vehicles

Article

Driver Injury Prediction and Factor Analysis in Passenger Vehicle-to-Passenger Vehicle Collision Accidents Using Explainable Machine Learning

Peng Liu [1], Weiwei Zhang [1,2,*], Xuncheng Wu [1], Wenfeng Guo [3] and Wangpengfei Yu [1]

[1] School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; liupeng@sues.edu.cn (P.L.); pengfeixiaoyu@smarvcte.com (W.Y.)
[2] School of Automotive Studies, Tongji University, Shanghai 201804, China
[3] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; gwf0330@163.com
* Correspondence: zhangww.2021@tsinghua.org.cn

Abstract: Vehicle accidents, particularly PV-PV collisions, result in significant property damage and driver injuries, causing substantial economic losses and health risks. Most existing studies focus on macro-level predictions, such as accident frequency, but lack detailed collision-level analysis, which limits the precision of severity prediction. This study investigates various accident-related factors, including environmental conditions, vehicle attributes, driver characteristics, pre-crash scenarios, and collision dynamics. Data from NHTSA's CRSS and FARS datasets were integrated and balanced using random over-sampling and under-sampling techniques to address severity-level data imbalances. The mRMR algorithm was employed for feature selection to minimize redundancy and identify key features. Five advanced machine learning models were evaluated for severity prediction, with XGBoost achieving the best performance: 84.9% accuracy, 84.85% precision, 84.90% recall, and an F1-score of 84.87%. SHAP analysis was utilized to interpret the model and conduct a comprehensive analysis of accident features, including their importance, dependencies, and combined effects on severity prediction. This study achieved high accuracy in predicting accident severity across all levels in PV-PV collisions. Moreover, by integrating the SHAP model interpretation method, we conducted detailed feature analysis at global, local, and individual case levels, thereby filling the gap in PV-PV accident severity prediction and feature analysis.

Keywords: explainable machine learning; driver injury severity prediction; accident datasets; feature extraction; accident feature analysis; SHAP

Academic Editor: Mohammed Chadli

Received: 20 March 2025
Revised: 24 April 2025
Accepted: 28 April 2025
Published: 3 May 2025

Citation: Liu, P.; Zhang, W.; Wu, X.; Guo, W.; Yu, W. Driver Injury Prediction and Factor Analysis in Passenger Vehicle-to-Passenger Vehicle Collision Accidents Using Explainable Machine Learning. *Vehicles* **2025**, *7*, 42. https://doi.org/10.3390/vehicles7020042

1. Introduction

Passenger vehicles (PVs), as motor vehicles designed to transport people and their belongings, have significantly enhanced the convenience of human mobility [1–3]. PVs are primarily categorized into two major groups: passenger cars and light trucks [4]. Among various vehicle types, PVs not only constitute the majority of the vehicle fleet but also account for the highest proportion of traffic accidents [5]. According to the 2024 analysis conducted by the National Highway Traffic Safety Administration on the 2022 Crash Report Sampling System data [6], more than 64% of police-reported crashes involved two-vehicle collisions, with passenger vehicle to passenger vehicle (PV-PV) crashes accounting for the highest proportion at 73.4%. Drivers are the most severely affected individuals in traffic accidents, representing 67.7% of all road traffic casualties in 2021. Therefore, understanding

the factors influencing collision occurrences and injury severity is crucial for enhancing traffic safety.

In the field of accident severity prediction, there are two distinct approaches. Macroscopic prediction examines accident severity from an overall road and traffic perspective, providing insights for broad-scale road management and policymaking. Microscopic prediction, conversely, delves into specific accident scenarios, enabling in-depth analysis to enhance driving safety. Predicting and analyzing the severity of driver injuries in PV-PV accidents represents a highly intricate task. It demands a comprehensive consideration of dual-vehicle characteristics, driver attributes, environmental conditions, causal factors, and the dynamics both before and during the collision [7]. As illustrated in Figure 1, in a representative accident scenario, a sedan attempting to overtake a cargo van from the left encountered a sudden pedestrian incursion into the lane. Obstructed by roadside greenery, the driver's view was impaired, delaying the detection of the pedestrian. The subsequent emergency braking and leftward swerve resulted in a rear-end collision with the cargo van, underscoring the complexity inherent in real-world accident scenarios. Consequently, for accurate accident prediction and analysis, it is essential to collect and analyze data across multiple aspects, spanning the occurrence, development, and consequences of the accidents. Omitting key information can lead to inaccurate or unreasonable predictions [8–10].

Despite the critical importance of predicting the severity of passenger vehicle-passenger vehicle (PV-PV) accidents, this area remains notably understudied, representing a significant gap in the literature. While there are studies on two-vehicle accidents, which are somewhat analogous to PV-PV accidents. Current research on two-vehicle collisions mainly relies on statistical methods to identify and measure factors affecting injury severity. For instance, Song et al. (2023) used a random parameters bivariate probability model to analyze UK truck-car collision data from 2017–2019, pinpointing key determinants of driver injury and suggesting safety improvements [11]. Gong et al. (2022) applied a Bayesian multivariate random parameters logit model to U.S. two-vehicle collision datasets from 2010–2018, revealing correlations between influencing factors and injury risk, as well as their heterogeneity and temporal variability [12]. Cerwick et al. (2014) compared mixed logit and latent class models using Iowa multi-vehicle heavy truck crash data from 2007–2012, finding that while the latent class model fit marginally better, the mixed logit model provided more accurate probability predictions [13]. However, these studies focus on analyzing individual accident characteristics, establishing relationships between specific features and severity, such as vehicle speed and collision outcomes. Their discrete analysis of features limits predictive power in real-world complexity. By treating features in isolation rather than considering their interactions, statistical methods fail to capture the full dynamics of accidents [14]. This highlights an urgent need for a more integrated approach to accurately predict PV-PV accident severity, underscoring a crucial research gap [15,16].

In recent years, machine learning (ML) has emerged as a powerful alternative to traditional methods for injury severity analysis [17,18], adept at extracting insights from large, complex, and heterogeneous datasets. Among ML approaches, ensemble machine learning (EML) and deep learning (DL) models are widely applied in accident severity prediction, but with distinct focuses. DL models are mainly used for macroscopic accident analysis, leveraging spatio-temporal data from images, videos, or sequences to predict accident occurrence and severity, thereby assisting traffic management in intervention and emergency response. For instance, Yu et al. (2021) developed a deep spatio-temporal graph convolutional network using Beijing's heterogeneous traffic data, significantly improving accident risk prediction and enabling early warnings and safer route planning [19]. Fares Alhaek et al. (2024) proposed a CNN-BiLSTM-based DL method that outperformed baselines in predicting traffic accident severity using UK city data [20].

Given the structured nature of accident severity datasets, typically in tabular format, EML models excel in handling such data [21]. By integrating multiple base models, EML methods outperform single models in capturing intricate patterns and relationships within data. This superiority is especially evident in accident severity prediction. Popular ensemble models, including random forest, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), categorical boosting (CatBoost), and adaptive boosting (AdaBoost), have been widely applied in the field of traffic accident severity prediction [22–28].

Ensemble machine learning (EML) algorithms boost predictive performance by integrating multiple decision tree models, reducing variance, bias, and overfitting for more accurate predictions across datasets. However, in high-dimensional accident feature spaces, the increased tree depth and node quantity of ensemble models raise complexity, severely undermining interpretability [29]. Thus, the "black-box" nature of EML models, despite their high predictive power, hinders understanding of accident causation. Explainable machine learning techniques like Shapley additive explanations (SHAP) [30] tackle this issue. Based on game-theoretic Shapley values, SHAP quantifies feature contributions and interaction effects. Its additive feature attribution suits tree-based EML models, enabling global and local interpretability [31]. Widely used in ensemble frameworks, SHAP is key for analyzing the link between accident severity and contributing factors, offering a new approach for passenger vehicle collision studies that balances prediction accuracy with in-depth model and feature interpretation.

Current research on accident severity prediction faces several challenges. First, natural accident datasets suffer from severe class imbalance [6,11,12], where fatal accident cases are significantly underrepresented compared to non-injury cases. This imbalance poses challenges for traditional resampling methods in effectively balancing severity levels [32]. Moreover, accident datasets often contain a mix of categorical and numerical features, which limits the applicability of Euclidean distance-based oversampling techniques such as the synthetic minority over-sampling technique (SMOTE) [33]. Thus, addressing data imbalance at the data collection stage is crucial. Second, existing studies primarily rely on empirical heuristics for feature selection [11], lacking rigorous scientific justification. Consequently, the representativeness and rationality of selected features remain in adequately explained.

To address the issue of class imbalance in accident severity datasets, this study first integrates the CRSS and FARS datasets from the National Highway Traffic Safety Administration (NHTSA). This integration reduces the severity imbalance ratio from 175.1 in CRSS and 10.2 in FARS to 5.1 in the combined dataset. Furthermore, a combination of oversampling and undersampling techniques is employed to achieve a balanced distribution of severity categories. For feature selection, traditional methods often rely on Pearson correlation coefficients to eliminate highly correlated and irrelevant variables. However, this approach is not suitable for accident datasets dominated by categorical features, as categorical values represent discrete labels rather than numerical relationships. To enhance the rationality of feature extraction, this study adopts the mRMR algorithm based on mutual information [34]. This algorithm has been widely applied in disease diagnosis [35,36]. Compared with other feature extraction methods, it demonstrates the best performance while requiring the fewest number of features [37]. Subsequently, this study applies ensemble machine learning algorithms—including CatBoost, AdaBoost, XGBoost, random tree, LightGBM, and decision trees—to accident severity prediction. By comparing the predictive performance of different models, the best-performing model is selected for SHAP-based interpretability analysis, facilitating insights into feature importance and improving the explainability of accident severity predictions.

The contributions of this paper can be summarized as follows:

(1) Integrated two distinct accident datasets to address imbalance issues and applied the mRMR algorithm for key feature extraction, reducing redundancy and improving model accuracy across all severity levels.

(2) Proposed a novel micro-level analytical approach for PV-PV accidents, incorporating dual-vehicle factors to simultaneously enhance prediction accuracy and model interpretability, advancing beyond traditional macro-level methods.

(3) Performed a systematic evaluation of state-of-the-art EML models for PV-PV accident severity prediction and integrated the optimal ensemble model with SHAP interpretation, enabling comprehensive feature analysis and addressing a critical research gap in traffic safety analytics.

The remainder of this paper is organized as follows. Section 2 describes the methodologies, encompassing data preprocessing techniques, feature extraction methods, predictive modeling approaches, and explainable ML algorithms. Section 3 presents the dataset's statistical characteristics and feature analysis. Section 4 details the experimental results, including model performance evaluation and critical factor analysis. Section 5 discusses the research implications and study limitations. Finally, Section 6 concludes the paper and proposes future research directions.

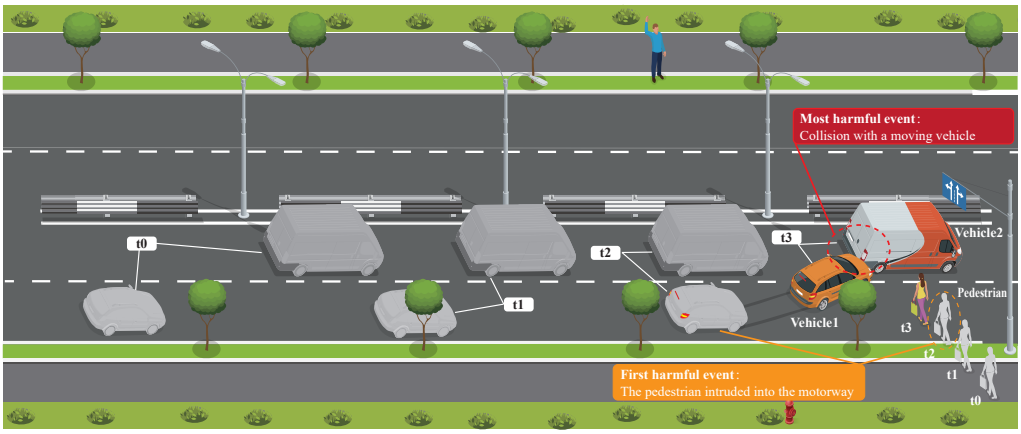

Figure 1. Schematic diagram of a traffic accident scene.

