

Article

Machine Learning for Liability Attribution in Pedestrians Involved in Traffic Crashes: Interpretability and Class Imbalance Solutions

Felisa C. Gragera-Peña ¹, Miguel A. Jaramillo-Morán ^{1,*}  and Alejandro Moreno-Sanfélix ^{1,2} 

¹ Escuela de Ingenierías Industriales, Universidad de Extremadura, Avenida de Elvas s/n, 06006 Badajoz, Spain; cgragera@unex.es (F.C.G.-P.); almorenos@alumnos.unex.es (A.M.-S.)

² Judicial Traffic Police of the Local Police of Badajoz, St. Gaspar Méndez, 2, 06011 Badajoz, Spain

* Correspondence: miguel@unex.es

Abstract

This paper proposes a Machine Learning (ML) framework designed to attribute liability between drivers and pedestrians in traffic crashes. This study applies classification algorithms and interpretability techniques to analyze judicial rulings related to pedestrian crashes in Badajoz, Spain, from 2015 to 2024. The primary objective is to identify recurring crash patterns and determine liability levels for the parties involved. Several classification algorithms were evaluated, including Support Vector Machines (SVM), Neural Network (NN), Decision Trees (DT), Boosted Trees (BT), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbors (K-NN), and Logistic Regression (LR). Among them, the quadratic-kernel SVM achieved the highest overall performance. To address the severe class imbalance of the data, stratified k-fold cross-validation and the Synthetic Minority Oversampling Technique (SMOTE) were applied to enhance the robustness and generalization capability of the model. A multiclass classification framework was implemented, and SHAP (SHapley Additive exPlanations) was integrated to improve interpretability by quantifying the contribution of each feature to the model's predictions. The analysis identified critical factors that play a significant role in determining liability outcomes: driver license status, crash location, lighting conditions, reaction time, and the presence of drugs or alcohol. This research aims to contribute to the legal domain. While most existing studies have focused on predicting injury severity, few have addressed liability attribution. This is a multifactorial task that requires a comprehensive analysis of judicial decisions. The results demonstrate that machine learning-driven liability attribution can support judicial decision-making and provide valuable insights for the development of proactive urban traffic safety strategies.

Keywords: class imbalance; explainable AI; pedestrian crashes; urban road safety

MSC: 68T05; 90B20; 91D99; 76A30



Academic Editors:
Momčilo Dobrodolac,
Dragan Lazarević and
Aleksandar Trifunović

Received: 27 May 2026

Revised: 26 June 2026

Accepted: 2 July 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Road traffic crashes are among the most critical global public health challenges. According to the World Health Organization (WHO), approximately 1.35 million people die in traffic crashes every year. Pedestrians account for a significant share of these fatalities, particularly in urban environments where vulnerability is greatest. Currently, road traffic injuries are the leading cause of death among individuals aged 5 to 29. Projections suggest that by 2030, traffic crashes will become the fifth leading cause of death worldwide [1]. In

the European Union, nearly 19,800 fatalities were recorded in 2024 (44 deaths per million inhabitants). Vulnerable road users, especially pedestrians, are the most affected group. It is estimated that around 270,000 pedestrians die in traffic crashes worldwide each year, accounting for about 23% of all road deaths [2]. In Spain, this trend persists: in 2024 there were 1785 road fatalities, including 206 pedestrians, who represented 42% of urban traffic victims. Although these figures show a slight decline compared to previous years, the magnitude of the problem underscores the urgent need for effective strategies to reduce mortality and injury severity [3].

It is noteworthy that Badajoz, the city where this study was conducted, received the Municipal Vision Zero Award in 2022 for achieving zero traffic fatalities among cities with over 100,000 inhabitants [4]. This achievement is consistent with the European Vision Zero strategy, which aims to eliminate all traffic deaths and serious injuries by 2050. The intermediate target is to reduce fatalities and serious injuries by 50% by 2030 compared to 2019 [5].

Pedestrians consistently account for most fatalities and severe injuries in urban crashes. It is essential to understand the primary causes, severity and liability attribution in crashes involving pedestrians in order to design evidence-based policies, preventive measures and judicial frameworks [6]. Despite extensive research on crash severity, liability attribution remains underexplored, particularly in pedestrian crashes. Although there is no universally accepted legal definition of liability, it is often described as ‘a statement expressing a negative judgement (legal disapproval) on the conduct of a person who has violated a rule within a given legal system’ [7]. Liability may be assigned exclusively to one or more of the parties involved in the crash, either individually or jointly, and with varying degrees of legal blame [8].

In this context, Machine Learning (ML), particularly classification algorithms, has become a powerful tool for modeling risk patterns, predicting crash severity and supporting data-driven decision-making in road safety. The aim of this work is to model the attribution of liability in traffic crashes involving pedestrians using ML classifiers and interpretability tools. The goal is to bridge the gap between legal reasoning and data-driven insights. The insights from this study may help courts and traffic authorities evaluate cases and support informed, consistent decision-making. The following research questions arise and will be discussed and answered throughout this study: “Can machine learning models accurately classify liability levels in pedestrian-involved crashes?”; “Does SMOTE improve minority-class classification in liability attribution?”; “Which crash-related variables contribute most strongly to model predictions?”; and “Can SHAP-based interpretation provide legally meaningful explanations for liability classification?”.

This work is primarily empirical. Existing models have been proposed to address an unresolved problem using well-established data preprocessing and classification tools. Additionally, a tool has been added to explain the results and provide authorities with the information needed to understand and interpret them, which facilitates responsibility assignment and generates actions that could help prevent traffic crashes.

The best tool for one problem may not be the best tool for another. There is no single tool that is the best for all classification problems. This is why there are so many classification tools. Several of them must be tested for a particular problem to determine which one adapts best to its circumstances (the problem itself, the other tools used, and the information to be obtained). For this reason, a detailed and systematic analysis of the classification tools has not been provided. Instead, only several well-established tools have been tested to ensure that the model proposed can be used to address similar problems and to provide a clear understanding of their robustness. An overly meticulous selection of the classification tool could diminish its generalizability.

This study aims to lay a solid foundation for a new line of research about which little is known: providing support for making final decisions about determining responsibility for traffic crashes. To this end, we began with simpler models to ensure their effectiveness and then progressed to more complex models. Moreover, more complex models require intensive parameter tuning, which makes traceability difficult. This is why we initially prioritize basic models, adjusting the balance between performance and interpretability.

The main contributions of this work are the following: the development of a machine learning-based framework for attributing liability in crashes involving pedestrians; the integration of techniques to address the severe class imbalance; the application of SHapley Additive exPlanations (SHAP) for model interpretability and transparency; and empirical analysis using court rulings from Badajoz (2015–2024).

The rest of this paper is structured as follows: Section 2 reviews related works; Section 3 describes the dataset, preprocessing steps, and methodology; Section 4 presents the experimental results; Section 5 discusses interpretability and key influencing factors; and Section 6 presents the conclusions.

2. Literature Review

Traditionally, research on traffic crashes has focused on predicting injury severity, often neglecting the legal attribution of liability, especially in pedestrian crashes. Many studies have examined the multifactorial nature of pedestrian crash severity. These studies have identified key variables, such as vehicle speed, lighting conditions, pedestrian behavior, and pre-crash maneuvers [9,10].

In recent years, ML techniques have become more prominent in traffic safety research due to their ability to process large datasets and reveal complex, non-linear relationships [11]. Several ML algorithms, including Decision Trees (DT), k-Nearest Neighbors (kNN), Naive Bayes (NB), and AdaBoost, have been used to accurately predict injury outcomes with low error rates. The factors emphasized in legal case analyses, such as vehicle speed, visibility, pre-crash maneuvers and pedestrian actions, closely correspond to the predictions provided by ML models [12]. This convergence suggests a significant intersection between legal reasoning and data-driven insights.

Among the ML algorithms usually used to analyze traffic crashes, Support Vector Machines (SVM) stand out for their ability to manage high-dimensional feature spaces and define non-linear decision boundaries [13]. These properties make them particularly suitable for complex liability attribution tasks, as evidenced by the results obtained in this study. Therefore, their inclusion and detailed analysis in the present research are fully justified. Consequently, this research uses SVM due to its proven ability to solve complex classification problems related to traffic safety [14]. SVM outperforms other machine learning tools, as the results of this research will demonstrate.

Recent studies have also explored unsupervised ML techniques to reveal hidden patterns in pedestrian crashes under varying lighting and environmental conditions. They have provided valuable information for both legal and policy decision-making [15]. However, class imbalance remains a persistent challenge since severe injuries and cases of exclusive pedestrian fault are relatively rare. This can bias predictions and degrade performance in minority classes [16]. To mitigate this issue, researchers have implemented strategies such as oversampling (e.g., SMOTE), cost-sensitive learning and class weighting, combined with stratified k-fold cross-validation to preserve class distribution during training and evaluation [17]. While prior research has focused on injury severity prediction, few studies address liability attribution, which involves complex legal reasoning and multi-factorial analysis.

Recent contributions emphasize the importance of robust evaluation metrics, including macro- and weighted-F1 scores and Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values, to ensure fairness and reliability in predictive models [18]. Furthermore, integrating explainable ML models, such as SHAP, has made it easier to identify influential variables in liability attribution. These advances enhance predictive performance and provide empirical evidence to support legal decision-making in civil and criminal contexts. Comparative analyses of SHAP show how these tools improve model transparency and help identify important factors in predicting traffic crashes. This reinforces their usefulness in legal and forensic applications [19].

3. Data and Methodology

3.1. Dataset and Data Preparation

This study is based on final judicial rulings derived from official reports issued by the Judicial Traffic Police (JTP) of the Badajoz Local Police (BLP) and judicial decisions from Spanish Judiciary (SJ). Specifically, 510 rulings involving pedestrian crashes in the city of Badajoz between 2015 and 2024 were analyzed. The dataset is considered objective, as it consists of concluded cases with no possibility of further legal action. Its reliability stems from the fact that those cases selected from the JTP of BLP underwent a judicial process and the Courts of First Instance made decisions consistent with the level of responsibility established by the JTP. These rulings are final and not subject to appeal. Cases selected from the SJ correspond to courts hierarchically superior to the Courts of First Instance and are likewise not subject to further appeal.

