOTERIA project, CVD and TDD are processed in a GDPR-compliant manner. Only anonymized (and in parts additionally aggregated) data from vehicles and anonymized mobile phone users are used. All results are presented solely in aggregated form, ensuring that no individual person can be identified or traced by time, location, or any combination of presented attributes.

The data treatment and procedure of the project were approved by the ethics institutes of the consortium members.

The authors have access to this data as members of the consortium and have complied with the terms of use of the data set by the providers.

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.

Acknowledgments

This research has been partially funded by the Horizon Europe Research & Innovation Programme, Spain under Grant Agreement No 101077433 [project SOTERIA (Systematic and orchestrated deployment of safety solutions in complex urban environments for ageing and vulnerable societies)]. This work has also been partially funded by the Spain, Ministry of Science, Innovation and Universities through the RENAISSANCE project [PID2022-140612OB-I00].

Data availability

All **road network, geometry, and infrastructure attributes** (e.g., number of lanes, speed limits, road class, crossings, traffic signals, or surface type) were derived from OpenStreetMap (https://www.openstreetmap.org), a collaborative open dataset released under the Open Database License (ODbL). The dataset was last accessed in November 2025.

Crash data were obtained from Madrid's Open Data Portal (https://datos.madrid.es/portal/site/egob), an open dataset provided by the City Council of Madrid, covering the period 2019–2022 and released under the ODbL. The dataset was last accessed in November 2025.

Connected vehicle data were provided by Vianova (https://www.vianova.io) for the year 2022 and were made available to the authors as part of the SOTERIA project (https://soteriaproject.eu).

Travel demand data were provided by Nommon (https://www.nommon.es) for the year 2022 and were also made available to the authors as part of the SOTERIA project (https://soteriaproject.eu).

Preprocessing summary: All data within 20 meters of the centerline of each road segment were geospatially aggregated to the OpenStreetMap road geometries. Annual Average Daily Traffic (AADT) was computed by summing all available daily travel-demand observations for each road segment and dividing by the number of days with available data.

Reproducibility: The code used in this study is publicly available at https://github.com/Jondiii/crash_explainability.git. Due to data-sharing restrictions, proprietary datasets (connected-vehicle and travel-demand data) were replaced with randomized surrogates in the public repository. As a result, running the pipeline with the public data will reproduce the methodology but not the exact numerical results reported in this paper.