2. Methods

The technical route of this study is shown in Figure 2, which is mainly divided into three parts: the first part is data collection and preprocessing, which aims to provide a structured data set that meets research needs to improve the accuracy and reliability of the model; the second part is model training and tuning, which aims to improve the predictive performance and generalization ability of the model by optimizing model parameters; the third part is to interpret and analyze the characteristics of the optimal model to reveal the key factors behind the model decision and enhance the interpretability and transparency of the model.

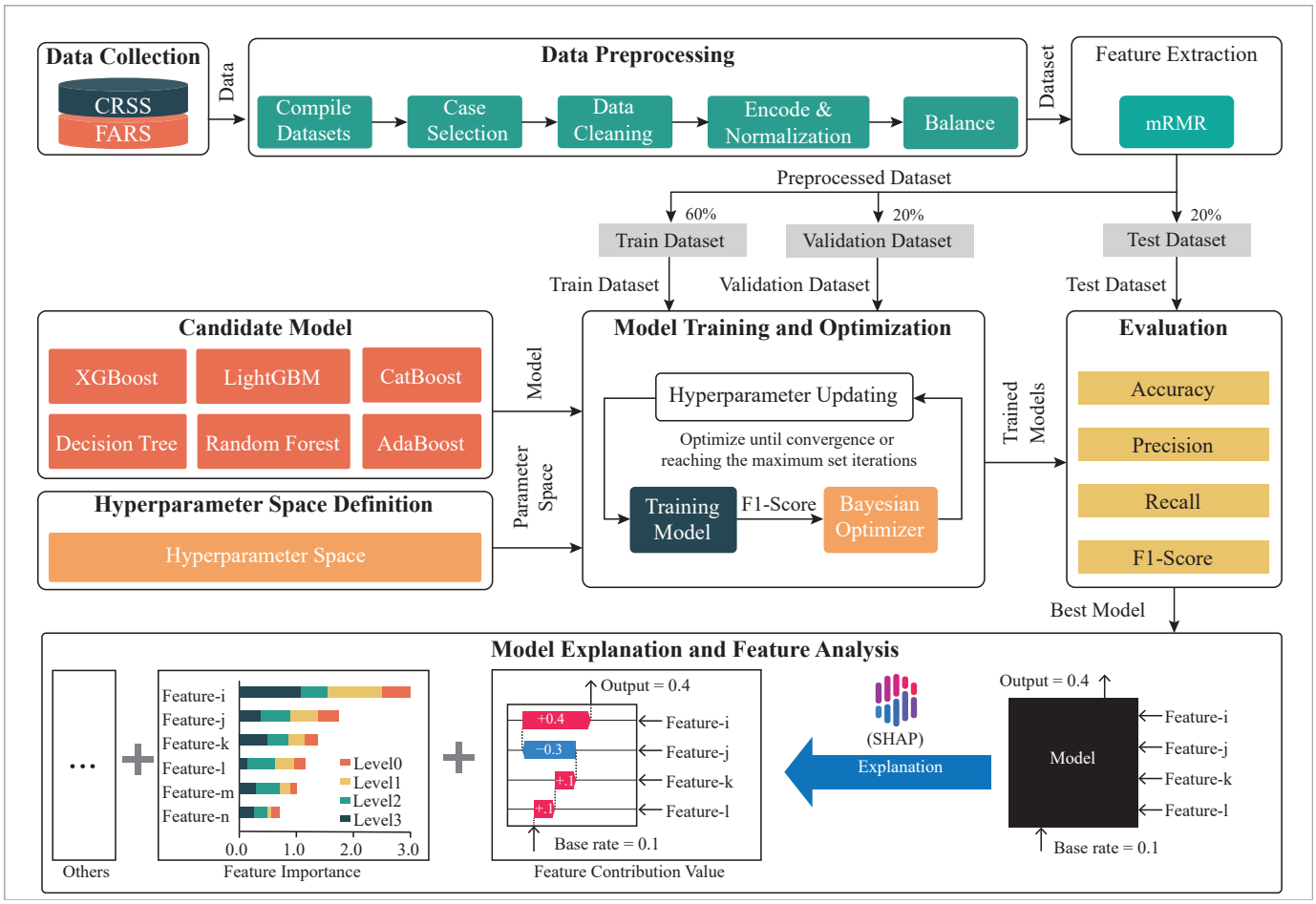

Figure 2. Technical framework diagram of this study.

2.1. Accident Data Collection and Preprocessing

2.1.1. Accident Data Collection

In accident severity prediction research, class imbalance is a common issue, particularly due to the relatively low number of fatal accident samples, which limits models' ability to predict severe accidents. Additionally, these datasets typically contain both categorical and numerical features, making traditional Euclidean distance-based SMOTE algorithms less effective for data balancing. To address dataset imbalance, this study integrates two complementary NHTSA datasets: CRSS and FARS. The FARS dataset, containing comprehensive fatal accident records, effectively compensates for the scarcity of severe cases in CRSS. This integration enhances the representation of high-severity accidents at the data source level, which improves the model's predictive accuracy for severe accident classification.

2.1.2. Dataset Preprocessing

Dataset Compilation

The accident-related data are extracted from multiple relational tables within both CRSS and FARS datasets, including accident records, person information, event sequences, and vehicle damage reports. These heterogeneous data sources are consolidated into a unified structured dataset through a systematic integration process. A composite primary key, comprising Accident ID and Driver ID, is implemented to ensure precise data linkage and maintain referential integrity across all integrated tables.

Case Selection

This study focuses on predicting driver injury severity in PV-PV collisions. Accordingly, vehicle collision cases involving exactly two passenger vehicles with corresponding driver characteristics were extracted from the integrated dataset, ensuring alignment with the research focus.

Data Cleaning

Features with excessive missing values were removed. Features with minimal missing values were imputed using a context-based approach. Cases lacking critical information, such as driver injury severity levels, were excluded to ensure data quality.

Encoding and Normalization

Categorical feature encoding in the original dataset was not directly related to injury severity. Although decision tree-based ensemble models are unaffected by encoding order, it influences the interpretability of SHAP-based feature analysis. Therefore, target encoding was applied, ranking categorical features in ascending order based on their mean injury severity levels. Numerical features were standardized to eliminate scale differences, ensuring model stability and convergence. Minor injury categories (e.g., Possible injury and Suspected minor injury) lacked sufficient distinction. Thus, injury severity levels were reclassified as follows:

No apparent injury → Level0
Possible injury and Suspected minor injury → Level1
Suspected serious injury → Level2
Fatal injury → Level3

Data Balancing

The integration of CRSS and FARS inherently mitigated class imbalance. To further balance the dataset, a combination of random oversampling and undersampling was applied [38]. The average sample size across injury severity levels was used as a reference: oversampling was conducted for underrepresented classes, while undersampling was applied to overrepresented classes. To prevent overfitting and enhance model generalization, categorical features remained unchanged during oversampling, whereas numerical features were perturbed by adding random noise in the range of −5% to +5%. The total dataset size remained constant before and after balancing.

2.2. Feature Selection

Feature selection is a crucial step in accident severity prediction, as redundant and irrelevant features can negatively impact model performance. In this study, the mRMR [34] method was employed for feature selection in accident severity prediction. Compared to traditional correlation-based feature selection methods (e.g., Pearson correlation), mRMR is more effective for datasets with categorical features, as it does not assume linear relationships. The mRMR algorithm is a widely used method for selecting an optimal subset of features by balancing two key objectives:

1. Maximum Relevance: The selected features should have the highest relevance to the target variable.

2. Minimum Redundancy: The selected features should be minimally correlated with each other to reduce redundancy.

2.2.1. Mutual Information (MI) Definition

The mutual information $I(X;Y)$ quantifies the dependency between two variables X and Y, and is defined as follows:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ represents the joint probability distribution, and $p(x)$ and $p(y)$ are the marginal distributions of X and Y, respectively.

2.2.2. Maximum Relevance Criterion

The relevance of a feature set S with respect to the target variable c is computed as the average mutual information between each feature X_i and c:

$$\mathcal{R}(S,c) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i;c) \tag{2}$$

where $|S|$ is the number of selected features. The goal is to maximize this relevance measure.

2.2.3. Minimum Redundancy Criterion

To ensure that the selected features are not highly correlated with each other, the redundancy term is defined as the average mutual information among selected features:

$$\mathcal{D}(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i; X_j) \tag{3}$$

The goal is to minimize this redundancy measure.

2.2.4. mRMR Optimization Objective

The final objective of mRMR is to simultaneously maximize relevance and minimize redundancy, which can be formulated as follows:

$$\max[\mathcal{R}(S,c) - \mathcal{D}(S)] \tag{4}$$

This ensures that the selected feature subset retains the most informative variables while reducing redundancy.

2.3. Prediction Model

Given the superior EML models in accident severity prediction, this study adopts them as the primary predictive approach. Specifically, five top-performing EML models—random forest, LightGBM, XGBoost, CatBoost, and AdaBoost—are selected, with the decision tree model serving as the baseline. A brief introduction to the principles and core formulas of each model is provided below.

2.3.1. Decision Tree

A decision tree is a tree-structured model for classification and regression [39], which recursively partitions the feature space to construct decision rules. Its core idea is to select the optimal splitting feature based on information Gain or Gini impurity.

Information Gain measures the improvement in dataset purity before and after splitting, calculated as follows:

$$\text{Information Gain} = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} H(D_v), \tag{5}$$

where $H(D)$ is the entropy of dataset D, and D_v is the subset where feature A takes the value v.

Gini Impurity measures the impurity of a dataset, calculated as follows:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{k} p_i^2, \tag{6}$$

where p_i is the proportion of samples of the $i-th$ class in dataset D.

2.3.2. Random Forest

Random forest is an ensemble learning method based on decision trees, designed to enhance generalization by constructing multiple decision trees and aggregating their predictions, proposed by Leo Breiman in 2001 [40]. Its core principle involves building diverse decision trees using bootstrap sampling and random feature selection. For classification tasks, random forest determines the final prediction through majority voting:

$$\hat{y} = \text{mode}(\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}), \tag{7}$$

where $\hat{y}$ is the prediction of the $i-th$ decision tree, and T is the number of trees.

2.3.3. LightGBM

LightGBM, proposed by Ke et al. in 2017 [41], is an advanced implementation of gradient boosting trees renowned for its efficiency. It accelerates training through two key techniques: the histogram-based decision tree algorithm, which discretizes features to reduce split-point evaluation complexity, and gradient-based one-side sampling (GOSS), which selectively samples data based on gradient magnitudes to focus on hard-to-learn instances. The objective function of LightGBM consists of a loss function and a regularization term, striking a balance between accurate prediction and overfitting prevention. The definition of its loss function is as follows:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{8}$$

where $\mathcal{L}$ is the loss function, and $\Omega(f_k)$ is the complexity regularization term for the k-th tree.

2.3.4. XGBoost

XGBoost, proposed by Tianqi Chen et al. in 2016 [42], is a highly efficient and powerful algorithm grounded in gradient boosting trees. It enhances the model's accuracy remarkably by leveraging a second-order Taylor expansion to approximate the objective function. This sophisticated approximation technique enables a more precise optimization process. Additionally, XGBoost incorporates regularization terms into its framework. These regularization terms play a crucial role in preventing the model from overfitting, ensuring that it generalizes well to new, unseen data. The objective function of XGBoost is formulated as follows:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \left(\gamma T_k + \frac{1}{2} \lambda \|w_k\|^2 \right), \tag{9}$$

where T_k is the number of leaf nodes in the k-th tree, w_k represents the vector of leaf node weights in the k-th tree, and γ and λ are regularization coefficients. The term $\|w_k\|^2$ is the L2 norm (squared) of the leaf node weights, serving as a regularization to penalize large weights and reduce model complexity, thus preventing overfitting.

2.3.5. CatBoost

CatBoost, introduced by Liudmila Prokhorenkova et al. in 2018 [43], is a sophisticated gradient boosting algorithm that stands out for its specialized optimization in handling categorical features. It employs a unique approach to process these features. Specifically, it utilizes target statistics for encoding categorical variables, which effectively captures the relationship between the categorical values and the target variable. Additionally, CatBoost incorporates ordered boosting as a safeguard mechanism to prevent target leakage, a common pitfall in machine learning models that can lead to overly optimistic performance estimates. For a categorical feature x, CatBoost encodes it using target statistics in the following manner:

$$\hat{x}_i = \frac{\sum_{j<i} \mathbb{I}_{x_j=x_i} \cdot y_j + a \cdot p}{\sum_{j<i} \mathbb{I}_{x_j=x_i} + a}, \tag{10}$$

where a is a smoothing parameter to control the impact of prior knowledge, and p is the global mean target value. The term $\mathbb{I}_{x_j=x_i}$ is an indicator function that equals 1 if $x_j = x_i$, and 0 otherwise.