The most relevant factors considered by the JTP of the BLP and the SJ to determine the degree of liability of pedestrians and drivers were identified, and crashes were classified into five liability categories based on the corresponding police reports and court rulings, as shown in Table 1. These categories were defined according to JTP and SJ rulings: Class A, driver fully liable (100%); Class B, liability is distributed as 75% driver and 25% pedestrian; Class C, both parties share equal liability (50–50%); Class D, the distribution is reversed (25% driver and 75% pedestrian); and Class E, pedestrian fully liable (100%).

Table 1. Classes and levels of responsibility used in this study from the original dataset.

Category	Liability Level (%)		Dataset
	Driver	Pedestrian	
A	100	0	61.71%
B	75	25	9.67%
C	50	50	7.75%
D	25	75	10.37%
E	0	100	10.51%

Fourteen factors consistently present across all cases were selected to determine liability levels. These factors were chosen because they systematically appeared in both judicial and police reports assessing liability in pedestrian crashes. Table 2 lists and briefly describes the most relevant factors used to characterize the crashes. They all were encoded as binary variables according to the four subsystems defined in the MOSES model [20,21].

The human subsystem includes six variables: H-1, driver attention, based on the match between the possible perception position (PPP) and the real perception position (RPP); H-2, normal reaction time (RT); H-3, driver alcohol consumption; H-4, driver drug use; H-5, pedestrian alcohol consumption; and H-6, pedestrian drug use. The technological subsystem comprises two variables: T-1, vehicle compliance with periodic technical inspection, and T-2, pedestrian clothing visibility. The structural subsystem includes two variables: S-1,

location of crash and S-2, lighting conditions. Finally, the normative subsystem consists of four variables: N-1, possession of a valid driving license; N-2, speed limits compliance; N-3, driving using mobile phone or similar; and N-4, pedestrian crossing while using a mobile phone or wearing music headphones or similar devices.

Table 2. Descriptive independent variable state.

Factor Name	Code	Description	Value	
Driver Attention	H-1	Driver attention while driving. PPP ^a and RPP ^a match/not match	1 (match)	0 (not match)
Reaction Time	H-2	The average RT ^b for a normal person is less than 0.75 s	1 (RT ^b ≤ 0.75 s)	0 (RT ^b > 0.75 s)
Driver Alcohol	H-3	Driver exceeds alcohol limit rate ^c	1 (No)	0 (Yes)
Driver Drugs	H-4	Driver Drugs	1 (No)	0 (Yes)
Pedestrian Alcohol	H-5	Pedestrian exceeds alcohol limit rate ^c	1 (No)	0 (Yes)
Pedestrian Drugs	H-6	Pedestrian drugs	1 (No)	0 (Yes)
Vehicle Inspection	T-1	Expired vehicle periodic technical inspection	1 (No)	0 (Yes)
Pedestrian Visibility	T-2	Visibility pedestrian clothing	1 (Yes)	0 (No)
Location	S-1	Influence area crash location ^d	1 (inside area) ^d	0 (outside area) ^d
Lighting Conditions	S-2	Lighting conditions	1 (Day/without glare)	0 (Night/glare)
Driver License	N-1	Expired or without driving license	1 (No)	0 (Yes)
Speed Limit	N-2	Exceeding the road limit speed	1 (No)	0 (Yes)
Driver Violation	N-3	Driving while using mobile phone or similar	1 (No)	0 (Yes)
Pedestrian Disattention	N-4	Crossing with mobile phone or with music headphones or similar	1 (No)	0 (Yes)

^a PPP: Possible Perception Point; PPR: Real Perception Point. ^b RT: Reaction Time. The average RT for a normal person is less than 0.75 s. ^c The legal alcohol limit in Spain is less 0.25 mg/L. ^d Pedestrian crossing or its influence area (approx. 5 m). Code: H: Human; T: Technological; S: Structural; N: Normative.

Taking all this information into account, each traffic crash report was transformed into a 14-component binary vector (1/0 values) and assigned to its corresponding liability class. Consequently, each ruling was represented as a binary 14-component vector with its associated class label, yielding a dataset of vector-class pairs suitable for ML processing.

Modelling, simulations, and result analysis were conducted using MATLAB software (The MathWorks, Inc., Natick, MA, USA), version R2025b, which provides robust tools for classification and data processing.

3.2. Data Preprocessing

A feature correlation analysis of the initial variables was performed to prevent multicollinearity, which could distort interpretations of their values and lead to erroneous conclusions in legal contexts. For this purpose, Pearson correlation coefficient, Cramér's *V* metric, and Variance Inflation Factor (*VIF*) were applied.

Pearson correlation coefficient can be represented as shown in Equation (1) [22]:

$$M = \begin{pmatrix} \rho(x, x) & \rho(x, y) \\ \rho(y, x) & \rho(y, y) \end{pmatrix}, \quad (1)$$

where $\rho(i,j)$ represents the quotient of their convolution and their corresponding standard deviations between variables i and j . This expression can be extended to the situation involving n variables.

Cramér's V is defined as shown in Equation (2) [23]:

$$V = \sqrt{\frac{X^2}{N(m-1)}}, \quad (2)$$

where X^2 is the chi-squared statistic value, N is the sample size, and m is the minimum of (rows-1) and (columns-1).

In these two techniques, the value of the correlation coefficient ranges from -1 to $+1$. A perfect direct relationship is indicated by a value of $+1$, while a perfect inverse relationship is indicated by a value of -1 . As the value approaches zero, there is less of a relationship (i.e., closer to no correlation). The correlation between the variables is related to how close the coefficient is to -1 or 1 [24].

VIF is defined as shown in Equation (3) [25]:

$$VIF_i = \frac{1}{1 - R_i^2}, \quad (3)$$

where R_i^2 is the coefficient of determination of the auxiliary regression of the variable X_i on the rest of the explanatory variables.

VIF values equal to or very close to 1 indicate that the variables are completely independent and not correlated. As this value increases beyond 1, so does the correlation. Therefore, for VIF values between 1 and 5, the correlation between the variables is moderate. The model is generally acceptable and does not require drastic adjustments. For VIF values between 5 and 10, the correlation is high and becomes severe for values greater than 10 [26,27].

3.3. Classification Procedure

This work uses a ML-based classification approach to assign responsibility levels to pedestrians and drivers in pedestrian crashes, as well as to assess the impact of various contributing factors on civil liability. The workflow is structured as shown in Figure 1. Several ML tools will be evaluated, and the most effective one will be selected for a comprehensive, in-depth analysis of its results.

The five-category distribution (Table 1) revealed significant class imbalance. According to the original judgments, the majority class (Class A, driver fully liable) accounted for 61.71% of all cases. The remaining cases were distributed among four minority classes—Class B (9.67%), Class C (7.75%), Class D (10.37%), and Class E (10.51%)—reflecting varying degrees of shared liability. Class imbalance is a well-known challenge in traffic crash datasets, particularly because cases involving exclusive driver fault are the most frequent. This is especially relevant given that pedestrians are vulnerable road users. Such imbalance poses significant challenges for traditional algorithms and requires the use of advanced supervised learning techniques combined with specialized preprocessing strategies to improve model performance [28]. Furthermore, the small number of samples available to construct the dataset exacerbates the problem. Dividing the dataset into training and testing subsets may lead to even more imbalanced class distributions within each subset. However, this partition is necessary for ML-based approaches, as models must be trained on one dataset to learn underlying patterns and subsequently evaluated on a separate dataset to assess their performance.

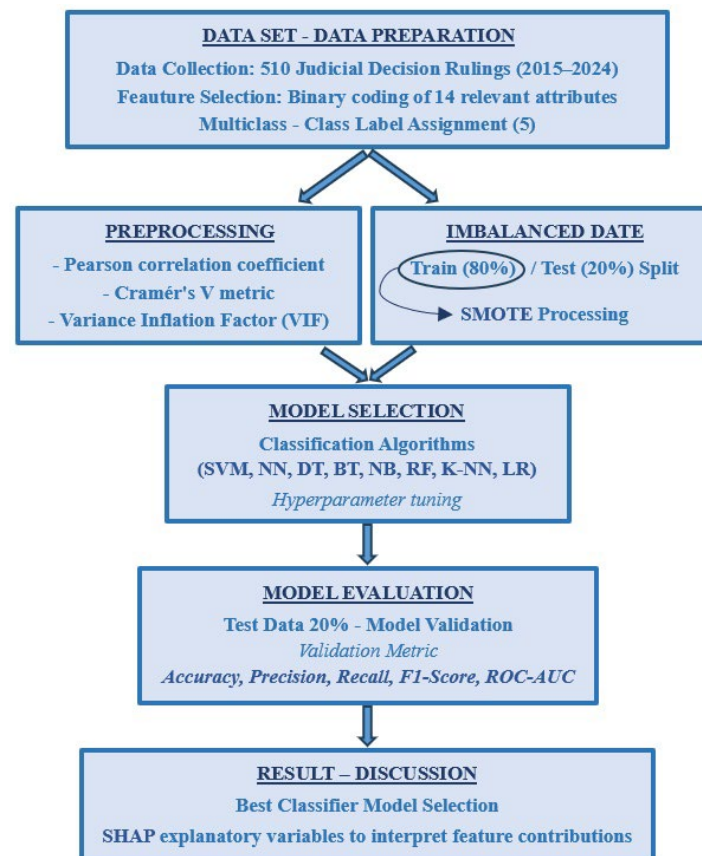


Figure 1. Summary of the proposed methodology flow.

In this work, the training set comprises 80% of the total data (408 pedestrian crashes), and the test set comprises the remaining 20% (102 crashes). Alternative data splits, such as 60/40, 70/30, and 75/25, were also evaluated, but they did not yield improved results.

Once the dataset is split into training and testing sets, preprocessing must be applied to the training set to address the issue of class imbalance. The Synthetic Minority Over-sampling Technique (SMOTE) was selected for this purpose. SMOTE generates synthetic samples for minority classes (Classes B, C, D and especially Class E) by interpolating existing instances, thereby creating a more balanced dataset by artificially increasing the number of elements of these classes, while preserving the original distribution patterns. SMOTE is renowned for its effectiveness in ML and data mining. Its application increases the size of the minority class, resulting in a more balanced training set. This strengthens the generalization capability of classification models for subsequent analyses [16]. As mentioned above, this procedure only applies to the training set.

It is important to note that the synthetic samples provided by SMOTE are not necessarily binary, as they result from interpolation. Analysis of these samples revealed that some contained non-binary components, though their values were very close to “0” or “1”. Therefore, these samples were processed to convert them into binary values. They were binarized to their nearest value (0 or 1). This resulted in two datasets for analysis. One dataset is Binary SMOTE, which contains only samples with binary components (non-binary components are rounded to 0 or 1). The other dataset was No-binary SMOTE, which included both binary and non-binary values. To ensure consistency, both datasets were analyzed, and it was verified that the synthetic non-binary values corresponded to legally significant conditions.