References

[1] WHO. Road traffic injuries report. 2023.
[2] Behboudi N, Moosavi S, Ramnath R. Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques. 2024, arXiv preprint arXiv:2406.13968.
[3] Ali Y, Hussain F, Haque MM. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. Accid Anal Prev 2024;194:107378.
[4] Chang L-Y, Wang H-W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accid Anal Prev 2006;38(5):1019–27.
[5] Iranmanesh M, Seyedabrishami S, Moridpour S. Identifying high crash risk segments in rural roads using ensemble decision tree-based models. Sci Rep 2022;12(1):20024.
[6] Santos K, Dias JP, Amado C. A literature review of machine learning algorithms for crash injury severity prediction. J Saf Res 2022;80:254–69.
[7] AlMamlook RE, Kwayu KM, Alkasisbeh MR, Frefer AA. Comparison of machine learning algorithms for predicting traffic accident severity. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology. JEEIT, IEEE; 2019, p. 272–6.
[8] Goswamy A, Abdel-Aty M, Islam Z. Factors affecting injury severity at pedestrian crossing locations with Rectangular RAPID Flashing Beacons (RRFB) using XGBoost and random parameters discrete outcome models. Accid Anal Prev 2023;181:106937.
[9] Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.
[10] Yu R, Abdel-Aty M. Utilizing support vector machine in real-time crash risk evaluation. Accid Anal Prev 2013;51:252–9.
[11] Hu J, Huang M-C, Yu X. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. Accid Anal Prev 2020;144:105665.
[12] Jiang F, Yuen KKR, Lee EWM. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid Anal Prev 2020;141:105520.
[13] Elgeldawi E, Sayed A, Galal AR, Zaki AM. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In: Informatics, vol. 8, MDPI; 2021, p. 79.
[14] Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 2020;415:295–316.
[15] a Ilemobayo J, Durodola O, Alade O, J. Awotunde O, T Olanrewaju A, Falana O, Ogungbire A, Osinuga A, Ogunbiyi D, Ifeanyi A, et al. Hyperparameter tuning in machine learning: A comprehensive review. J Eng Res Rep 2024;26(6):388–95.
[16] Johnsson C, Laureshyn A, De Ceunynck T. In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. Transp Rev 2018;38(6):765–85.
[17] Olszewski P, Szagała P, Rabczenko D, Zielińska A. Investigating safety of vulnerable road users in selected eu countries. J Saf Res 2019;68:49–57.
[18] Zhang S, Abdel-Aty M. Real-time crash potential prediction on freeways using connected vehicle data. Anal Methods Accid Res 2022;36:100239.
[19] Abou Elassad ZE, Mousannif H, Al Moatassime H. A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. Transp Res Part C: Emerg Technol 2020;118:102708.
[20] Silva PB, Andrade M, Ferreira S. Machine learning applied to road safety modeling: A systematic literature review. J Traffic Transp Eng (English Edition) 2020;7(6):775–90.
[21] Singh D, Das P, Ghosh I. Bridging conventional and proactive approaches for road safety analytic modeling and future perspectives. Innov Infrastruct Solut. 2024;9(5):128.
[22] Shbeeb L. Road safety performance index: A tool for crash prediction. Cogent Eng 2022;9(1):2124637.
[23] Simmachan T, Boonkrong P. Effect of resampling techniques on machine learning models for classifying road accident severity in Thailand. J Curr Sci Technol 2025;15(2):99.
[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.
[25] Morris C, Yang JJ. Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling. Accid Anal Prev 2021;159:106240.
[26] Carvalho M, Pinho AJ, Brás S. Resampling approaches to handle class imbalance: a review from a data perspective. J Big Data 2025;12(1):71.
[27] Molnar C. Interpretable machine learning. Lulu. com; 2020.
[28] Wen X, Xie Y, Jiang L, Li Y, Ge T. On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. Accid Anal Prev 2022;168:106617.
[29] Das A, Abdel-Aty MA. A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. Saf Sci 2011;49(8–9):1156–63.
[30] Iranitalab A, Khattak A. Comparison of four statistical and machine learning methods for crash severity prediction. Accid Anal Prev 2017;108:27–36.

[31] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR; 2015, p. 448–56.
[32] Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. J Amer Statist Assoc 1963;58(302):415–34.
[33] Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1, IEEE; 1995, p. 278–82.
[34] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.
[35] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput System Sci 1997;55(1):119–39.
[36] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;1189–232.
[37] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: ACM; 2016, p. 785–94.
[38] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[39] Glenn WB, et al. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
[40] Naeini MP, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 29, (1). 2015.
[41] Moran PA. Notes on continuous stochastic phenomena. Biometrika 1950;37(1/2):17–23.
[42] Brochu E, Cora VM, De Freitas N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. 2010, arXiv preprint arXiv:1012.2599.
[43] Shapley LS, et al. A value for n-person games. 1953.
[44] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. 2017, Curran Associates, Inc.
[45] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[46] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst 2011;24.
[47] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. 2019.
[48] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):2522–5839.
[49] Bansal P, Kockelman KM, Singh A. Assessing public opinions of and interest in new vehicle technologies: An austin perspective. Transp Res Part C: Emerg Technol 2016;67:1–14.
[50] Spurlock CA, Sears J, Wong-Parodi G, Walker V, Jin L, Taylor M, Duvall A, Gopal A, Todd A. Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the San Francisco Bay Area. Transp Res Part D: Transp Environ 2019;71:283–301.
[51] Islam Z, Abdel-Aty M. Traffic conflict prediction using connected vehicle data. Anal Methods Accid Res 2023;39:100275.