2.3.6. AdaBoost

AdaBoost, first proposed by Yoav Freund and Robert E. Schapire in 1997 [44], is a prominent ensemble learning method that operates on the principle of weighted voting. This algorithm works by iteratively adjusting the weights assigned to individual samples in the dataset. Through these iterative weight adjustments, AdaBoost trains a sequence of weak classifiers, each of which is focused on different aspects of the data based on the adjusted sample weights. Eventually, it aggregates the predictions of these weak classifiers to form a more accurate and robust final prediction.In the t-th iteration of the AdaBoost process, the sample weights are updated in the following way:

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp(-\alpha_t y_i h_t(x_i)), \tag{11}$$

where $w_i^{(t+1)}$ is the weight of sample i in the $t+1$-th iteration, $w_i^{(t)}$ is the weight of sample i in the t-th iteration, α_t is the weight of the t-th weak classifier, y_i is the true label of sample i (with values ± 1), and $h_t(x_i)$ is the prediction of the t-th weak classifier for sample x_i. The weight α_t of the weak classifier is computed as

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \tag{12}$$

where ϵ_t represents the error rate of the weak classifier in the t-th iteration.

2.4. Model Interpretation and Feature Analysis

SHAP is a unified framework for interpreting machine learning models by assigning each feature a contribution value based on its impact on the model's prediction, The theory of SHAP can be found in this article [30]. It is rooted in cooperative game theory and Shapley values, ensuring fair and consistent feature attribution. SHAP offers several key advantages: Additivity : SHAP values decompose the model's prediction into the sum of contributions from each feature, providing a clear and interpretable explanation. Consistency: If a feature's impact on the model's prediction increases, its SHAP value will also increase, ensuring reliable interpretations. Efficiency: For tree-based models, Tree SHAP computes SHAP values in polynomial time, making it significantly faster than traditional methods. Interpretability: SHAP values provide both local (individual predictions) and global (overall feature importance) interpretability.

2.4.1. Shapley Values

Shapley values distribute the contribution of each feature fairly by considering all possible feature subsets. For a model f and feature i, the Shapley value ϕ_i is computed as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (v(S \cup \{i\}) - v(S)), \tag{13}$$

where F is the set of all features, $|F|$ is the total number of features, S is a subset of features excluding feature i, $|S|$ is the number of features in subset S, $v(S)$ and $v(S \cup \{i\})$ are the model outputs for feature subsets S and $S \cup \{i\}$, respectively, and the weight $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ represents the probability of the feature subset S in all possible permutations.

2.4.2. Additivity Property

SHAP values satisfy the additivity property, meaning the model's prediction can be decomposed into the sum of the SHAP values of all features plus a baseline value:

$$f(x) = E(f(x)) + \sum_{i=1}^{M} \phi_i, \tag{14}$$

where $E(f(x))$ is the baseline value (typically the mean prediction over the training data), and M is the total number of features. This property ensures that SHAP values provide a clear and interpretable explanation of the model's output.

2.4.3. Consistency

SHAP values are consistent: if a feature's impact on the model's prediction increases, its SHAP value will also increase. This ensures reliable interpretations across different models and datasets.

2.4.4. Efficiency for Tree Models (Tree SHAP)

For tree-based models (e.g., decision trees, random forests, gradient boosting), Tree SHAP computes SHAP values efficiently in polynomial time by leveraging the tree structure. The SHAP value for feature i in a single tree is as follows:

$$\phi_i = \sum_{v \in \text{Paths}(i)} (\text{PathProb}(v) \cdot (\text{Value}(v \cup \{i\}) - \text{Value}(v))), \tag{15}$$

where Paths(i) is the set of paths where feature i is used in a split, PathProb(v) is the probability of path v, and Value(v) is the prediction value at the end of path v.

3. Data Description

This study uses crash data from the CRSS and FARS databases maintained by the NHTSA from 2016 to 2022. The CRSS provides a nationally representative sample of police-reported crashes, including those resulting in property damage, injuries, or fatalities, while the FARS records fatal crashes where at least one fatality occurs within 30 days of the accident. The two databases collectively include 594,773 crash cases, covering single-vehicle, two-vehicle, and multi-vehicle collisions, with PV-PV collisions being the most frequent, accounting for 225,461 cases. A detailed breakdown of the raw data is presented in Table 1. More information on these databases can be found at NHTSA CRSS (https://www.nhtsa.gov/crash-data-systems/fatality-analysis-reporting-system, accessed on 25 December 2024) and NHTSA FARS (https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system, accessed on 25 December 2024).

Table 1. Accident data statistics from the CRSS and FARS databases (2016–2022).

Database	Participants	Case Number		Driver Numbers for Different Accident Vehicles					
		Total	PV-PV	Total	PV	Motorcycle	Truck	Bus	Others
CRSS	1	121,288	-	121,288	107,169	8585	3768	418	1348
	2	204,916	175,910	409,832	379,053	9341	14,218	1590	5630
	>2	26,938	-	87,755	83,664	883	2226	122	860
FARS	1	134,982	-	134,982	111,851	14,360	5462	370	2939
	2	86,718	49,551	173,436	132,888	18,678	17,512	580	3778
	>2	19,931	-	69,080	57,565	3685	6361	177	1292
CRSS + FARS	-	594,773	225,461	996,373	872,190	55,532	49,547	3257	15,847

3.1. Descriptive Statistics of Accident Data

To ensure data completeness and reliability, accident cases missing critical feature information were removed. Since this study focuses on predicting the injury severity of the subject vehicle's driver, certain attributes of the other party in the collision were considered less relevant. For instance, the seat belt usage of the other driver is unlikely to significantly impact the subject driver's injury severity. Consequently, some cases were retained despite incomplete information on the other driver, resulting in instances where only one driver in a collision was analyzed.

After preprocessing, the final dataset comprises 116,998 accident cases and 182,296 drivers, ensuring that each retained case contains sufficient information for injury severity prediction. A detailed summary of the dataset after preprocessing is presented in Tables 2–4 and Figures 3 and 4.

Table 2. Statistical description of accident environmental characteristics.

Variables	(%)	Injury Severity (%)				Variables	(%)	Injury Severity (%)			
		0	1	2	3			0	1	2	3
Database						Divided, Unprotected (2)	27.7	50.2	31.2	6.6	12
CRSS	63.3	68.7	27.7	3.2	0.4	Divided, Protected (3)	13.4	58	22.8	5.6	13.6
FARS	36.7	5	25.5	18.4	51.1	One-Way (4)	7.4	72.5	19.9	3.1	4.6
CRSS+FARS	100	45.3	26.9	8.8	19	Roadway Profile (TWY_PRFL)					
Location at rural or urban (RUR_URBN)						Level (1)	85.9	47.5	26.9	8.1	17.5
Rural (1)	67.9	53	25.6	7.7	13.7	Grade (2)	7.2	36.1	27.7	11.4	24.8
Urban (2)	32.1	29.1	29.5	11.3	30.2	Uphill (3)	2.4	28.8	25.7	15	30.5
Light condition (LGT_COND)						Downhill (4)	2.7	31.2	24.5	13.6	30.7
Daylight (1)	69.8	50.6	26.8	7.2	15.4	Hillcrest (5)	1.5	23.2	27.3	15.2	34.3
Dawn (2)	1.6	34	25	12	29.1	Sag (6)	0.2	26.6	31.2	11.6	30.7
Dusk (3)	2.6	47.6	24.8	9.3	18.3	Roadway Alignment (TWY_ALGN)					
Dark-Lighted (4)	15	41.1	30.5	9.4	19	Straight (1)	91.6	47.2	27.4	8.1	17.4
Dark-Not Lighted (5)	11	19.2	22.9	17.5	40.4	Curve Right (2)	4.6	31.1	19.1	14.9	35
Atmospheric Conditions (ATM_COND)						Curve Left (3)	3.8	18	24	19.5	38.5
Clear (1)	74.4	46.3	26.6	8.6	18.5	Type of Intersection (TWY_INTT)					
Cloudy (2)	14.8	43.1	27.6	9.3	20.1	Four-Way Intersection (1)	33.7	45.9	33.3	6.8	14
Rain (3)	8.6	44.5	27.6	8.6	19.2	T-Intersection (2)	14.4	46.2	34.1	6.2	13.5
Snow (4)	1.2	41.3	26.8	11.1	20.9	Roundabout (3)	0.2	84.4	13.4	0.6	1.6
Fog (5)	0.6	17.7	24.6	17.6	40.1	Five-Point, or more (4)	0.2	55.5	29.7	5.9	8.9
Other (6)	0.3	22	26.4	17.3	34.4	Y-Intersection (5)	0	47.2	33.3	2.8	16.7
Trafficway Description (TWY_TYPE)						Other Type (6)	0.4	29.8	29.8	11.3	29.1
Not Divided (1)	51.5	35.6	26.6	11.6	26.2	Non-intersection (7)	51.1	44.7	20.6	10.9	23.9

Table 3. Statistical description of collision and pre-collision characteristics.

Variables	(%)	Injury Severity (%)				Variables	(%)	Injury Severity (%)			
		0	1	2	3			0	1	2	3
Attempted Avoidance Maneuver (DRV_MANU)						Intersect Paths (Straight Path) (7)	14.3	26.1	40.6	9.9	23.5
No Avoidance Maneuver (1)	23.7	62.6	17.6	5.7	14.2	Opposite Direction (Angle Sideswipe) (8)	4.2	12.4	28.4	16.5	42.8
Braking (2)	3.5	51.6	32.8	7.5	8.2	Head-On (9)	14.7	2.2	16.1	27.1	54.6
Unknown (3)	68	40.7	29.5	9.3	20.5	Other Types (10)	0.3	8.2	23.4	10.9	57.5
Accelerating (4)	0.1	39.5	29.3	10.9	20.4	Travel Speed of This Vehicle (TRV_SPD1)					
Braking and Steering (5)	1	24.7	35.9	18.3	21	0 mph	16.2	84.6	11.3	1.4	2.7
Releasing Brakes (6)	0	22.9	41.7	2.1	33.3	0–20 mph	19	82.2	8.8	2.2	6.8
Braking and Unknown Steering Direction (7)	0.2	13.4	42.9	18.8	25	21–40 mph	27.1	38.1	41.5	7.2	13.2
Steering (8)	3.5	20.9	28.6	19	31.5	41–60 mph	28.2	15.5	34.2	16.2	34.1
Accelerating and Steering (9)	0	26.1	19.5	16.3	38.2	>60 mph	9.4	13.7	25.9	17.4	43.1
Pre-Impact Stability (PIM_STAB)						Travel Speed of Other Vehicle (TRV_SPD2)					
Tracking (1)	92.6	48	27.1	8.2	16.7	0 mph	10	73	21.2	2.4	3.4
Skidding Longitudinally (2)	1.4	26.5	31.1	14.5	27.9	0–20 mph	18.9	79.3	15.9	2.4	2.3
Skidding Laterally (3)	1.1	5.6	8.5	14.9	70.9	21–40 mph	30	46.8	35.8	6.4	11
Other (4)	4.9	10.2	24.9	17.2	47.7	41–60 mph	30.9	23.1	28.4	14.7	33.8
Crash Type (ACC_TYPE)						>60 mph	10.1	18.2	21.4	16.2	44.2
Rear End (1)	32.1	74.9	18.6	2.1	4.4	Damage Area Count (DMG_ARCT)					
Same Direction (Angle, Sideswipe) (2)	7	79	11	1.8	8.2	1	55.9	52.9	26.1	6.9	14
Miss Control (3)	1.2	90.1	6.5	1.1	2.4	2–5	35.2	44.1	30.9	8.4	16.6
Turn Into Path (4)	12.2	42.1	37.2	5.9	14.8	6–9	4.7	4.7	20.1	22.2	53
Turn Across Path (5)	13.5	36	43.8	7.9	12.2	10–12	4.2	1.1	10.4	21.9	66.6
Opposite direction (Forward Impact) (6)	0.4	1.7	34.7	21.9	41.7						

After applying the mRMR algorithm, the feature set was reduced from 65 to 33, achieving a balance between relevance and redundancy. At this selection point, feature redundancy is minimized, and further increasing the number of features does not yield noticeable improvements in predictive accuracy. The retained features encompass key accident-related factors, ensuring both robust model performance and meaningful interpretability in accident analysis. The correlation among the selected features, assessed using Cramér's V [45], is visualized in Figure 5. This refinement enhances model performance by ensuring the inclusion of only the most relevant and non-redundant features, which are categorized into three groups:

(1) Environmental Features (seven features): These features are related to weather conditions and accident location. Table 2 presents all seven environmental features.