Both training sets were evaluated, and it was found that the one with non-binary values (No-binary SMOTE) produced slightly better results. Therefore, this set was selected for training the models presented in this work.

After addressing the issue of sample imbalance, the small sample size must also be considered. K-fold cross-validation is typically recommended for this issue because it is easy to program and improves accuracy and robustness. This algorithm divides the training set into k subsets of similar size. In each iteration, one subset is used for validation, while the remaining $k-1$ subsets are used for training. This process is repeated k times so that each subset serves as the validation set exactly once. The algorithm summarizes the performance of each simulation. This process is repeated several times with different sets of hyperparameters for the ML model. The optimal set of hyperparameters is selected based on the best performance achieved during cross-validation. These hyperparameters define the final model configuration, which is subsequently trained on the entire training set. The model's performance is then evaluated on the independent testing set. To mitigate the effect of imbalanced classes, a stratified k-fold cross-validation is applied, meaning the number of elements of each class is kept constant in the k subsets. Several values of k were tested, and the best results were obtained with $k = 10$. This value is usually recommended in most works [17]. Other cross-validation techniques were also tested, such as Nested Cross-Validation or Leave-One-Out Cross-Validation, but they yielded worse results than stratified cross-validation.

Several ML tools were evaluated in this work, including Support Vector Machines (SVM), Neural Network (NN), Decision Trees (DT), Boosted Trees (BT), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbors (K-NN), and Logistic Regression (LR). Although all models were tested, only the best-performing one is analyzed in detail to prioritize interpretability and methodological clarity. The SVM model achieved the highest performance and will be the focus of the following sections.

All models were evaluated using global overall accuracy, macro-recall, and macro-F1 score during testing. The best-performing model, SVM, was further analyzed using the confusion matrix and the accuracy, precision, recall, and F1 score metrics for each class. Additionally, ROC curves and the corresponding AUC values (by class and average) were generated. Finally, the SHAP framework was applied to interpret the model, analyze feature contributions, and identify the most relevant variables.

3.4. Support Vector Machine (SVM)

SVM is a supervised ML algorithm primarily designed for binary classification, although it can be extended to multiclass problems through strategies such as One-vs-One (OvO) or One-vs-Rest (OvR). The main objective of an SVM is to identify the optimal separating hyperplane that maximizes the margin between classes [29]. This hyperplane is defined by:

$$w^T x - b = 0, \quad (4)$$

where w is the normal vector to hyperplane, and b is the bias term (offset from the origin). The optimization goal is to minimize $\left[\frac{1}{2}w^T w\right]$, which is equivalent to maximizing the margin. To allow for non-linear classification and margin violations, a soft-margin formulation introduces slack variables ξ_i and a regularization parameter C . They control the trade-off between margin size and classification error. In this way the optimization problem becomes:

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \right), \quad (5)$$

subject to $y_i(w^T x_i - b) \geq 1 - \xi_i; \xi_i \geq 0$.

When the data are not linearly separable in the original feature space, SVM uses a kernel function to project data into a higher-dimensional space where linear separation becomes possible. Instead of explicitly computing the transformation, kernels efficiently compute inner products in the transformed space, enabling the algorithm to handle complex, non-linear decision boundaries [30].

To address the multiclass classification problem using binary classifiers, the OvO and OvR strategies were evaluated. The OvO approach provided superior performance. Details of the optimized SVM configuration are provided in Table 3.

Table 3. Definition of the model hyperparameters.

Quadratic Kernel SVM	
Hyperparameter	Value
Kernel function	Polynomial (quadratic) $K(x, x') = (x^T x' + c)^2$
Kernel Scale	Automatic
Box Constraint Level	1
Multiclass Coding	One-vs-One (OvO)
Standardize Data	Yes

3.5. Performance Metrics

A set of standard evaluation metrics—accuracy, recall, precision, and F1-score—was used to assess the predictive performance of the classification models. The mathematical formulations of these metrics are described by Equations (6)–(9), as summarized in Table 4. Although these metrics were originally developed for binary classification, they can be generalized to multiclass problems by computing them for each class using strategies such as OvO or OvR. Accuracy quantifies the proportion of correctly classified instances among all samples; however, it can be misleading in imbalanced datasets because it tends to be dominated by the majority class. Recall evaluates the model’s ability to correctly identify positive cases, which is critical when false negatives carry severe consequences. Precision reflects the proportion of true positives among all predicted positives, with higher precision implying fewer false positives. Finally, F1-score gives the harmonic mean of precision and recall, a balanced measure that is especially useful when the class distributions are skewed [31].

Table 4. Performance metrics for evaluating the ML classifiers.

Metrics	Expressions
Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$
Recall	$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$
Precision	$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$
F1 score	$\text{F1 score} = \frac{(2*TP)}{(2*TP+FN+FP)} \quad (9)$

TP: True Positive, TN: True Negative. FP: False Positive, FN: False Negative.

To ensure a robust evaluation under severe class imbalance and to facilitate model comparison, macro-averaged metrics (macro-recall, macro-precision, and macro-F1 score) were used to assess the performance of all ML models tested. Accuracy is straightforward to interpret and enables comparison across models. However, macro-recall mitigates bias toward dominant classes, gives equal weight to all classes, and ensures a balanced performance evaluation, which is essential when working with imbalanced datasets. The macro-F1 score provides a balanced measure by averaging the F1 scores across all classes, combining precision and recall into a single harmonic mean metric [32].

Additionally, ROC curves and the corresponding AUC values were used to evaluate performance per class. ROC curves plot the true positives rate against the false positives rate at various decision thresholds. Curves closer to the upper-left corner indicate superior performance, while curves near the diagonal suggest random behavior. Higher AUC values (approaching 1) reflect superior classification performance. All ROC curves and their corresponding AUCs for each class were plotted together to enable comparative analysis. In this regard, average curves were also used to assess the performance of the best performing multiclass classifier (the SVM model in this study). Three types of AUC are shown by these curves: macro-AUC, which represents the classifier's performance when treating all classes equally and preventing the majority class from dominating the evaluation; micro-AUC, which shows the classifier's overall behavior regardless of imbalance; and weighted-AUC, which accounts for the actual class distribution, assigning greater influence to more frequent classes.

3.6. Model Interpretation

After classifying the dataset, interpreting the results is essential to extract meaningful insights and ensure transparency in predictive analytics. Interpretability is critical for trust and accountability in ML-based decision-making. Among the most widely adopted techniques, SHAP stands out as a robust framework grounded in cooperative game theory [33]. The adoption of SHAP over other alternative interpretability techniques, such as Local Interpretable Model-agnostic Explanations (LIME), is justified by its strong theoretical foundation and consistency. SHAP assigns an importance value to each feature for a specific prediction, enabling both local interpretability (individual predictions) and global interpretability (overall model behavior). This dual capability makes SHAP particularly valuable for complex models [34].

The SHAP framework decomposes a model's prediction into a baseline value plus the sum of feature contributions:

$$f(x) = g(f(x)) + \sum_{i=1}^M \Phi_i z_i', \quad (10)$$

where $f(x)$ denotes the prediction for sample x , $g(f(x))$ represents the model's average predicted value (baseline), M is the number of input features, Φ_i is the SHAP value for the i -th feature—it measures its marginal contribution—and $z_i' \in \{0,1\}$ specifies whether the feature is present (1) or absent (0) [35].

SHAP values are computed by evaluating the marginal contribution of each feature across all possible subsets of features. The classifier is repeatedly trained on all possible subsets of features S obtained from the full set F , withholding one feature at a time. The contribution of each feature is calculated by comparing predictions with and without the feature, normalized by the cardinality of sets F and S . This behavior is described by the expression:

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]. \quad (11)$$

Positive SHAP values indicate that increasing the feature value raises the predicted outcome, while negative values imply the opposite. Graphical representations, such as bar plots showing the mean absolute SHAP value for each feature and a summary plot displaying the SHAP distribution across samples, improve interpretability. Both visualization techniques are used in this work to identify the most influential variables in liability attribution for pedestrian crashes, thereby providing evidence-based insights into explana-

tory variables (the variables selected to characterize the pedestrian crash). By quantifying feature contributions, SHAP improves interpretability without compromising predictive performance, supporting transparent and reliable decision-making in transportation safety research. Recent studies highlight the growing relevance of SHAP in engineering and transportation contexts. These studies demonstrate SHAP's role in improving transparency and compliance in safety-critical applications [36,37].

The parameters used in SHAP implementations are shown in Table 5. These parameters were selected based on their ability to achieve the best predictive performance.

Table 5. SHAP parameters.

Description	Parameters
Query dataset	Validation
Number of query points	200
Function	Kernel SHAP algorithm
Observation samples	100
Maximum predictor subsets	1024

4. Results

4.1. Preprocessing Results, Pearson Correlation Coefficient, Cramér's V Metric, and VIF Value

Preprocessing was performed using a Pearson matrix of the dataset and a Cramér *V* metric. The result obtained with Pearson's matrix (Figure 2) is corroborated by Cramér's *V* metric. Both methods identified only two pairs of variables that could be considered correlated (values greater than 0.5 or less than -0.5): H-3/H-4 and H-5/H-6. However, this correlation is not particularly strong, as both methods yield values only slightly above 0.5 or below -0.5 . For the Pearson matrix, the values for the H-3/H-4 and H-5/H-6 variables are 0.5374 and 0.5993, respectively, and for the Cramér *V* metric, they are 0.566 and 0.638, respectively. All other variable pairs exhibit correlation values below 0.5 and above -0.5 . This indicates low levels of association and suggests that multicollinearity is not a significant concern within the dataset.

VARIABLES	H-1	H-2	H-3	H-4	H-5	H-6	T-1	T-2	S-1	S-2	N-1	N-2	N-3	N-4
H-1	1.0000	0.0000	-0.0068	0.0248	-0.0153	-0.0450	0.0159	-0.0294	-0.0288	-0.0103	0.0022	-0.0068	0.0046	0.0363
H-2	0.0000	1.0000	0.0000	0.0314	-0.0394	-0.0039	-0.0122	0.2824	0.2575	-0.0993	0.2431	0.0000	0.0039	-0.2471
H-3	-0.0068	0.0000	1.0000	0.5374	0.3199	0.3397	-0.0964	-0.0117	0.2061	0.0391	0.0039	-0.0039	-0.0078	-0.0001
H-4	0.0248	0.0314	0.5374	1.0000	0.2673	0.3267	-0.0982	0.0045	-0.1587	0.0512	0.0274	0.0038	0.0159	-0.0004
H-5	-0.0153	-0.0394	0.3199	0.2673	1.0000	0.5993	-0.1020	0.0255	0.0039	0.0249	0.2094	-0.0351	0.0071	-0.0378
H-6	-0.0450	-0.0039	0.3397	0.3267	0.5993	1.0000	-0.1317	0.0452	0.1285	0.0136	0.0951	-0.0390	0.0268	-0.0653
T-1	0.0159	-0.0122	-0.0964	-0.0982	-0.1020	-0.1317	1.0000	-0.0702	-0.0448	0.0631	0.0335	-0.2507	-0.0305	0.0164
T-2	-0.0294	0.2824	-0.0117	0.0045	0.0255	0.0452	-0.0702	1.0000	0.0720	-0.0048	0.1923	0.0197	-0.0002	-0.4077
S-1	-0.0288	0.2575	0.2061	-0.1587	0.0039	0.1285	-0.0448	0.0720	1.0000	-0.1347	-0.3250	0.0130	0.1591	0.0197
S-2	-0.0103	-0.0993	0.0391	0.0512	0.0249	0.0136	0.0631	-0.0048	-0.1347	1.0000	0.0947	0.0232	0.3547	0.1127
N-1	0.0022	0.2431	0.0039	0.0274	0.2094	0.0951	0.0335	0.1923	-0.3250	0.0947	1.0000	0.0039	-0.2666	-0.2589
N-2	-0.0068	0.0000	-0.0039	0.0038	-0.0351	-0.0390	-0.2507	0.0197	0.0130	0.0232	0.0039	1.0000	0.0000	-0.0001
N-3	0.0046	0.0039	-0.0078	0.0159	0.0071	0.0268	-0.0305	-0.0002	0.1591	0.3547	-0.2666	0.0000	1.0000	0.0354
N-4	0.0363	-0.2471	-0.0001	-0.0004	-0.0378	-0.0653	0.0164	-0.4077	0.0197	0.1127	-0.2589	-0.0001	0.0354	1.0000

Figure 2. Pearson's correlation matrix. Darker blue indicates higher correlations. Lighter blue indicates lower correlations.

Table 6 shows the correlation results based on the *VIF* values. The variables with the highest *VIF* values are H-3, H-4, H-5, H-6, S-1, and N-1. In all cases, the values are close to 1, placing them within the lower range of moderate correlation and very close to the threshold indicating no correlation.

Table 6. Correlation of variables based on the *VIF* value.

Variable Code	VIF Value
H-1	1.0065
H-2	1.3995
H-3	1.7510
H-4	1.7579
H-5	1.6968
H-6	1.7388
T-1	1.1136
T-2	1.2768
S-1	1.7078
S-2	1.2973
N-1	1.6352
N-2	1.0778
N-3	1.3422
N-4	1.3062

Overall, the results obtained from the three metrics (Pearson correlation coefficient, Cramér's *V*, and *VIF*) indicate that the level of correlation among the 14 variables is not significant. Therefore, all 14 variables originally defined for the study were retained for the subsequent analysis.

4.2. Descriptive Results

Table 7 summarizes the performance evaluation of the tested ML models. It shows the values of four key metrics used to compare performance of the ML tools tested and described in Section 3.5: accuracy, macro-recall, macro-precision, and macro-F1 score. These metrics correspond to the results obtained using the training dataset (80% of the total dataset). Model hyperparameters were optimized using stratified 10-fold cross-validation to achieve optimal results. Other techniques, such as Nested Cross-Validation (Macro-F1 score with SVM/Quadratic/No-binary SMOTE: 0.5437+ / −0.171) or Leave-One-Out Cross-Validation (Macro-F1 score with SVM/Quadratic/No-binary SMOTE: 0.5136+ / −0.269), were evaluated to address class imbalance. However, these approaches did not improve accuracy and resulted in increased computational cost.

As shown in Table 7, the generation of synthetic samples for the minority classes through SMOTE improves model performance across all evaluated metrics. The SVM model with a quadratic kernel achieves the best overall results. The SVM model achieves a test accuracy of 81.37% when the No-binary SMOTE technique is used, compared to 71.57% without SMOTE. Furthermore, the other metrics considered also show substantial improvement: macro-recall increases from 39.64% to 59.20%, macro-precision increases from 37.79% to 56.76%, and macro-F1 score increases from 38.29% to 57.29%. These results indicate that the SVM model exhibits strong classification capabilities in the high-dimensional feature space defined by the dataset and provides reliable predictions in non-linear classification scenarios. It is worth noting that the values achieved by macro-recall and macro-F1 score are very similar.

Table 7. Results of the ML classifiers metrics for the No-SMOTE and SMOTE-enhanced test data.

Model	Accuracy			Macro-Recall			Macro-Precision			Macro-F1 Score		
	No SMOTE	Binary SMOTE	No Binary SMOTE	No SMOTE	Binary SMOTE	No Binary SMOTE	No SMOTE	Binary SMOTE	No Binary SMOTE	No SMOTE	Binary SMOTE	No Binary SMOTE
SVM (Quadratic)	0.7157	0.7856	0.8137	0.3964	0.5876	0.5920	0.3779	0.5275	0.5676	0.3829	0.5698	0.5729
SVM (Gaussian)	0.7051	0.7576	0.7647	0.3790	0.4547	0.4603	0.3680	0.4167	0.4593	0.3604	0.4246	0.4591
SVM (Cubic)	0.6863	0.7267	0.7551	0.3429	0.4064	0.4337	0.3578	0.3987	0.4246	0.3116	0.3697	0.3746
SVM (Linear)	0.7255	0.7314	0.7451	0.3861	0.4097	0.4111	0.3925	0.4064	0.4216	0.3246	0.3637	0.3637
NN (MLP)	0.7345	0.7398	0.7416	0.3689	0.3978	0.4031	0.3598	0.3872	0.3996	0.3547	0.4148	0.4281
DT	0.7157	0.7246	0.7353	0.3575	0.3743	0.3889	0.3478	0.3794	0.3924	0.3101	0.3643	0.3942
BT (AdaBoost)	0.7036	0.7269	0.7319	0.3621	0.3598	0.3674	0.3602	0.3847	0.3713	0.3704	0.4214	0.4324
NB	0.7059	0.7235	0.7252	0.2876	0.3361	0.3477	0.2947	0.3236	0.3387	0.2649	0.3365	0.3461
RF	0.7059	0.7111	0.7187	0.3290	0.3673	0.3575	0.3313	0.3368	0.3487	0.2563	0.2874	0.2705
K-NN	0.7278	0.7147	0.7128	0.3275	0.3682	0.3889	0.3305	0.3698	0.3889	0.3304	0.3672	0.3842
LR	0.6918	0.7179	0.7052	0.3310	0.3597	0.3773	0.3476	0.3763	0.3838	0.2901	0.3314	0.3542

The SVM model with a quadratic kernel, which provides the best results, will be analyzed further by providing detailed per-class metrics. These metrics are shown in Tables 8 and 9.

Table 8. Test results: confusion matrix and per-class metrics for the SVM model. Data: No-SMOTE.

Test Confusion Matrix							
True Class	Test Predicted Class						
	A	B	C	D	E		
A	65	3	2	1	1		
B	1	2	3	3	0		
C	1	2	1	1	2		
D	0	1	1	3	2		
E	0	0	2	3	2		
Per-Class Metrics Test Results							
Accuracy	0.9118	0.8725	0.8627	0.8824	0.9020	Macro-accuracy	0.8863
Precision	0.9701	0.2500	0.1111	0.2727	0.2857	Macro-precision	0.3779
Recall	0.9028	0.2222	0.1429	0.4286	0.2857	Macro-recall	0.3964
F1 Score	0.9353	0.2353	0.1250	0.3333	0.2857	Macro-F1 score	0.3829

Table 9. Test results: confusion matrix and per-class metrics for the SVM model. Data: No-binary SMOTE -enhanced.

Test Confusion Matrix							
True Class	Test Predicted Class						
	A	B	C	D	E		
A	68	2	1	1	0		
B	1	4	1	3	0		
C	1	2	2	2	0		
D	0	0	1	3	3		
E	0	0	0	1	6		
Per-Class Metrics Test Results							
Accuracy	0.9412	0.9118	0.9216	0.8922	0.9608	Macro-accuracy	0.9255
Precision	0.9714	0.5000	0.4000	0.3000	0.6667	Macro-precision	0.5676
Recall	0.9444	0.4444	0.2857	0.4286	0.8571	Macro-recall	0.5920
F1 Score	0.9577	0.4706	0.3333	0.3529	0.7500	Macro-F1 score	0.5729

As shown in Table 8, the SVM model with a quadratic kernel achieves very good metrics for class A. However, its performance is considerably lower for classes B, C, D, and E. Precision, recall and F-1 scores are generally below 30% for these classes, while macro-precision, macro-recall and macro-F-1 scores remain below 40%. This behavior can be attributed to class imbalance and the dominant representation of cases involving full driver responsibility (Class A) in the original dataset (Table 1).

Nevertheless, the application of No-binary SMOTE (see Table 9) results in significant improvement in all metrics. Performance increases for all classes, particularly for the minority classes. Macro-accuracy increases from 88.63% to 92.55%, and macro-precision, macro-recall and macro-F1 scores are now above 50%. This is primarily due to the improvement of metrics in classes B, C, D and E, which now reach values above 30% in most cases. These improvements ensure better model performance across all classes when SMOTE is applied, particularly to the minority classes in the original dataset.

Figure 3 displays the ROC curves and the corresponding AUC values. In these approaches, each class is treated as positive while the remaining classes are treated as negative, generating one ROC curve per class—five curves in Figure 3a. Aggregated metrics (macro, micro, and weighted averages) are shown in Figure 3b, providing a single representative value for the multiclass scenario. Specifically, the micro-average compiles all predictions into a single binary classification task. This offers an overall performance measure, but it is potentially biased toward the majority class. The macro-average computes independently for each class and then calculates an unweighted mean, ensuring equal importance across classes. In contrast, the weighted-average applies class-specific weights based on sample size, providing a more representative evaluation of imbalanced datasets [38].