(2) Collision and Pre-Collision Features (eight features): These features describe the driver's evasive actions before impact and the collision dynamics. Table 3 lists six of these features, while the initial impact areas of both vehicles (IMP_ARE1 and IMP_ARE2) are shown in Figure 4.

(3) Participant Features (18 features): These features include driver and vehicle-related characteristics for both parties involved in the collision. Table 4 presents 14 of these features, the age distributions of the two drivers (DRV_AGE1 and DRV_AGE2) are illustrated in Figure 4, and the vehicle brand statistics of the involved vehicles are shown in Figure 3.

This selection process ensures that the retained features contribute meaningfully to accident severity prediction while minimizing redundancy and noise in the dataset.

Table 4. Statistical description of vehicle and driver characteristics for both parties in the collision.

Variables	(%)	Injury Severity (%)				Variables	(%)	Injury Severity (%)			
		0	1	2	3			0	1	2	3
The gender of this driver (DRV_SEX1)						<1	4.7	54.8	21.5	6.5	17.2
Male (1)	55.9	44.7	23.7	9.5	22	1–10	55.9	49.3	24.7	7.6	18.4
Female (2)	44.1	46.2	30.8	7.9	15.2	11–20	34.1	39.6	29.8	10.5	20.1
Drug involvement of this driver (DRV_DUG1)						>20	5.3	32.4	35.3	12.7	19.7
No (1)	96.6	46.8	27.2	8.3	17.7	The Weight of This Vehicle (VEH_WGT1)					
Yes (2)	3.4	3.9	17.1	22.5	56.6	1700–2700 lbs.	8.5	38.2	21.3	8.2	32.5
Alcohol Test Result of this driver (DRV_ALC1)						2701–3700 lbs.	51.9	44.8	25.9	8.4	21
0 mg/dL	93.7	48	27.4	8.2	16.5	3701–4700 lbs.	26.3	48.3	28.3	9.2	14.2
1–80 mg/dL	1	2.7	14.8	15.9	66.5	4701–5700 lbs.	11.2	47.4	30.6	10.1	11.9
>80 mg/dL	5.4	7.7	18.7	19	54.6	>5700 lbs.	2.1	40.7	34.6	10.9	13.9
The Other Driver's Gende (DRV_SEX2)						The Weight of Other Vehicle (VEH_WGT2)					
Male (1)	58	41	26.6	9.8	22.7	1700–2700 lbs.	8	47.77	35.4	9.45	7.38
Female (2)	42	51.4	27.2	7.5	13.9	2701–3700 lbs.	50.1	48.77	28.8	8.91	13.52
Drug Involvement of Other Driver (DRV_DUG2)						3701–4700 lbs.	27	44.53	24.96	8.47	22.04
No (1)	96.6	46.7	26.9	8.3	18.2	4701–5700 lbs.	12.4	35.65	19.91	8.51	35.93
Yes (2)	3.4	7.8	25.9	22.9	43.5	>5700 lbs.	2.5	25.54	14.95	9.9	49.61
Alcohol Test Result of other driver (DRV_ALC2)						Vehicle Body Type of This Vehicle (VEH_BDY1)					
0 mg/dL	93.7	47.7	26.8	8	17.6	Pickup (1)	15.7	45	27.7	10.4	16.9
1–80 mg/dL	0.9	6.6	29	23	42	SUV (2)	30	50.2	28.7	8.4	12.7
>80 mg/dL	5.4	11.6	27.2	21.2	40.1	Minivan (3)	2.1	69.1	27.8	2.8	0.3
Seat Belt Type and Usage Status (SAF_REST)						Cargo Van (4)	0.3	80.8	17.5	1.5	0.2
Not Used (1)	12.7	9.9	14.8	16.9	58.4	VAN (5)	1.9	11	30.2	18.8	40
Two–Point (2)	0.9	41.3	35.2	8.4	15.1	Sedan (6)	40.7	43.2	25.8	8.5	22.5
Three–Point (3)	84.7	50.9	28.4	7.5	13.2	Coupe (7)	3.7	38.6	22.9	9.4	29.1
Others (4)	1.7	34	33	16	17	Wagon (8)	0.7	41.8	23	6.6	28.7
Air bag deployment (SAF_ARBG)						Hatchback (9)	4.1	41.1	24.4	7.9	26.6
Not Deployed (1)	58	71.2	17.8	3	8.1	Convertible (10)	0.8	39.1	21.8	8.5	30.6
Curtain (2)	0.1	10.2	30.5	15.3	44.1	Vehicle Body Type of Other Vehicle (VEH_BDY2)					
Side (3)	1.2	19	39	8.2	34	Pickup (1)	17.3	34.9	20.8	9.2	35.1
Front (4)	1.2	18.8	38.7	8.2	34.4	SUV (2)	29.8	48.2	25	7.8	19
Combined (5)	16	10.1	39	16.7	34.2	Minivan (3)	2.1	67.1	28.9	3.6	0.4
Other (6)	24.4	8.7	39.8	17.4	34	Cargo Van (4)	0.4	68.3	27.6	3.9	0.2
The Age of This Vehicle (VEH_AGE1)						VAN (5)	2	9.2	23	18	49.8
<1	5	53.7	30.3	6.9	9.1	Sedan (6)	39.5	48.3	29.9	9.1	12.8
1–10	56.8	49.7	29.3	7.8	13.1	Coupe (7)	3.6	42.1	31.4	10.5	16
11–20	32.7	39.5	23.6	10.4	26.6	Wagon (8)	0.7	46.1	31.7	10	12.2
>20	5.4	26.7	17.3	12.1	43.9	Hatchback (9)	3.8	48.8	30.9	8.6	11.7
The Age of Other Vehicle (VEH_AGE2)						Convertible (10)	0.8	41.9	34.3	10.3	13.6

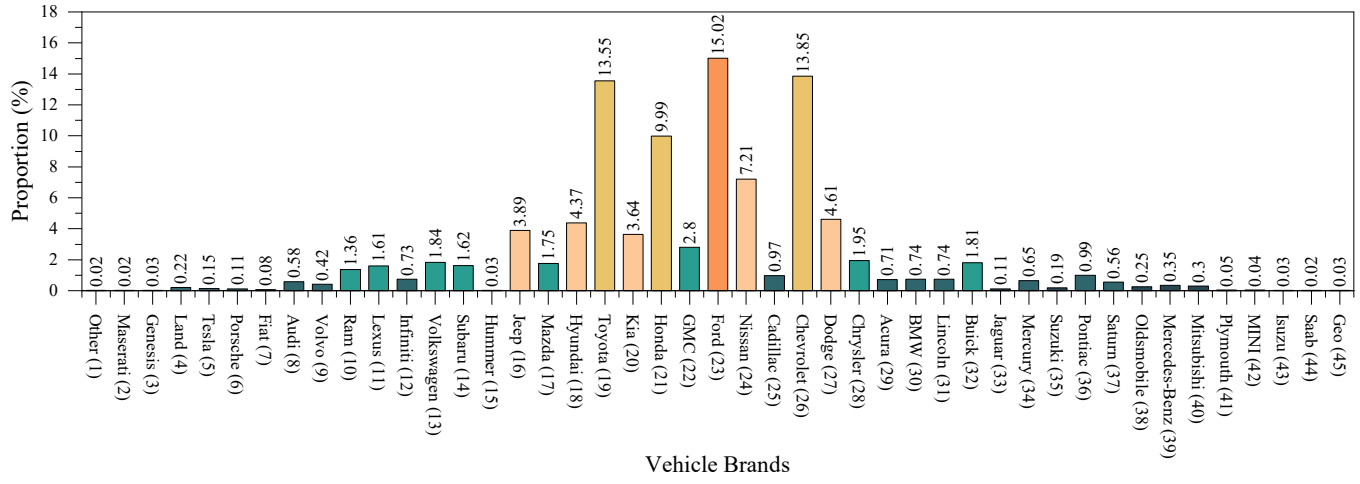

Figure 3. Distribution of vehicle brands (VEH_MKE) in the dataset.

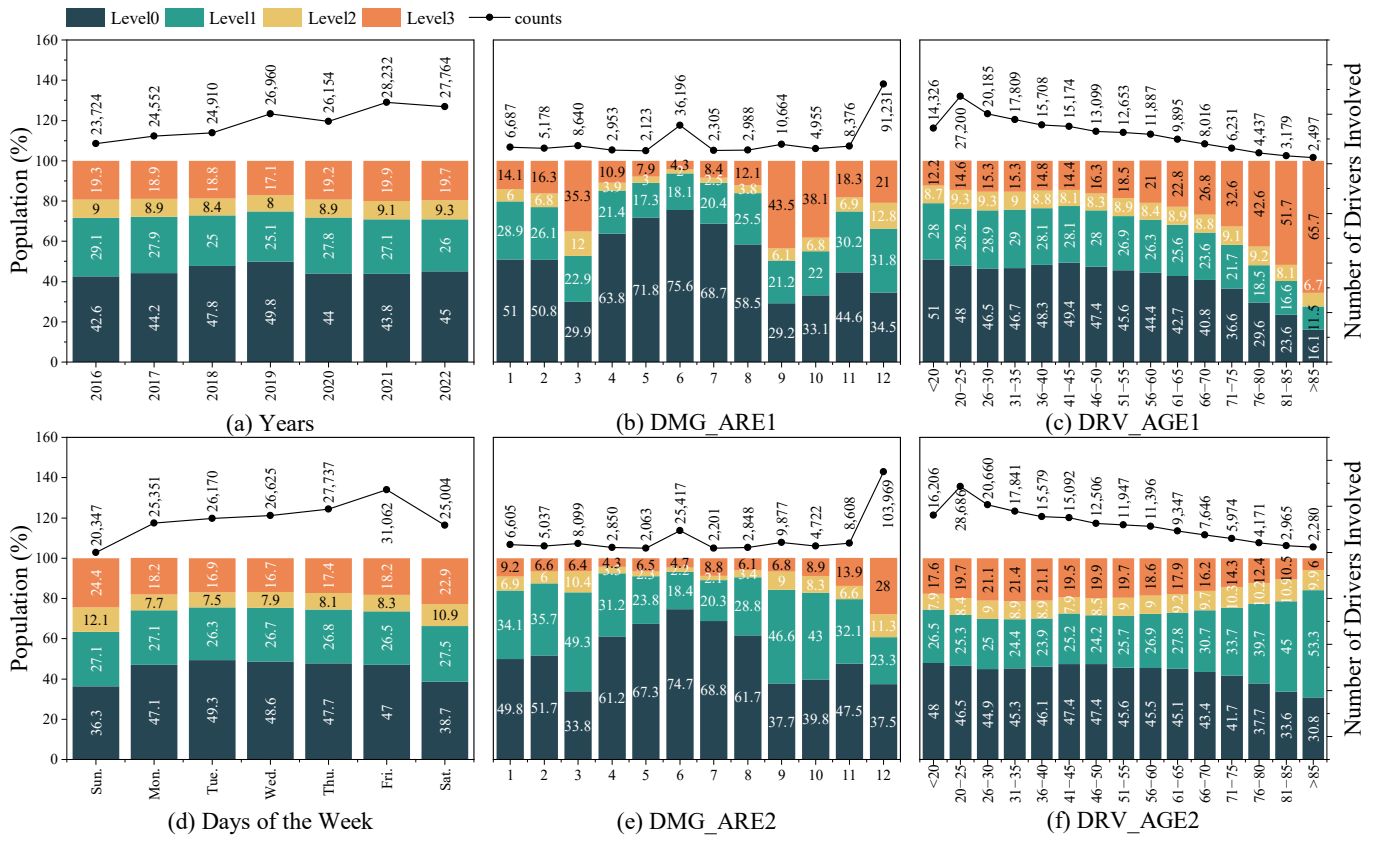

Figure 4. Stacked chart of collision characteristics and severity. (**a**) Annual distribution. (**b**) Initial impact area of this vehicle (IMP_ARE1). (**c**) Driver age of this vehicle (DRV_AGE1). (**d**) Days of the week. (**e**) Initial impact area of the other vehicle (IMP_ARE2). (**f**) Driver age of the other vehicle (DRV_AGE2).

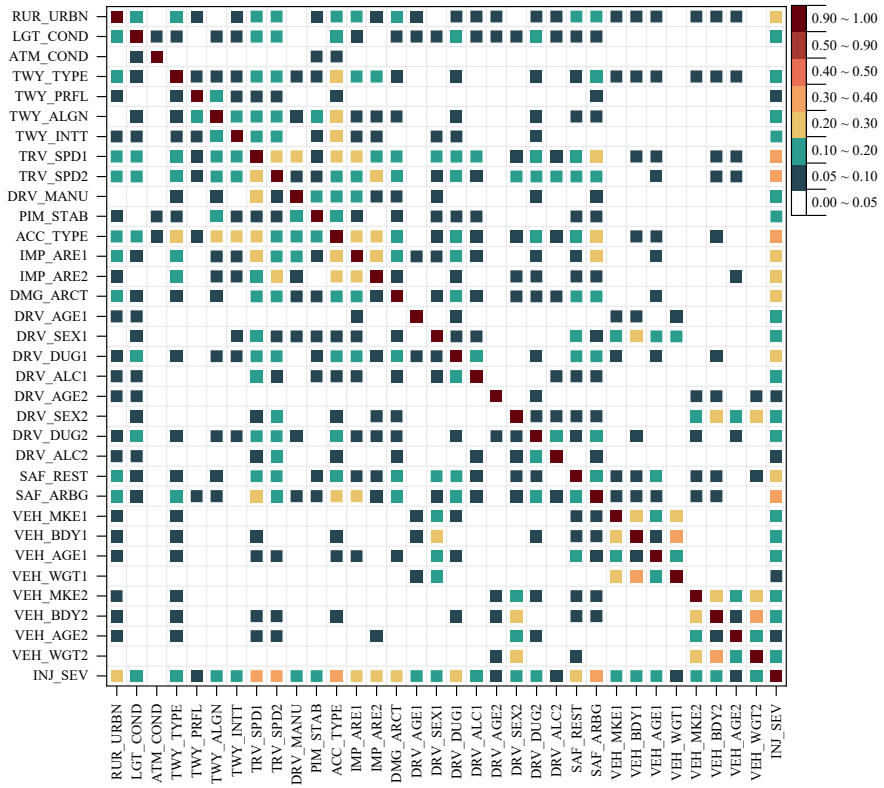

Figure 5. Cramér's V correlation matrix of selected accident features.

3.2. Feature Statistical Description

Building on the statistical analysis of environmental factors, accidents most frequently occur during daytime, under clear weather, on straight and level roads, and on two-way undivided roadways outside intersections—scenarios where vehicles spend most of their travel time. While crash frequency is high due to greater exposure, severe injuries are less common, as good visibility, well-maintained pavement, and predictable traffic flow enhance safety. However, two-way undivided roads stand out as an exception, exhibiting a relatively higher fatality risk despite their prevalence, likely due to the increased likelihood of high-speed head-on collisions.

Shifting focus to collision dynamics, pre-collision features play a crucial role in accident severity. Braking upon detecting danger is associated with a lower risk of fatal injuries (8%), whereas vehicle instability before impact—such as skidding laterally—correlates with a substantially higher severity risk (70.9%), likely due to factors such as slippery roads, excessive speed, and driver errors. Accident type (ACC_TYPE) also influences injury outcomes: rear-end collisions, though the most frequent, are linked to a relatively lower fatal injury risk (4.4%), while head-on collisions present the highest risk due to greater collision energy. Impact location further affects injury severity, with side impacts at the 9, 10, and 3 o'clock positions exhibiting significantly higher fatal injury risks (43.5%, 38.1%, and 35.3%, respectively), likely due to limited energy absorption and intrusion into the driver's space.

It is important to note that since this study integrates two datasets with different class distributions—one covering all accident types and the other focusing on high-severity cases—the reported percentages indicate relative risk levels rather than absolute probabilities. Being struck by the front end (12 o'clock position) of the opposing vehicle poses the highest risk (28%). Similarly, higher pre-crash speeds (TRV_SPD1, TRV_SPD2) are strongly correlated with increased injury severity, while rollovers and loss-of-control crashes—often involving multiple impacts with objects such as the ground, trees, or guardrails—further elevate the risk of severe injuries.

An analysis of driver and vehicle characteristics identifies additional factors influencing the severity of the injury. Male drivers exhibit a higher risk of fatal injury than female drivers, with Cramér's V [45] analysis revealing significant associations between gender and vehicle attributes, including VEH_AGE1, VEH_BDY1, VEH_MKE1, and TRV_SPD. Furthermore, impairment due to alcohol consumption or drug influence is a critical factor associated with increased injury severity.

Finally, protective measures and vehicle attributes play a key role in reducing injury severity. Seat belt use and airbag deployment significantly lower injury severity, with non-deployed airbags typically seen in low-energy collisions that result in minor injuries. Heavier vehicles generally offer better protection due to lower collision accelerations, with vehicle weight also linked to structural attributes like height, width, and length that affect crash safety. Additionally, collisions involving minivans or cargo vans tend to result in less severe injuries, possibly due to factors such as professional drivers and lower urban speeds.

4. Experimental Results

To predict accident severity, five EML models and a decision tree model were employed, treating the prediction as a multi-class classification task. Model performance was evaluated using multiple metrics, as shown in Figure 6. Since accuracy alone does not provide a complete assessment, particularly in imbalanced classification problems, precision, recall, and F1-score were also used to ensure a more comprehensive evaluation.

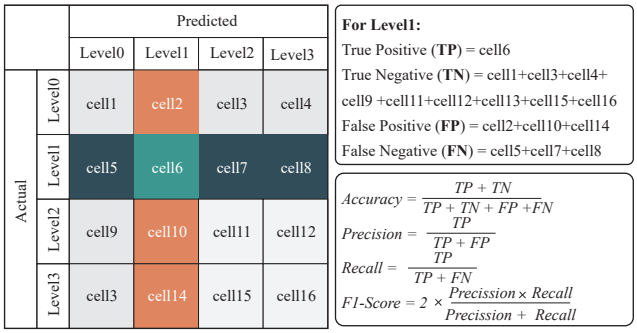

Figure 6. Performance Metrics for Multiclass Classification Models.

Despite merging the CRSS and FARS databases, class imbalance persisted, with severity levels distributed as follows: Level 0 (45.3%), Level 1 (26.9%), Level 2 (8.8%), and Level 3 (19%). To mitigate this, random oversampling and undersampling techniques were applied, using the average class size as a benchmark. Classes exceeding the benchmark were downsampled, while those below it were upsampled. For synthetic data generation, categorical features remained unchanged, while numerical features were modified by adding random noise within ±10% of the original values to enhance data variability while preserving statistical integrity.

Bayesian optimization was employed to determine the optimal hyperparameters for each model, offering advantages such as reduced computational cost and efficient exploration of the hyperparameter space. Model performance was evaluated both before and after data balancing, with hyperparameters tuned separately for each scenario. The dataset was split into 60% training, 20% validation, and 20% testing sets.

Following data balancing, all models demonstrated significant performance improvements. As shown in Table 5, prior to balancing, XGBoost achieved the highest accuracy (79.52%) and precision (71.8%), while CatBoost excelled in recall (68.6%) and F1-score (68.69%). After balancing, XGBoost outperformed all models across all four metrics (accuracy = 84.9%, precision = 84.85%, recall = 84.9%, F1-score = 84.87%), followed by random forest (accuracy = 83.74%, precision = 83.63%, recall = 83.77%, F1-score = 83.66%). However, decision tree exhibited a slight accuracy decrease of 0.73% after balancing, despite notable improvements in precision, recall, and F1-score. This suggests that the model initially favored predicting the majority class in the imbalanced dataset. The redistribution of class instances reduced this bias, enhancing its ability to classify minority classes correctly, while causing a slight accuracy drop.

4.1. Accident Feature Analysis

Based on the model performance evaluation, XGBoost demonstrated superior performance across all metrics and was therefore selected as the target model for interpretability analysis in this study. The SHAP method, based on Shapley values (as detailed in Section 2), was employed to systematically interpret its decision-making mechanisms.

Global SHAP analysis quantitatively assessed the contributions of individual features to predictions across the four accident severity classes. Figure 7a–d illustrates the SHAP value distributions for different severity categories. The horizontal axis represents the range of SHAP values, indicating feature influence intensity (positive values support the prediction of the corresponding severity class, while negative values indicate opposing effects). The vertical axis lists feature variables, with scatter point colors representing feature magnitudes (for numerical variables) or encoded categorical labels (for nominal variables)

Table 5. Statistical performance of accident severity prediction models.

Models	Injury Severity	Imbalanced Data (%)				Balanced Data (%)			
		Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
XGBoost	Level 0	87.93	90.81	89.34		85.19	87.19	86.17	
	Level 1	72.17	74.02	73.08		79.12	76.99	78.04	
	Level 2	51.55	20.95	29.79		86.66	85.88	86.27	
	Level 3	75.57	88.07	81.34		88.42	89.55	88.98	
	Average	71.80	68.46	68.39	79.52	84.85	84.90	84.87	84.90
Random Forest	Level 0	87.85	90.07	88.95		84.69	86.85	85.76	
	Level 1	71.01	74.28	72.61		79.74	74.47	77.01	
	Leve l2	54.96	15.33	23.97		86.00	87.88	86.93	
	Level 3	72.74	88.33	79.78		88.11	89.75	88.92	
	Average	71.64	67.01	66.33	78.81	84.63	84.74	84.66	83.74
CatBoost	Level 0	88.16	90.20	89.17		84.38	86.43	85.39	
	Level 1	71.60	74.19	72.87		75.83	75.05	75.44	
	Level 2	49.39	23.59	**31.93**		81.07	79.31	80.18	
	Level 3	75.82	86.42	80.77		84.68	85.38	85.03	
	Average	71.24	68.60	68.69	79.22	81.49	81.54	81.51	81.52
LightGBM	Level 0	87.88	90.28	89.06		83.87	86.03	84.93	
	Level 1	71.18	72.30	71.74		73.84	74.27	74.06	
	Level 2	42.18	24.21	30.76		77.72	73.80	75.71	
	Level 3	75.32	83.86	79.36		81.04	82.67	81.85	
	Average	69.14	67.66	67.73	78.31	79.12	79.19	79.14	79.15
AdaBoost	Level 0	88.81	84.61	86.66		85.03	77.86	81.29	
	Level 1	63.38	77.87	69.88		69.39	78.04	73.46	
	Level 2	51.16	20.92	29.70		80.52	74.09	77.17	
	Level 3	77.01	82.07	79.46		80.65	83.89	82.24	
	Average	70.09	66.37	66.43	76.62	78.90	78.47	78.54	78.45
Decision Tree	Level 0	84.50	86.85	85.66		80.26	81.56	80.91	
	Level 1	63.95	63.95	63.95		64.50	65.64	65.07	
	Level 2	37.56	24.15	29.40		65.36	66.99	66.17	
	Level 3	69.71	70.59	70.15		75.53	71.07	73.23	
	Average	63.93	61.39	62.29	72.00	71.42	71.32	71.34	71.27

Analytical results indicate that TRV_SPD1 and TRV_SPD2 make the greatest contribution to Level 0 predictions, suggesting that the travel speeds of both collision parties play a crucial role in determining accident severity by influencing kinetic energy transfer. Lower velocities (cool-toned colors) increase the likelihood of Level 0 predictions, whereas higher velocities (warm-toned colors) lower the probability of being classified as Level 0.

Further investigation reveals that most features exhibit nonlinear relationships between their values and SHAP contributions. Most features exhibit a bimodal distribution, where lower-value ranges correspond to predominantly negative SHAP values, transitioning to positive values at higher ranges. Additionally, certain features display polarity reversal across different severity classes. This phenomenon suggests that the model captures context-dependent variations in feature contributions. Notably, vehicle speed parameters are strongly correlated with accident severity, supporting the validity

of the model's decision-making process and its consistency with fundamental collision dynamics principles.