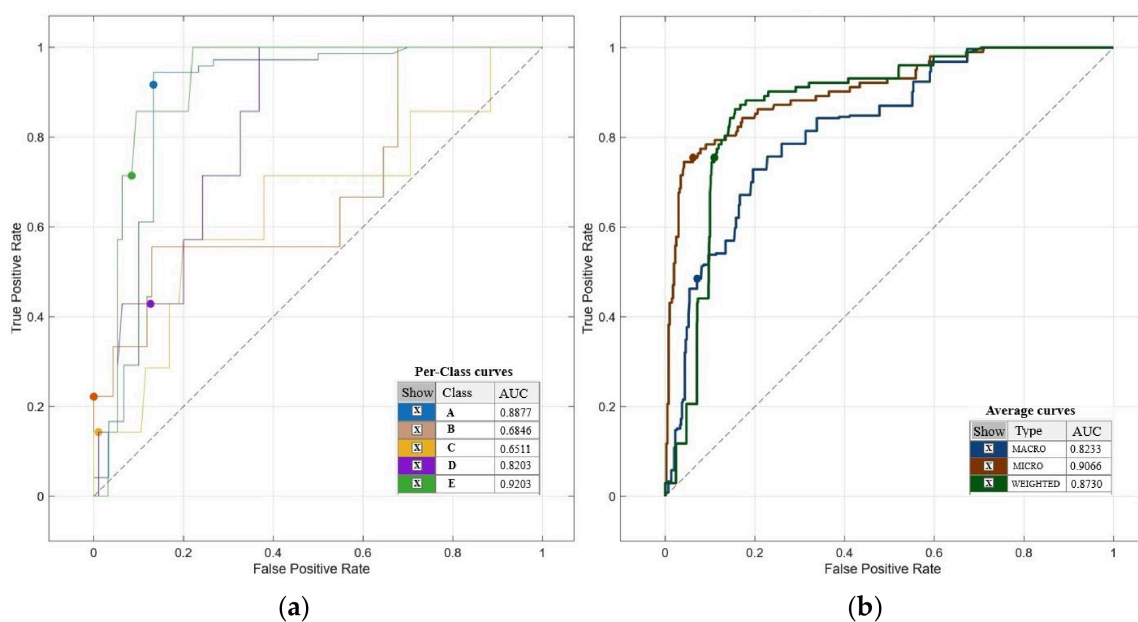


Figure 3. ROC curve and AUC values for SVM model data. No-binary SMOTE-enhanced. (a) Per-Class Curves (b) Average Curves.

4.3. Interpretation of Results with SHAP

After identifying and analyzing the best-performing model, SHAP was used to analyze the relative importance of each input variable in the classification process carried out with this model. Figures 4 and 5 illustrate the interpretability analysis using SHAP. SHAP provides a detailed quantification of feature importance by assigning an impact value to each feature for a given prediction. This approach enhances transparency by explaining how individual variables influence the model's decisions. SHAP values rank features according

to their contribution, offering actionable insights into the decision-making process of the SVM model. Figure 4 shows the global importance of each feature, highlighting those with the greatest influence on model predictions across all classes. Figure 5 provides class-specific SHAP summary plots for each liability category (Classes A-E), thereby enabling interpretation at both global and local levels. These visualizations provide a comprehensive understanding of how feature contributions vary between classes and support transparent and reliable decision-making in liability attribution for pedestrian crashes.

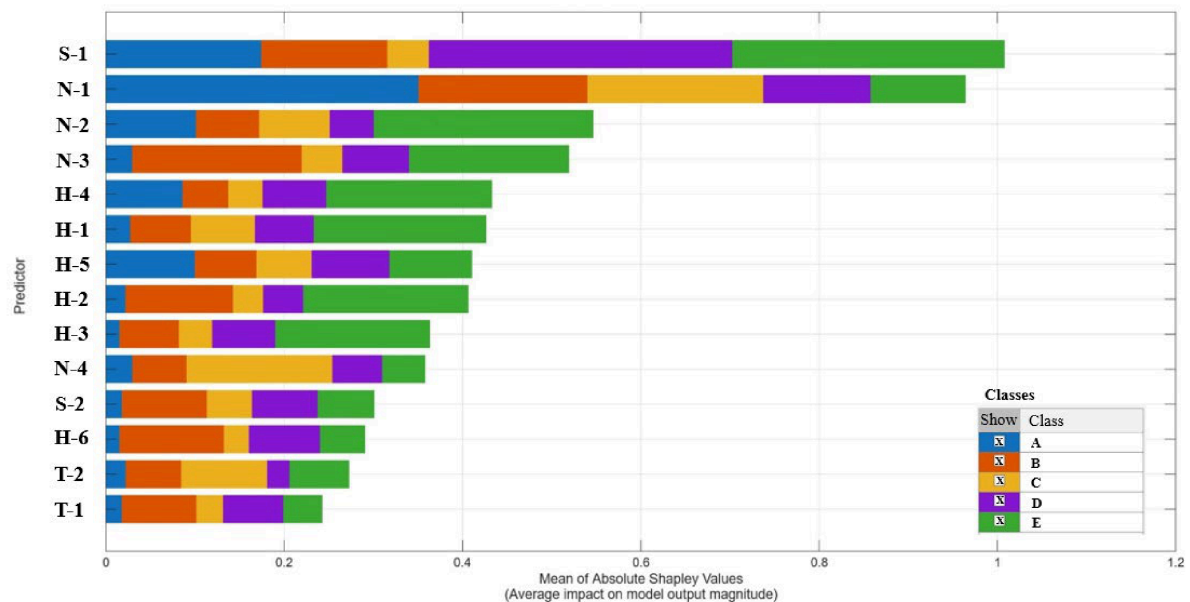


Figure 4. Feature importance derived from SHAP Values.

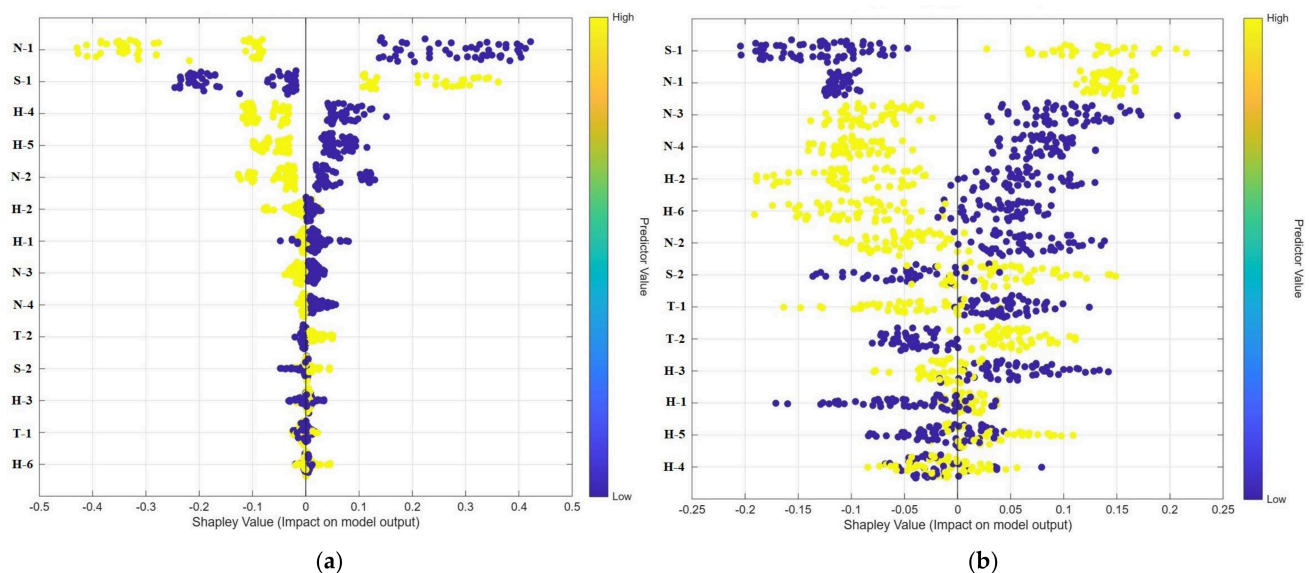


Figure 5. Cont.