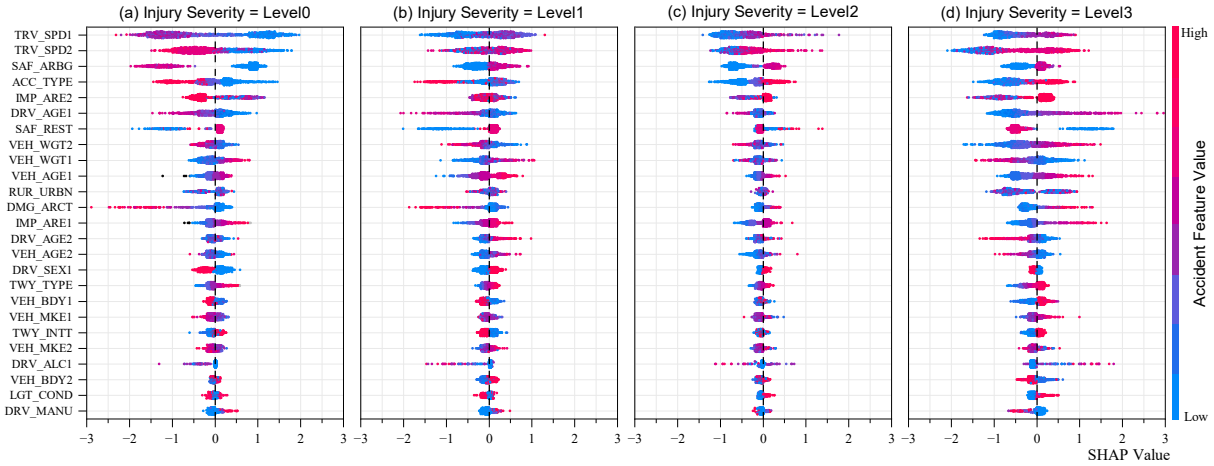

Figure 7. Global SHAP value distribution of accident features.

4.1.1. Feature Importance Analysis

Feature importance analysis is essential for understanding model decision-making and guiding accident analysis. While SHAP value distributions show how individual feature values impact predictions, they do not quantify the overall significance of features. To address this, the mean absolute SHAP value (Mean(|SHAP|)) is used to measure the average contribution of each feature across all predictions. The feature importance ranking results are shown in Figure 8, with the feature importance for another eight types of accidents presented in Figure A1. TRV_SPD1, TRV_SPD2, ACC_TYPE, and IMP_ARE2 are identified as the most influential features in the model. However, globally important features do not always align with those most critical for specific severity levels. For instance, the primary predictors for Level 3 severity are TRV_SPD2, DRV_AGE1, and SAF_ARBG.

Moreover, to assess the impact of feature importance on model performance, features were progressively removed in ascending order of importance, followed by model retraining and evaluation. The relationship between feature count and model accuracy, illustrated in Figure 9, demonstrates that retaining high-importance features results in significant performance improvements, while low-importance features contribute less substantially.

Although global rankings identify key predictors, they may overlook features with strong localized effects. For example, DRV_ALC1 ranks 22nd globally but exhibits widely dispersed SHAP values, indicating that despite its low frequency, it has a substantial impact on accident severity. This underscores the necessity of considering both overall and scenario-specific importance for a more comprehensive assessment of accident severity determinants.

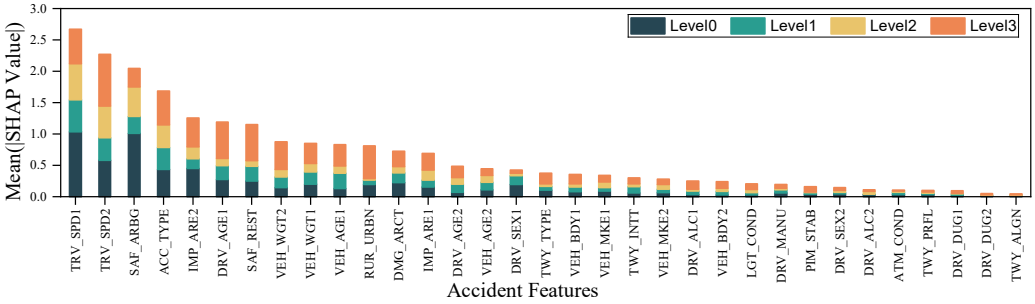

Figure 8. Global feature importance ranking.

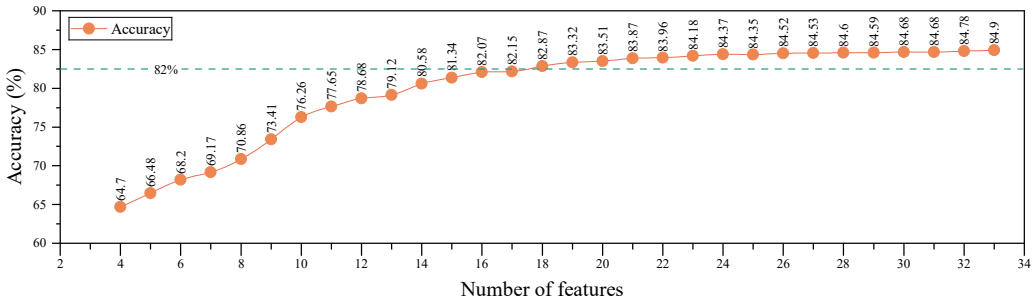

Figure 9. Relationship between feature numbers and model performance.

4.1.2. Feature Dependency Analysis

In this study, feature dependency analysis is performed using the SHAP dependence plot, which captures the marginal effect of individual features on the target variable while illustrating the distribution and variability of SHAP values. Each point on the plot represents the predicted target value for a specific observation in the dataset. The analysis focuses on Level 3 severity, representing fatal accidents—the most critical crash severity category. Thus, feature dependency analysis primarily examines feature behavior at this level to provide deeper insights into their contributions to fatal crash outcomes. Specifically, four key features—TRAV_SPD1, IMP_ARE1, VEH_MKE1, and VEH_BDY1—are analyzed. Additionally, the system autonomously identifies the feature with the strongest interaction within the dataset, as shown in Figure 10.

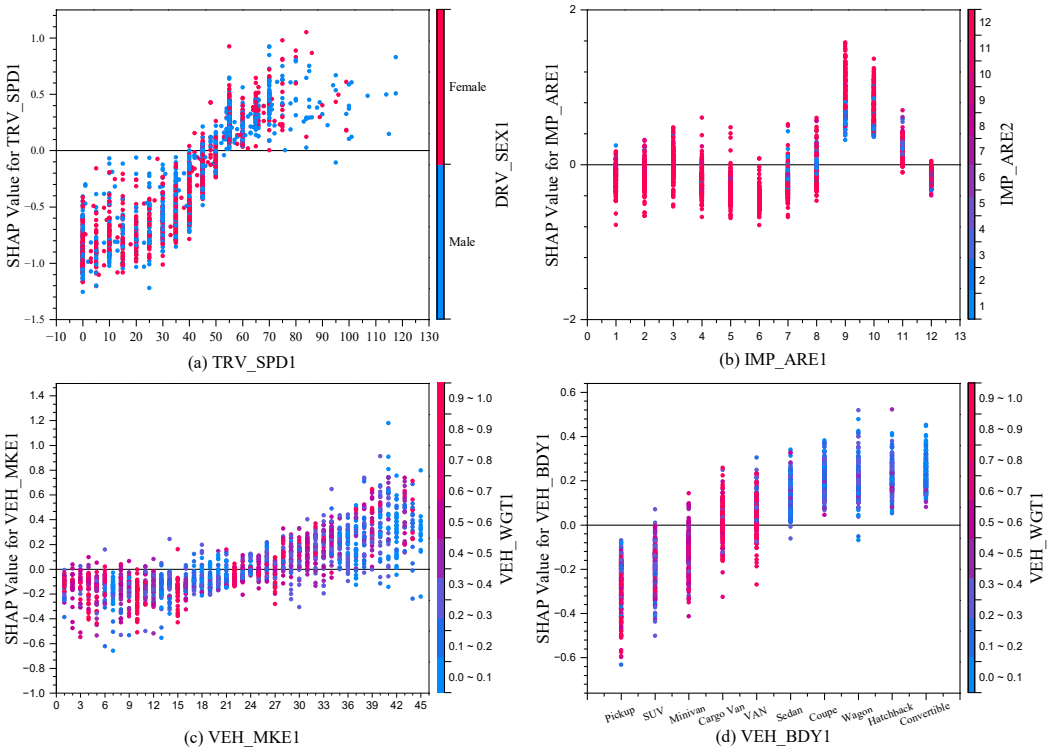

Figure 10. The dependence of accident features.

Figure 10a presents the interaction between TRV_SPD1 and DRV_SEX1. As TRV_SPD1 increases, the SHAP value shifts from negative to positive along the y-axis, with a transition threshold between 40 mph and 55 mph, indicating a critical speed range for predicting fatal accidents. The scatter point colors represent driver gender, revealing nearly identical SHAP distributions for male and female drivers. Although DRV_SEX1 exhibits a globally bimodal

impact, its narrow SHAP distribution suggests a negligible overall influence compared to speed, whose broader distribution signifies a more substantial effect.

Figure 10b illustrates the interaction between IMP_ARE1 and IMP_ARE2, underscoring the strong association between impact location and fatal crash outcomes. Collisions at the 9 o'clock and 10 o'clock positions contribute most significantly to fatal injury predictions, aligning with real-world crash dynamics. These zones, located near the driver, absorb less impact energy, leading to severe structural deformation and reduced survival space. The effect is exacerbated when these areas are struck by another vehicle's 11 o'clock or 12 o'clock positions, where impact forces are typically highest.

Figure 10c examines the influence of VEH_MKE1 and VEH_WGT1, demonstrating brand-specific variations in vehicle safety performance. The x-axis represents encoded vehicle brand labels (refer to Figure 4), allowing for a direct comparison of brand influence on accident severity. Scatter point colors denote vehicle curb weight, highlighting how certain manufacturers specialize in distinct vehicle categories. For example, Hummer (15) focuses on off-road vehicles, while RAM (10) primarily produces light trucks—both featuring high curb weights and reinforced body structures. Conversely, Volkswagen (13) and Tesla primarily manufacture passenger cars with lower curb weights but high safety ratings, aligning with their real-world safety reputations.

Figure 10d explores the relationship between VEH_BDY1 and VEH_WGT1, emphasizing the significant variation in crash survivability across vehicle body types. Light trucks exhibit lower susceptibility to fatal crashes compared to passenger vehicles (PVs), which are more prone to severe injuries. Scatter point colors indicate that light trucks generally have higher curb weights than PVs. Among body types, pickup trucks demonstrate the highest safety performance, whereas convertibles rank lowest, underscoring the critical role of vehicle structure and weight in crash outcomes.

From the above-mentioned feature dependencies, it is clearly evident that the vehicle's driving speed, collision location, brand, and type all have an impact on the severity of driver injuries. For drivers, in complex traffic environments, it is crucial to reduce the driving speed to a safe range and remain fully focused while driving, enabling them to promptly take risk-mitigation measures. As for vehicle manufacturers, they should first enhance the safety design indicators and safety test standards of vehicles. Additionally, efforts should be made to reinforce the vulnerable parts of vehicles, such as the sides of the body. This comprehensive approach can effectively improve vehicle safety performance and better protect the safety of drivers.

4.1.3. Bivariate SHAP Contribution Analysis

Given that SHAP values exhibit additivity, the sum of individual feature SHAP values equals their cumulative contribution to the model's prediction. This property allows for the evaluation of the combined impact of two or more features on the target variable by aggregating their SHAP values. In this study, bivariate SHAP contribution analysis investigates the joint influence of feature pairs on model predictions by analyzing the sum of their SHAP values. This approach uncovers feature interactions and their collective impact on accident severity classification. By visualizing the aggregated SHAP contributions through a scatter heatmap, patterns emerge that show how specific feature combinations lead to higher or lower predicted severity levels, providing deeper insights into interdependent factors driving model decisions.

Figure 11 systematically explores how multiple factor combinations affect injury severity. The first set of sub-figures, (a–d), highlights the significant role of TRV_SPD1 and TRV_SPD2. With the SHAP value-colored scatter points, we can clearly observe a direct correlation: as vehicle speeds increase, so does the predicted injury severity.

Contrastingly, sub-figures (e–h) shift the focus to DRV_AGE1 and VEH_AGE1. Here, the data reveals that elderly drivers, especially those over 80, are more prone to severe (Level 3) injuries. Intriguingly, this age-related risk threshold decreases with the increasing age of the vehicle, suggesting that newer vehicles offer better protection to older drivers.

Finally, the last part of the figure, (i–l), examines the relationship between VEH_WGT1 and VEH_WGT2. In comparison to the speed-and age-related factors, the results show that heavier vehicles have an advantage in reducing injury severity during collisions, presenting a different but equally important pattern in injury risk assessment.

This analysis effectively reveals the joint influence of feature pairs on accident severity outcomes. The clustering of colors within the heatmaps highlights distinct support, rejection, and transition regions. Red regions indicate the support domain, where predictions are more likely to align with the corresponding category, while deep blue regions represent the rejection domain, where predictions are less likely to belong to the specified category. Light green and light yellow regions correspond to transition zones, where predictions exhibit intermediate tendencies and neither strongly support nor reject the predicted outcomes. The clarity of these clusters underscores the contribution of these feature interactions in enhancing model prediction accuracy.

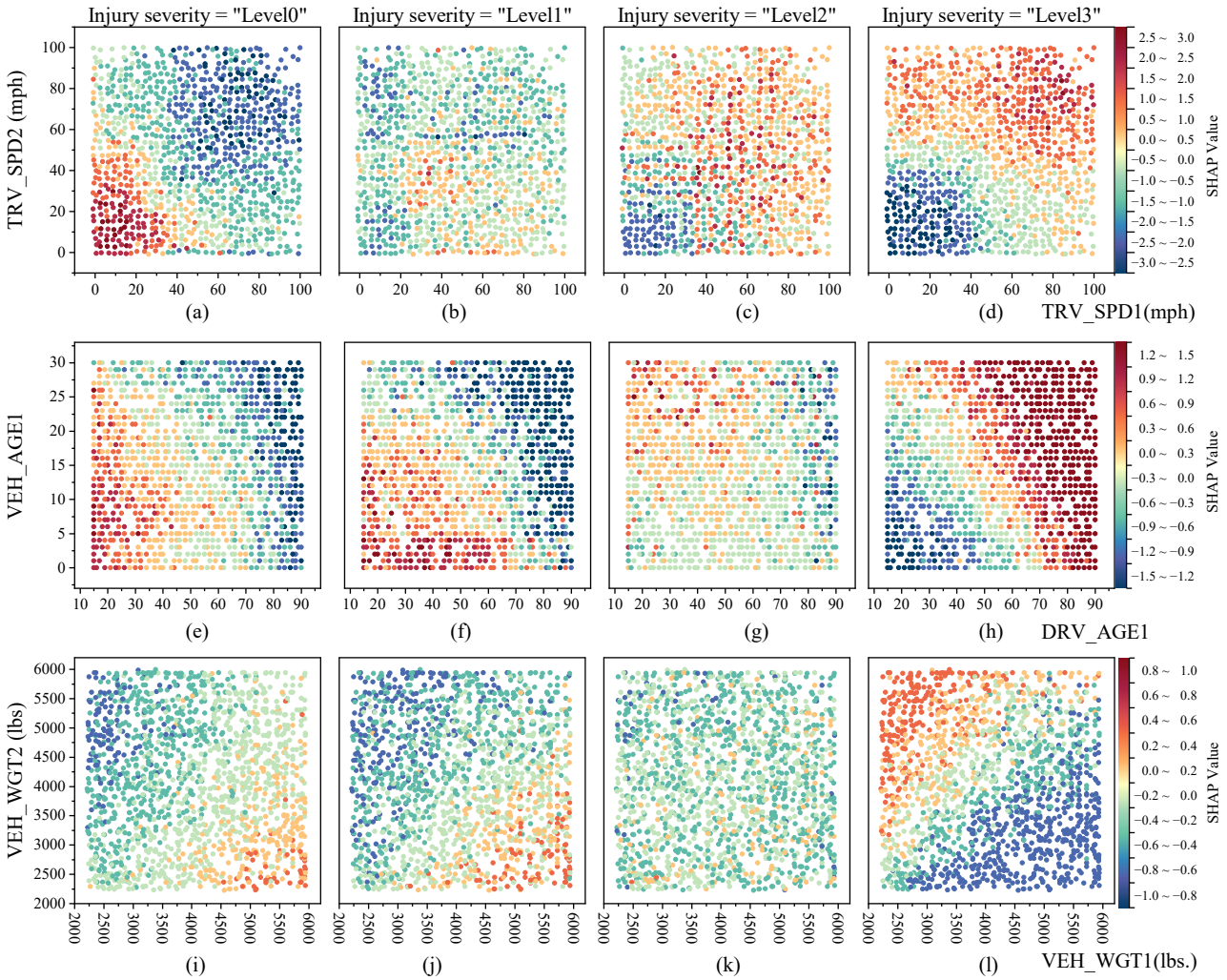

Figure 11. Scatter plot of combined SHAP impact on accident severity. (**a**–**d**) TRV_SPD1 and TRV_SPD2, (**e**–**h**) DRV_AGE1 and VEH_AGE1, and (**i**–**l**) VEH_WGT1 and VEH_WGT2, with all plots showing combined SHAP values for injury severity levels (Level 0–Level 3).

4.1.4. Instance-Level SHAP Interpretation

Instance-Level SHAP Interpretation decomposes the SHAP values for a single prediction instance, revealing the specific contribution of each feature to the final prediction and explaining why the model made that particular decision. Figure 12 presents the SHAP value waterfall plot for a randomly selected accident case from the dataset, where the actual injury severity is classified as "Level 3". The x-axis represents the SHAP values, while the y-axis lists the features in descending order of importance. The feature labeled 'Others' represents the cumulative SHAP values of lower-contributing features, ensuring a more concise visualization.

The SHAP value waterfall plot provides an intuitive and transparent representation of the model's decision-making process, illustrating how individual features contribute to the final prediction. The baseline value, $E(f(x)$, representing the expected model output across the dataset, is computed as 0.360, 0.568, 0.555, and 0.465 for injury severity levels 0 to 3, respectively. By aggregating the SHAP values of each feature relative to this baseline, the cumulative SHAP value for each severity level is obtained. The final predicted probabilities for each severity category are then derived using the Softmax function.

This instance-level interpretation enhances the explainability of the model by highlighting the key factors that influence the classification outcome. Such insights are particularly valuable for understanding the underlying risk factors in individual accident cases and can aid in data-driven decision-making for traffic safety analysis and policy development.

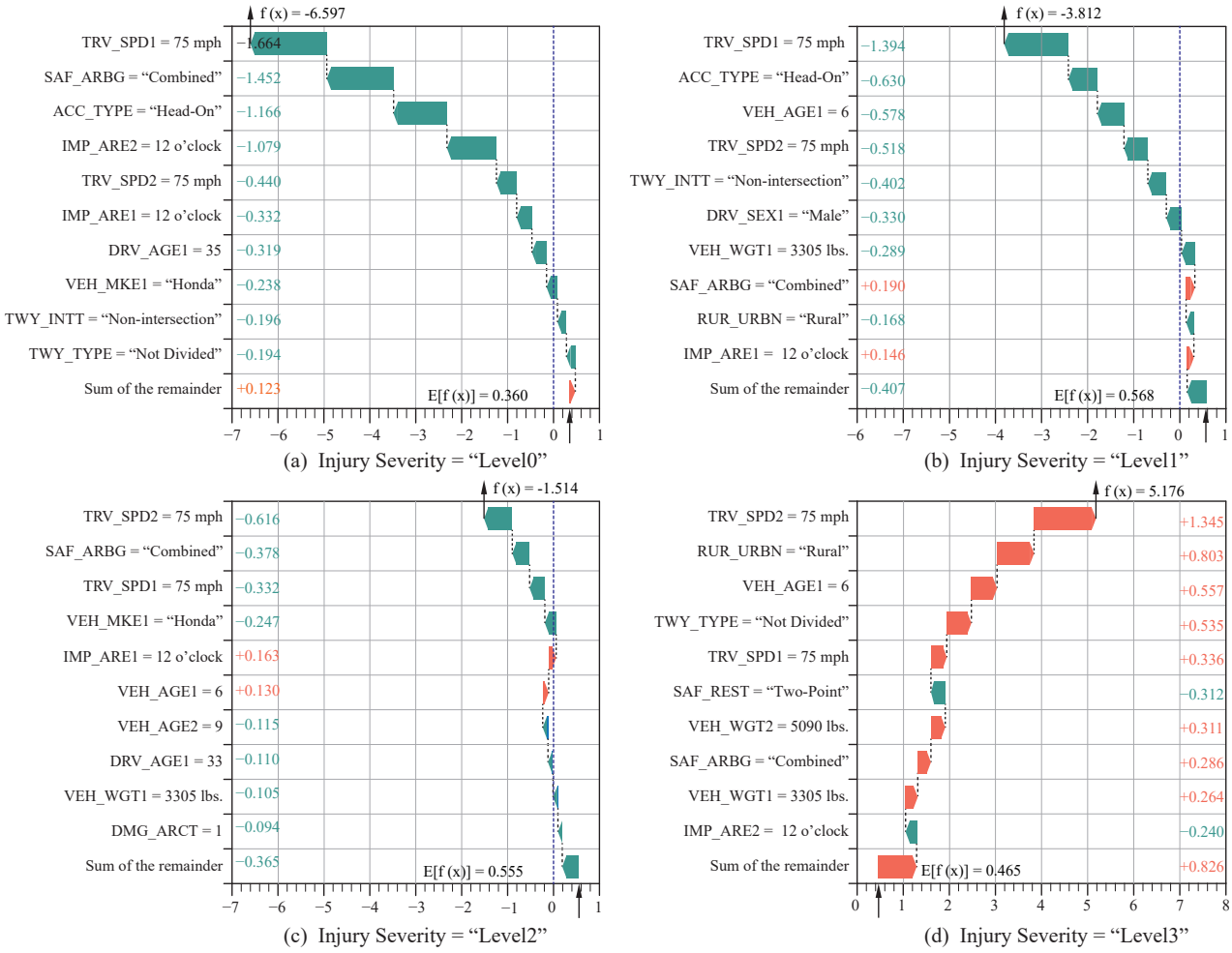

Figure 12. Local SHAP waterfall plot for a single multi-class prediction case (True label: 'Level 3').

5. Implication and Limitations

5.1. Implications

This study focuses on driver injury severity prediction and factor analysis in PV-PV collisions, addressing a critical research gap in this domain. By leveraging interpretable machine learning techniques, we not only enhance prediction accuracy but also provide deeper insights into the key factors influencing injury severity, thereby informing traffic safety improvements and accident prevention strategies.

Moreover, by integrating the CRSS and FARS datasets, we have significantly mitigated the issue of class imbalance, reducing it by two orders of magnitude. This enhancement strengthens the model's ability to predict high-severity injury cases, which are often underrepresented in imbalanced datasets. Given that accurately identifying severe injury cases is essential for emergency response planning and policy formulation, this improvement substantially increases the practical applicability of our findings.

Furthermore, this study employs SHAP to interpret the machine learning model, enabling a precise quantification of each feature's impact, magnitude, and direction on injury severity. This interpretability not only enhances the model's credibility but also provides actionable insights for stakeholders, such as policymakers and vehicle manufacturers, facilitating the development of targeted safety interventions.

5.2. Limitations

While the study yielded satisfactory results, it has several aspects that warrant further consideration. Traffic accident analysis and injury severity prediction are inherently complex. On one hand, crucial elements such as driver psychological state and fatigue are challenging to quantify within structured datasets, and the collision dynamics, especially the influence of secondary impacts on injury severity, are inadequately captured in conventional data. On the other hand, in applying SHAP for feature analysis, although it successfully clarified the individual impact of features, it lacks the ability to account for the dynamic interactions among them.

5.3. Future Work

Future research could enhance injury severity prediction through multi-pronged approaches. Firstly, incorporating vision-based methods to automatically extract both macro-level features like vehicle trajectories and collision impact forces, as well as micro-level features such as driver facial expressions and pre-crash behavior from accident videos. This approach would enable a more comprehensive understanding of the accident process, overcoming the limitations of relying solely on structured datasets that capture only a single moment in time. Secondly, to address the issue of analyzing dynamic feature interactions, future studies could explore advanced techniques that can better model the complex relationships between features.