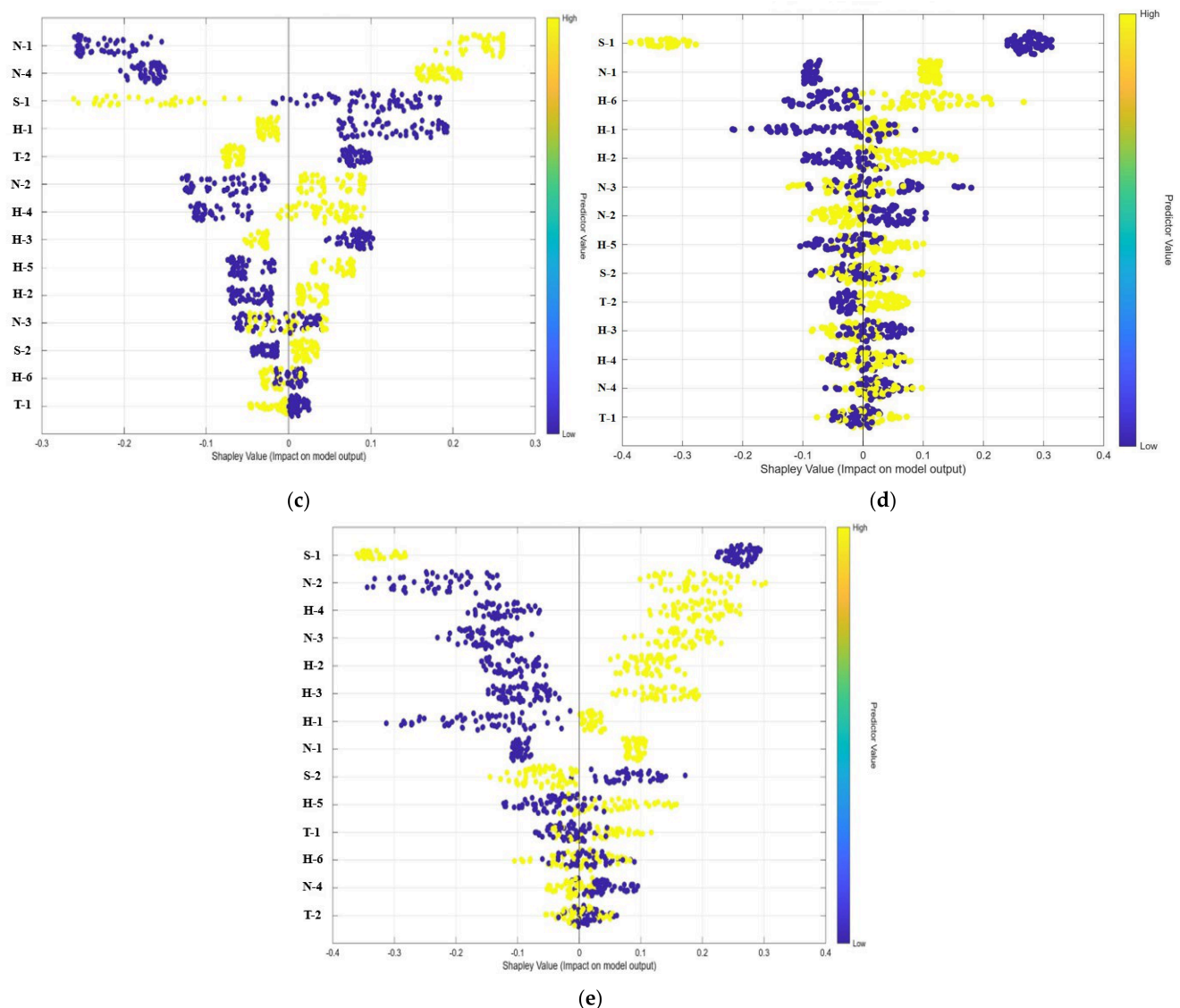


Figure 5. SHAP summary plot for each liability. (a) Class A, (b) Class B, (c) Class C, (d) Class D and (e) Class E.

5. Discussion

As mentioned earlier in Section 4.1, the preprocessing results revealed relatively moderate correlations between some variables, including H-3, H-4, H-5, and H-6. This finding is consistent with the nature of the domain. However, the observed correlation levels do not indicate severe redundancy among the 14 variables. While there is some degree of dependence, it is minimal and does not significantly impact the stability of the model or invalidate SHAP-based interpretations. Therefore, it can be concluded that the association between these variables is weak. Consequently, reducing the number of variables does not significantly improve the results of this study. Nevertheless, given the presence of correlated features, SHAP values should be interpreted with caution, as they reflect shared or conditional contributions rather than strictly independent effects.

This section offers a thorough interpretation of the results, with a focus on the SVM classifier, which was identified as the most effective model for liability classification. As shown in Table 7, the SVM achieved an overall accuracy of 81.37%, a macro-recall of 59.20%, a macro-precision of 56.76%, and a macro-F1 score of 57.29%. These metrics are superior to those obtained without applying SMOTE, as also verified by analyzing Tables 8 and 9. These macro-average metrics confirm substantial predictive reliability across all classes,

ensuring balanced performance even under class imbalance. Furthermore, as illustrated in Figure 3b, the SVM model demonstrates very good discriminative capacity, with a macro-average AUC of 82.33%, a micro-average AUC of 90.67%, and a weighted-average AUC of 87.31%. The class-specific AUC values in Figure 3a range from 92.03% (Class E) to 65.11% (Class C). These values demonstrate the model's acceptable ability to distinguish levels of responsibility in traffic crashes involving pedestrians.

Figure 3 shows the ROC curves, which validate the robustness of the classifier and reveal a consistent trade-off between true positive and false positive rates across thresholds. This is particularly relevant for imbalanced datasets because it ensures that the model achieves high accuracy while consistently identifying true instances of each liability class. Classes A (88.77%) and E (92.03%) have the highest accuracy rates. This suggests that the model performs best in clear-cut cases, such as when the responsibility lies entirely with the driver or pedestrian. However, the model requires further refinement for ambiguous scenarios, such as those in Classes B, C, and D, where responsibility is shared by the involved parties in varying percentages. The comparatively lower metrics and AUC values of the model suggest that it has difficulty classifying samples in these classes consistently and accurately.

As shown in Figure 4, the results provided by SHAP indicate that certain explanatory variables (EVs), such as pedestrian crash location and possession of a driving license, are highly informative in distinguishing responsibility levels, especially in Classes D, E and A, respectively. However, these EVs are less effective in borderline cases, such as Classes B (70/25 driver-pedestrian) and D (25/75 driver-pedestrian). In these cases, contextual interpretation plays a more significant role because they can be confused with Classes A (100/0 driver-pedestrian) and E (0/100 driver-pedestrian), respectively. The remaining EVs have a smaller influence on assigning responsibility levels. They have a low influence on assigning a crash to Class A but a higher influence on assigning a crash to Class D, indicating that assigning full responsibility to drivers (Class A) is largely determined by EVs location and possession of a driver license, while assigning full responsibility to the pedestrian requires evaluating most of the EVs. In this context, Class C (50/50 driver-pedestrian) also requires special attention because it represents complex scenarios that demand expert review, as there is no clear way to identify this case. Assignment to this class requires a more refined approach from competent authorities to ensure accurate classification and legal interpretation.

SHAP analysis provides transparency by quantifying each feature's contribution to predictions. When interpreting the results provided by the SVM model for each liability class, SHAP value plots reveal how EVs influence the model's predictions. SHAP values offer a consistent framework for interpreting how each feature contributes to predicting a given class. This provides a transparent view of the factors considered by the model when assigning liability levels. As illustrated in Figure 4 and previously discussed, possession of a driving license and crash location are the most influential features, as they have the highest average SHAP values. These results suggest that these two variables play a critical role in the model's decision-making process regarding liability distribution. The remaining variables have significantly lower SHAP values, ranging from 0.50 to 0.25. Compare these with the values close to 1 shown by location and driving license: this indicates that they have less influence on the process.

Figure 5 summarizes SHAP distributions, showing how each EV affects the probability of assigning a specific liability class. The figure also reveals coherent patterns across parties' behaviors and contextual conditions. In the plot, each feature is represented by a horizontal row, and individual data points are colored according to the feature's value (yellow indicating higher values and blue indicating lower values). Each point reflects the

direction and magnitude of the feature's impact on the model's output. Points to the right indicate an increase in the predicted probability of the positive class, while points to the left indicate a decrease. Clustering on one side of the axis suggests a consistent influence of a feature, while wide dispersion indicates variability and potential interaction with other features. The order of the rows for each class reflects the aggregated magnitude of SHAP values across all observations, providing a ranked overview of feature importance in the model's decision-making process.

To properly understand the meaning of each summary plot, it is important to keep in mind how binary values are attributed to each EV, as described in Table 2. Assigning different levels of responsibility to the driver or pedestrian depends on accomplishing each of the corresponding factors. For instance, assigning a value of "1" to the factor "location" (the crash occurred within the area of influence of the pedestrian crossing) increases the probability of assigning a higher level of responsibility to the driver. Conversely, assigning a value of "0" reduces that probability. From the pedestrian's point of view, the probabilities are opposite.

For Class A, driver-related variables such as driver license possession, location, driver violation, inadequate speed adaptation, low attention, delayed reaction, and alcohol/drug consumption by the driver exert the strongest positive influence on assigning a higher level of responsibility to the driver, while alcohol/drug consumption by the pedestrian and favorable visibility reduce it. In contrast, Class E is mainly explained by pedestrian violations (location, distraction, alcohol/drug consumption, and visibility). At the same time, driver diligence increases the pedestrian responsibility. Intermediate classes display mixed signatures. Class B involves cases of significant driver negligence combined with minor pedestrian contributions, demonstrating substantial horizontal dispersion (i.e., interactions). In contrast, Class D reflects the predominance of pedestrian fault with limited driver involvement. Finally, Class C involves a wide range of risks, with both the driver and the pedestrian exhibiting negligent behavior. In these cases, expert human analysis is advisable to determine if both parties are equally responsible, as the SHAP value analysis shown in Figure 4 also suggests.

Overall, the model performs well in the extreme classes, Class A (100% driver liability) and Class E (100% pedestrian liability). High recall values in these classes indicate reliable identification. In contrast, performance is notably lower in the intermediate classes. Class C (50–50 liability) poses the greatest challenge due to its inherent ambiguity, as reflected by its low recall value (see Table 9). The corresponding confusion matrix (Table 9) shows that most misclassifications occur among these classes. Class D (25% driver liability and 75% pedestrian liability) has the highest number of misclassifications. This pattern aligns with the SHAP analysis, which reveals more balanced and less distinctive feature contributions in the intermediate classes. This confirms their reduced separability and higher uncertainty. In this sense, SHAP is an insightful means of interpreting the results from a complex algorithm such as SVM. This technique allows one to evaluate the importance and direction of each feature's impact on the model's outcome and capture the complex, nonlinear joint impacts of the features. This is especially useful for decision-makers determining liability levels in traffic crashes. Furthermore, the results confirm the SVM classifier's suitability for the legal liability classification task. SVM offers high predictive performance and interpretability, both of which are essential for applications in forensic and policy-oriented contexts. The goal is not to replace human operators, but to provide them with useful, precise tools that facilitate their challenging work of making complex decisions.

The goal of this work is to predict liability levels in pedestrian crashes based on judicial rulings, and then to evaluate the influence of the factors considered in those predictions. Little research is available in this field. Therefore, the results of this work cannot be directly

compared to those reported in other works, though meaningful extrapolation can be made using relevant sources. The performance metrics obtained in this research largely align with those reported in previous studies in other fields [39–43] and in road safety [17,44–48]. Despite the differences in the objectives and specific characteristics of each study, the results achieved show that the model proposed in this work is a reliable classification tool that can also provide useful explanations of the results.

Similarly, when comparing these studies, certain features consistently emerge as relevant factors for model implementation, such as lighting conditions, driving licenses, and alcohol/drug consumption. However, caution is required when interpreting these comparisons. Differences in problem formulation, feature sets, and target variables limit direct comparability and may lead to misleading conclusions if contextual factors are ignored. While this research focuses on liability attribution based on judicial decisions, prior studies have examined different scenarios. Despite these limitations, the convergence of key explanatory variables across different research contexts underscores their relevance for predictive modeling in traffic safety. This suggests that, although the modeling objectives differ, certain behavioral and environmental factors are critical determinants of crash outcomes and liability attribution. Future research should explore standardized benchmarks and shared datasets to enable more rigorous cross-study comparisons and enhance the generalizability of findings.

6. Conclusions

This study shows that an SVM with a quadratic kernel outperforms other algorithms at predicting liability attribution in pedestrian crashes. This finding confirms the SVM's suitability for modeling high-dimensional, non-linear decision boundaries, which are often found in legal and behavioral datasets. Furthermore, analyzing the dataset using the Pearson's correlation coefficient, Cramer's *V* metric, and VIF techniques indicates that the variables are not significantly correlated. Consequently, reducing the number of variables does not significantly improve the results.

The process of assigning responsibilities includes preparing the data, encoding the features as binary, balancing the classes using SMOTE, and training multiple ML models (SVM, NN, DT, BT, NB, RF, K-NN, and LR) with cross-validation. The best-performing model was evaluated using accuracy, F1-score, precision, recall, and ROC-AUC. Then, the results were interpreted with SHAP for feature contribution analysis. SHAP enhances transparency by explaining feature-level contributions to model predictions, thereby improving interpretability. This capability is essential for building trust in AI-assisted legal processes because it allows stakeholders to validate algorithmic outputs and ensure their alignment with legal principles.

SHAP is a useful tool for interpreting the results of complex algorithms, such as SVM. It evaluates the importance and direction of each feature's impact on the model's outcome by capturing the combined, nonlinear impacts of features. This is especially useful for decision-makers determining liability levels in traffic crashes. The goal is not to replace human operators, but rather to provide them with useful and accurate tools that facilitate their arduous task of making complex decisions. Furthermore, using SHAP to interpret data identifies the most influential variables in a pedestrian crash. Policymakers can use this information to understand each pattern, enabling critical evaluation rather than blind reliance. Consequently, they can adopt new measures to reduce traffic crashes and increase road safety.

From a jurisprudential perspective, integrating ML models into liability attribution processes has significant potential to complement judicial decision-making by providing data-driven insights. These models can identify patterns in court rulings and highlight in-

fluent factors, such as lighting conditions, crash location, reaction time, and alcohol/drug involvement. This provides empirical evidence to support legal reasoning. However, ML-based predictions should remain tools that support decisions rather than substitute for judicial discretion. First, ML models should support judicial decision-making. As previously mentioned, the SVM-SHAP framework can assist judges and legal professionals by offering objective information on attribution of responsibility and minimizing inconsistencies. SHAP analysis, on the other hand, identifies high-impact factors, such as driver alcohol or drug use and driver and pedestrian violations. These factors can inform road safety policies and preventive measures. They are essential for developing new strategies to reduce traffic crashes. Finally, integrating interpretable ML models into traffic control and urban planning workflows can improve the fairness and efficiency of mobility systems, enhance road design, and promote sustainability.

In conclusion, this study shows that ML, particularly SVM, integrated with interpretability tools like SHAP, can substantially improve liability attribution analysis in pedestrian crashes. While these models cannot replace human judgment, they can support legal reasoning by promoting consistent, data-driven transparent decision-making. Incorporating interpretable ML models into legal workflows can promote fairness and efficiency in urban mobility governance. This research can help competent authorities determine liability in pedestrian crashes. The model distinguishes clear-cut cases from those requiring more thorough, specialized evaluation. It brings us closer to the concept of a “robot judge” by helping judges and JTPs make more efficient and objective final decisions in similar cases. This would also enable them to allocate their time and resources to more complex cases requiring in-depth investigation by specialists. Furthermore, identifying the most influential variables in pedestrian crashes would allow policymakers to allocate their resources more effectively and efficiently to new measures that improve road safety.

6.1. Limitations

When interpreting the results, several limitations of this study should be recognized. First, the dataset does not contain a large number of pedestrian crashes. Additionally, the instability of minority groups is due to an imbalanced database. Ideally, a larger, more balanced database would be used; however, obtaining data containing personal information is often complicated by strict data protection regulations. The time period covers the most up-to-date and complete database that each institution could provide and to which we had access (Badajoz City).

Another limitation is binary variables. Some variables, such as the location of pedestrian crashes (inside or outside a crosswalk) or whether the car has a valid inspection certificate, have two precise states (0 or 1). However, other variables could yield better results with more than two states. For instance, variables such as alcohol level (higher levels increase risk due to longer reaction times), lighting conditions (which impact visibility), and speed (higher speeds increase risk) could be graded. In this sense, SMOTE provides nonbinary values in synthetic samples. Therefore, some variables describing crashes should probably also have non-binary values. This could be a more general and precise way of describing traffic crashes and deserves further study. However, not every traffic crash has a unique set of variables, and extracting this information is arduous. For this study, 14 variables were used, which were the variables that could be obtained from the initial data provided by the participating institutions.

Additionally, the ML models trained on such data may risk implicitly learning and reproducing existing decision patterns, including potential systemic biases present in the underlying legal and administrative processes. Therefore, we conceptualize the model as a decision support tool rather than a decision-making system, and its results should be interpreted within a legal and institutional context. This underscores the importance

of exercising caution when applying the model to other jurisdictions with different legal practices or interpretive frameworks.

On the other hand, extracting simplified decision rules from techniques such as SHAP analysis is a promising approach when a traffic crash involves a small, consistent number of variables. However, the resulting inference system would likely be impractical because the more rules there are, the less applicable the model will be.

Other limitations that could be included are the potential for noise in the labels of court rulings, the lack of external validation by competent authorities to certify the reliability of the results obtained, the ability to adapt the tool to the specific legal norms in each jurisdiction, the ability to omit confounding factors, and the questionable realism of the synthetic samples generated by SMOTE.

6.2. Future Research Directions

To advance this line of research and generate more generalizable and robust conclusions, future studies should aim to expand the scope and depth of the analysis.

A key priority should be to increase the sample size and include broader geographic representation by incorporating data from other cities, and even other countries, to capture greater jurisdictional diversity. This also involves increasing the number of variables and ensuring that they are not binary.

In any case, it is necessary to properly process all input variables using new techniques to explore correlations and combinations of variables that could improve accuracy. Techniques such as recursive feature elimination, L1 regularization, and principal component analysis (PCA) are ideal for validating the non-redundancy of the selected features. Additionally, preprocessing alternatives such as SMOTE-NC, class weighting, cost-sensitive SVM, random subsampling, ADASYN, Borderline-SMOTE, or balanced ensemble methods could be useful for addressing class imbalance.

Future research should incorporate multi-jurisdictional datasets to validate the model across diverse legal systems and cultural contexts. Additionally, exploring the interaction between pedestrian and driver behavior and crashes is essential. This can be done by comparing it with other studies and developing behavioral questionnaires to identify influencing variables [49]. Furthermore, research should investigate hybrid approaches that combine ML with rule-based legal reasoning to achieve normative alignment. Interactive decision-support systems should also be developed for judges and legal practitioners to visualize model explanations and simulate alternative scenarios. Finally, there should be greater integration with intelligent transportation systems to enable proactive risk assessment and real-time liability prediction.

Future research should include a rigorous evaluation of the predictive confidence of the most effective model. In this study, the model is an SVM. Calibrating the model's predictive probability, defining a high uncertainty threshold, and explicitly indicating cases of high uncertainty are important parameters to consider in new analyses.

6.3. Core Ethical Principles

This work is not meant to replace human operators; rather, it is meant to support them. The principle of human oversight remains intact. The model's results support decision-making processes but do not replace the judgment of judges, prosecutors, traffic officers, or experts. A duly trained human authority in the field of road safety must always make the final decision.

The models used are supported by scientific literature and ensure transparency and explainability. The interpretability results can reproduce the research enables mechanisms that justify and audit these results. This allows legal professionals to identify the most influential factors in each final decision. It facilitates proper justification and potential review.

Additionally, these models can be periodically evaluated to identify and address potential biases stemming from the training data or the variables used as legal regulations evolve and jurisdictional guidelines are updated. This ensures objectivity.

Finally, the quality, integrity, and representativeness of the data are guaranteed. The models have been trained using verified and validated information from reliable databases. This minimizes errors and improves the robustness of the final results. Similarly, data privacy and protection principles have been observed to ensure the confidentiality of the information used throughout the entire process. Consequently, a larger database is unavailable for this research because the originating institutions have guaranteed these principles.

Author Contributions: Conceptualisation, F.C.G.-P., M.A.J.-M. and A.M.-S.; methodology, F.C.G.-P., M.A.J.-M. and A.M.-S.; software, F.C.G.-P. and A.M.-S.; validation, F.C.G.-P., M.A.J.-M. and A.M.-S.; formal analysis, F.C.G.-P.; investigation, F.C.G.-P. and A.M.-S.; resources, A.M.-S.; data curation, F.C.G.-P. and A.M.-S.; writing—original draft preparation, F.C.G.-P.; writing—review and editing, F.C.G.-P., M.A.J.-M. and A.M.-S.; visualisation, F.C.G.-P., M.A.J.-M. and A.M.-S.; supervision, F.C.G.-P., M.A.J.-M. and A.M.-S.; project administration, F.C.G.-P., M.A.J.-M. and A.M.-S.; funding acquisition, F.C.G.-P. and M.A.J.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has been co-financed at 85% by the European Union, European Regional Development Fund (FEDER “Una Manera de Hacer Europa”), and the Government of Extremadura, grant number GR24104, Management Authority. Ministry of Finance.

Data Availability Statement: Some of the data related to this study have not been deposited in a public repository, are confidential and are available on request in the database of the Judicial Traffic Police of the Badajoz Local Police. The rest of the data are available in the database of the Spanish Judiciary (<https://www.poderjudicial.es/search/indexAN.jsp>, accessed on 20 January 2025).

Acknowledgments: The authors would like to thank the Local Police of Badajoz and the Spanish Judiciary. We also thank the reviewers for their comments to improve this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SVM	Support Vector Machines
NN	Neural Network
MLP	Multi-Layer Perceptron
DT	Decision Trees
BT	Boosted Trees
NB	Naïve Bayes
RF	Random Forest
K-NN	K-Nearest Neighbors
LR	Logistic Regression
SMOTE	Synthetic Minority Oversampling Technique
SHAP	SHapley Additive exPlanations
ML	Machine Learning
JTP	Judicial Traffic Police
BLP	Badajoz Local Police
SJ	Spanish Judiciary
VIF	Variance Inflation Factor
ROC	Receiver Operating Characteristic Curve
AUC	Area Under Curve
Ovo	One-vs-One
OvR	One-vs-Rest
EVs	Explanatory Variables

References

1. World Health Organization (WHO). *Global Status Report on Road Safety*; WHO: Geneva, Switzerland, 2024; Available online: <https://www.who.int/publications/i/item/> (accessed on 26 February 2026).
2. European Road Safety Observatory. *Annual Statistical Report on Road Safety in the EU*; European Commission: Brussels, Belgium, 2023; Available online: https://transport.ec.europa.eu/background/road-safety-statistics-2023_en (accessed on 7 February 2026).
3. European Documentation Centre. *2024 Sees 3% Drop in EU Road Fatalities, Yet Progress Remains Slow*; European Commission: Brussels, Belgium, 2025; Available online: https://ec.europa.eu/commission/presscorner/detail/es/ip_25_789 (accessed on 22 January 2026).
4. Onda Cero. El Ayuntamiento de Badajoz gana el premio Visión Zero Municipal al no Haber Registrado Fallecidos por Accidentes en 2021. Available online: https://www.ondacero.es/emisoras/extremadura/badajoz/noticias/ayuntamiento-badajoz-gana-premio-vision-zero-municipal-haber-registrado-fallecidos-accidentes-2021_202210246356c8559719c700015b8ddb.html (accessed on 16 December 2025).
5. European Parliament. *Road Safety Policy Framework 2021–2030: Recommendations on Next Steps Towards, Vision Zero, 2021*; European Parliament: Brussels, Belgium, 2021; Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021IP0407> (accessed on 7 January 2026).
6. Casado-Sanz, N.; Guirao, B.; Attard, M. Analysis of the Risk Factors Affecting the Severity of Traffic Accidents on Spanish Crosstown Roads: The Driver’s Perspective. *Sustainability* **2020**, *12*, 2237. [CrossRef]
7. Hochstatter, J. Legal Liability. In *Global Encyclopedia of Public Administration, Public Policy, and Governance*; Springer: Berlin/Heidelberg, Germany, 2018.
8. Sanz Encinar, A. *El Concepto Jurídico de Responsabilidad en la Teoría General del Derecho*; Universidad Autónoma de Madrid: Madrid, Spain, 2000.
9. Ahmed, S.; Hossain, M.A.; Bhuiyan, M.M.I.; Ray, S.K. A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity. In Proceedings of the 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), London, UK, 20–22 December 2021; pp. 390–397.
10. Shrinivas, V.; Bastien, C.; Davies, H.; Daneshkhan, A.; Hardwicke, J. Parameters influencing pedestrian injury and severity—A systematic review and meta-analysis. *Transp. Eng.* **2023**, *11*, 100158. [CrossRef]
11. Behboudi, N.; Moosavi, S.; Ramnath, R. Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of ML Techniques. *arXiv* **2024**. [CrossRef]
12. Cappelli, G.; Nardoian, S.; D’Apuzzo, M.; Nicolosi, V. Pedestrian Crash Severity Prediction and Contributory Factors Analysis by Using Machine Learning Methods. In *Computational Science and Its Applications ICCSA*; Springer: Berlin/Heidelberg, Germany, 2025. [CrossRef]
13. Jing, N.; Yang, M.; Cheng, S.; Dong, Q.; Xiong, H. An efficient SVM-based method for multi-class network traffic classification. In Proceedings of the 30th IEEE International Performance Computing and Communications Conference, Orlando, FL, USA, 17–19 November 2011; pp. 1–8.
14. Chen, C.; Zhang, G.; Qian, Z.; Tarefder, A.; Tian, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accid. Anal. Prev.* **2016**, *90*, 128–139. [CrossRef] [PubMed]
15. Qawasmeh, B.; Oh, J.-S.; Kwizile, V.; Qawasmeh, D.; Al Tawil, A.; Aldalqamouni, A. Analyzing daytime/nighttime pedestrian crash patterns in Michigan using unsupervised machine learning techniques and their potential as a decision-making tool. *Open Transp. J.* **2024**, *18*, e26671212352718. [CrossRef]
16. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
17. Adeel, M.; Asad Khattak, A.J.; Mishra, S.; Thapa, D. Enhancing work zone crash severity analysis: The role of synthetic minority oversampling technique in balancing minority categories. *Accid. Anal. Prev.* **2024**, *208*, 107794. [CrossRef] [PubMed]
18. Abokadr, S.; Azman, A.; Hamdan, H.; Amelina, N. Handling Imbalanced Data for Improved Classification Performance: Methods and Challenges. In Proceedings of the 3rd International Conference Emerging Smart Technologies and Applications, Taiz, Yemen, 10–11 October 2023; pp. 1–8.
19. Zhang, X.; Xue, Q.; Guo, W.; Tan, J. Enhancing model transparency: A comparative analysis of SHAP and LIME in explaining traffic accident prediction models. In Proceedings of the 2024 International Conference Artificial Intelligence and Autonomous Transportation, AIAT, Beijing, China, 5–7 December 2024; pp. 48–56.
20. Campón Domínguez, J.A. *El Diseño de Una Base de Datos de Investigaciones en Profundidad Sobre Atropellos a Peatones*. Ph.D. Thesis, Universidad Carlos III, Madrid, Spain, 2015.
21. Moreno-Sanfélix, A.; Gragera-Peña, F.C.; Jaramillo-Morán, M.A. An improvement of the conceptual system of the sequential events model of road crashes (i-MOSES). *Heliyon* **2024**, *10*, e37268. [CrossRef] [PubMed]
22. Rodgers, J.L.; Nicewander, W.A. Thirteen ways to look at the correlation coefficient. *Am. Stat.* **1988**, *42*, 59–66. [CrossRef]

23. Jansen, M.; Claeskens, G. Cramér–Rao Inequality. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011. [\[CrossRef\]](#)
24. Coppersmith, D.; Hong, S.J.; Hosking, J.R.M. Partitioning Nominal Attributes in Decision Trees. *Data Min. Knowl. Discov.* **1999**, *3*, 197–217. [\[CrossRef\]](#)
25. Kim, J.H. Multicollinearity and misleading statistical results. *Korean J. Anesthesiol.* **2019**, *72*, 558–569. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Belsley, D.A. A Guide to using the collinearity diagnostics. *Comput. Sci. Econ. Manag.* **1991**, *4*, 33–50. [\[CrossRef\]](#)
27. Kutner, M.H.; Nachtsheim, C.J.; Neter, J. *Applied Linear Regression Models*, 4th ed.; McGraw-Hill Irwin: Columbus, OH, USA, 2004.
28. Barysè, D.; Sarel, R. Algorithms in the court: Does it matter which part of the judicial decision-making is automated? *Artif. Intell. Law* **2023**, *32*, 117–146. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Schölkopf, B.; Smola, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
30. Zhang, S.; Liu, Q.; Fan, M.; Mu, W.; Feng, J. Multi-view least squares support vector classifiers with the principles of complementarity and consensus. *Neurocomputing* **2025**, *657*, 131647. [\[CrossRef\]](#)
31. Gaudreault, J.G.; Branco, P.; Gama, J. An analysis of performance metrics for imbalanced classification. In *Proceedings of the International Conference on Discovery Science*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12986, pp. 67–77. [\[CrossRef\]](#)
32. Takahashi, K.; Yamamoto, K.; Kuchiba, A.; Koyama, T. Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Appl. Intell.* **2022**, *52*, 4961–4972. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Shapley, L.S. A value for n-person games. In *Contributions to the Theory of Games*; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–317.
34. Lunderg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 4–9 December 2017.
35. Kashifi, M.T. Investigating two-wheelers risk factors for severe crashes using an interpretable machine learning approach and SHAP analysis. *IATSS Res.* **2023**, *47*, 357–371. [\[CrossRef\]](#)
36. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. [\[CrossRef\]](#)
37. Lu, J.; Huang, Z.; Yang, L. Road Traffic Accident Severity Prediction under Unbalanced Data. In *Proceedings of the 4th International Conference Computer, Big Data and Artificial Intelligence (ICCBD+AI)*, Guiyang, China, 15–17 December 2023; pp. 650–654.
38. Bugaj, M.; Wrobel, K.; Iwaniec, J. Model explainability using SHAP values for LightGBM predictions. In *Proceedings of the IEEE XVIIth International Conference Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, Polyana, Ukraine, 12–16 May 2021; pp. 102–106.
39. Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V. Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access* **2021**, *9*, 39707–39716. [\[CrossRef\]](#)
40. Dablain, D.; Krawczyk, B.; Chawla, N.V. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 6390–6404. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Ahmad Khan, A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778. [\[CrossRef\]](#)
42. Gabrielli, G.; Melioli, A.; Bertini, F. Corporate financial distress prediction: A machine learning approach in the era of big data. *J. Account. Organ. Change* **2025**, *22*, 31–65. [\[CrossRef\]](#)
43. Agrawal, R.; Hamdare, S.; Ghosh, D.; Ghenand, K.; Gupta, T.; Patel, P.; Chauhan, S.; Upasani, S.; Agrawal, P. Improving Predictive Performance in Telecom Churn Modeling with Hybrid SMOTE and GAN-Based Synthetic Data Generation. *Int. J. Comput. Intell. Syst.* **2026**, *19*, 141. [\[CrossRef\]](#)
44. Alshriem, M.; Yang, Y. Prediction of Large-Scale Traffic Accident Severity in Qatar: A Binary Reformulation Approach for Extreme Class Imbalance with Interpretable AI. *Future Transp.* **2026**, *6*, 88. [\[CrossRef\]](#)
45. Saleem, J.; Islam, R.; Altas, I.; Islam, M.Z. Enhancing Darknet Traffic Classification: Integrating Traffic-Aware SMOTE and Adaptive Weighted Feature Aggregation. *J. Cybersecur. Priv.* **2026**, *6*, 68. [\[CrossRef\]](#)
46. Chen, S.; Cui, B.; Chang, A. An adaptive data rebalancing framework for real-time traffic risk prediction. *Sci. Rep.* **2026**, *16*, 8882. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Díaz-Aparicio, J.; Rodríguez-Esparza, E.; Fajardo-Calderín, J.; Onieva, E. Studying the impact of data preprocessing, hyperparameter tuning and machine learning algorithms in crash prediction explainability. *Array* **2026**, *30*, 100743. [\[CrossRef\]](#)

48. Moreno-Sanfélix, A.; Gragera-Peña, F.C.; Jaramillo-Morán, M.A. Evaluation of the level of responsibility in pedestrian crashes using machine learning algorithms. *Sci. Rep.* **2026**, *16*, 12093. [[CrossRef](#)] [[PubMed](#)]
49. Esmaili, A.; Aghabayk, K.; Parishad, N.; Stephens, A.N. Investigating the interaction between pedestrian behaviors and crashes through validation of a pedestrian behavior questionnaire (PBQ). *Accid. Anal. Prev.* **2021**, *153*, 106050. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.