6. Conclusions

The interpretability of prediction models. By constructing an XGBoost prediction model, it demonstrates optimal performance in terms of indicators such as accuracy (84.9%), precision (84.85%), recall (84.9%), and F1-score (84.87%). It achieves a high prediction accuracy across all four severity levels and particularly outperforms existing studies in predicting minority-class samples. While reducing the risk of overfitting, this model can extract the accident causation mechanisms embedded in the accident data through training, exhibiting strong generalization ability. In the aspect of feature engineering, this study focuses on the specific scenarios of PV-PV collision accidents, screening out key accident features and eliminating irrelevant variables, which enhances the practicality and

pertinence of the model. Additionally, the SHAP-based model interpretability analysis is carried out at three levels: global feature importance, local prediction interpretation, and specific case validation. The results show that the prediction logic of the model is consistent with the statistical trends of the data, and the conclusions of the case analysis conform to the actual patterns, further validating the reliability of the model. The accident feature analysis of this study visually demonstrates the direction and extent of the influence of different features on the severity of driver injuries. The results can support policymaking, accident investigation, and vehicle safety analysis. For example, from the analysis results of the dependence of accident features on vehicle brands, the relationship between different vehicle brands and the severity of accidents is clearly revealed. These results can reflect the safety characteristics of vehicles from different brands in the actual environment, guiding regulatory authorities to supervise automakers with poor safety performance. Future research can further expand the data scale, optimize the model structure, and integrate multimodal data to improve prediction accuracy and generalization ability.

Author Contributions: Conceptualization, P.L., W.Y., and W.G.; methodology, P.L., W.Z., X.W., and W.G.; software, P.L., W.G., and W.Z.; validation, P.L. and X.W.; formal analysis, P.L. and W.Y.; resources, W.Y. and P.L.; investigation, P.L., W.Z., and X.W.; data curation, P.L. and W.Y.; writing—original draft, P.L., W.Z., and W.G.; writing—review and editing, P.L., X.W., and W.Z.; visualization, P.L. and W.Y.; supervision, P.L., W.Z. and W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are derived from the publicly available NHTSA crash dataset, which can be accessed at https://www.nhtsa.gov/data/crash-data-systems, accessed on 25 December 2024. However, the processed dataset used in this study is available upon request from the corresponding author due to confidentiality and personal data protection regulations.

Conflicts of Interest: Authors Jiejie Xu, Wangpengfei Yu and Yang Chen are employed by the Shanghai Intelligent Vehicle cooperating Innovation Center. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

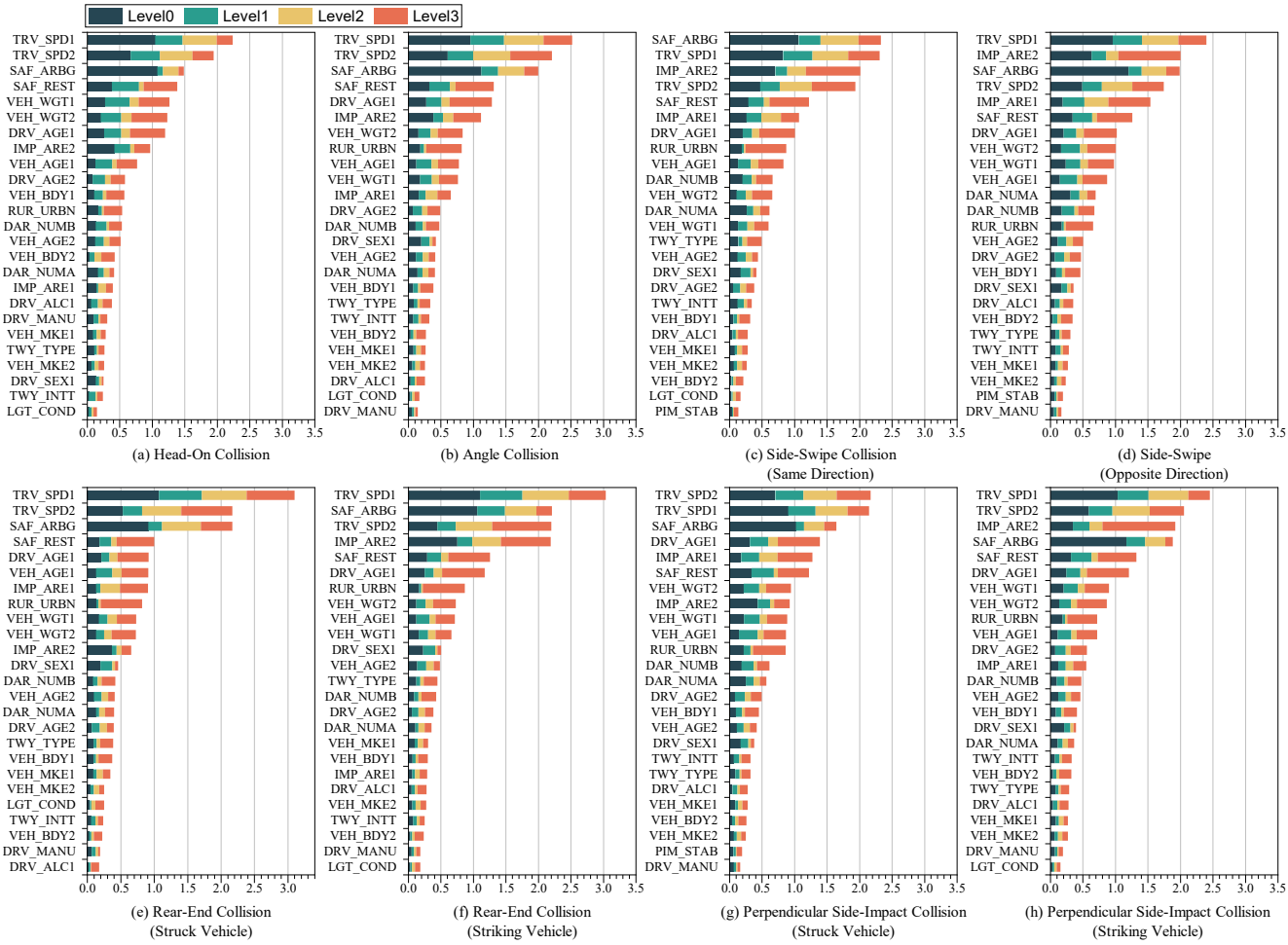

Figure A1. The ranking of the feature importance of eight typical accident types.

References

1. European Automobile Manufacturers' Association (ACEA). *ACEA Pocket Guide 2024–2025*; Report No. N/A; European Automobile Manufacturers' Association: Brussels, Belgium, 2024.
2. International Transport Forum (ITF). *Road Safety Annual Report 2024*; OECD Publishing: Paris, France, 2024.
3. Guo, W.; Li, J.; Song, X.; Zhang, W. A game-theoretic driver steering model with individual risk perception field generation. *Accid. Anal. Prev.* **2025**, *211*, 107869. [CrossRef]
4. International Organization for Standardization. *Road Vehicles–Types–Terms and Definitions* (ISO Standard No. 3833:1977). 1977. Available online: https://www.iso.org/standard/9389.html (accessed on 25 December 2024).
5. World Health Organization. *Global Status Report on Road Safety 2023*; World Health Organization: Geneva, Switzerland, 2023; Licence: CC BY-NC-SA 3.0 IGO.
6. National Center for Statistics and Analysis. *Traffic Safety Facts 2022: A Compilation of Motor Vehicle Traffic Crash Data*; Report No. DOT HS 813 656; National Highway Traffic Safety Administration: Washington, DC, USA, 2024.
7. Guo, W.; Song, X.; Zhang, W.; Li, J.; Wu, X. Game-Theoretic Shared Control Strategy for Cooperative Collision Avoidance Under Extreme Conditions. *IEEE Trans. Veh. Technol.* **2024**, *74*, 246–262. [CrossRef]
8. Chand, A.; Jayesh, S.; Bhasi, A.B. Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. *Mater. Today Proc.* **2021**, *47*, 5135–5141. [CrossRef]
9. Pourroostaei Ardakani, S.; Liang, X.; Mengistu, K.T.; So, R.S.; Wei, X.; He, B.; Cheshmehzangi, A. Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability* **2023**, *15*, 5939. [CrossRef]
10. Setiadi, D.R.I.M.; Islam, H.M.M.; Trisnapradika, G.A.; Herowati, W. Analyzing preprocessing impact on machine learning classifiers for cryotherapy and immunotherapy dataset. *J. Future Artif. Intell. Technol.* **2024**, *1*, 39–50. [CrossRef]

11. Song, D.; Yang, X.; Yang, Y.; Cui, P.; Zhu, G. Bivariate joint analysis of injury severity of drivers in truck-car crashes accommodating multilayer unobserved heterogeneity. *Accid. Anal. Prev.* **2023**, *190*, 107175. [CrossRef]
12. Gong, H.; Fu, T.; Sun, Y.; Guo, Z.; Cong, L.; Hu, W.; Ling, Z. Two-vehicle driver-injury severity: A multivariate random parameters logit approach. *Anal. Methods Accid. Res.* **2022**, *33*, 100190. [CrossRef]
13. Cerwick, D.M.; Gkritza, K.; Shaheed, M.S.; Hans, Z. A comparison of the mixed logit and latent class methods for crash severity analysis. *Anal. Methods Accid. Res.* **2014**, *3*, 11–27. [CrossRef]
14. Sheikh, M.S.; Peng, Y. Modeling collision risk for unsafe lane-changing behavior: A lane-changing risk index approach. *Alex. Eng. J.* **2024**, *88*, 164–181. [CrossRef]
15. Zhang, J.; Jin, M.; Wan, C.; Dong, Z.; Wu, X. A Bayesian network-based model for risk modeling and scenario deduction of collision accidents of inland intelligent ships. *Reliab. Eng. Syst. Saf.* **2024**, *243*, 109816. [CrossRef]
16. Liu, Z.; Chen, Y.; Xia, F.; Bian, J.; Zhu, B.; Shen, G.; Kong, X. TAP: Traffic Accident Profiling via Multi-Task Spatio-Temporal Graph Representation Learning. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–25. [CrossRef]
17. Santos, K.; Dias, J.P.; Amado, C. A literature review of machine learning algorithms for crash injury severity prediction. *J. Saf. Res.* **2022**, *80*, 254–269. [CrossRef]
18. Shaik, M.E.; Islam, M.M.; Hossain, Q.S. A review on neural network techniques for the prediction of road traffic accident severity. *Asian Transp. Stud.* **2021**, *7*, 100040. [CrossRef]
19. Yu, L.; Du, B.; Hu, X.; Sun, L.; Han, L.; Lv, W. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing* **2021**, *423*, 135–147. [CrossRef]
20. Alhaek, F.; Liang, W.; Rajeh, T.M.; Javed, M.H.; Li, T. Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach. *Knowl.-Based Syst.* **2024**, *286*, 111406. [CrossRef]
21. Jamal, A.; Zahid, M.; Tauhidur, Rahman, M.; Al-Ahmadi, H.M.; Almoshaogeh, M.; Farooq, D.; Ahmad, M. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *Int. J. Inj. Control Saf. Promot.* **2021**, *28*, 408–427. [CrossRef]
22. Yan, M.; Shen, Y. Traffic accident severity prediction based on random forest. *Sustainability* **2022**, *14*, 1729. [CrossRef]
23. Wu, S.; Yuan, Q.; Yan, Z.; Xu, Q. Analyzing accident injury severity via an extreme gradient boosting (XGBoost) model. *J. Adv. Transp.* **2021**, *2021*, 3771640. [CrossRef]
24. Otchere, D.A.; Ganat, T.O.A.; Ojero, J.O.; Tackie-Otoo, B.N.; Taki, M.Y. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *J. Pet. Sci. Eng.* **2022**, *208*, 109244. [CrossRef]
25. Guo, M.; Yuan, Z.; Janson, B.; Peng, Y.; Yang, Y.; Wang, W. Older pedestrian traffic crashes severity analysis based on an emerging machine learning XGBoost. *Sustainability* **2021**, *13*, 926. [CrossRef]
26. Dong, S.; Khattak, A.; Ullah, I.; Zhou, J.; Hussain, A. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2925. [CrossRef]
27. Zahid, M.; Habib, M.F.; Ijaz, M.; Ameer, I.; Ullah, I.; Ahmed, T.; He, Z. Factors affecting injury severity in motorcycle crashes: Different age groups analysis using CatBoost and SHAP techniques. *Traffic Inj. Prev.* **2024**, *25*, 472–481. [CrossRef] [PubMed]
28. 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 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), London, UK, 20–22 December 2021; pp. 390–397.
29. 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. [CrossRef] [PubMed]
30. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, Su. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
31. Ahmed, S.; Hossain, M.A.; Ray, S.K.; Bhuiyan, M.M.I.; Sabuj, S.R. A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance. *Transp. Res. Interdiscip. Perspect.* **2023**, *19*, 100814. [CrossRef]
32. Boo, Y.; Choi, Y. Comparison of mortality prediction models for road traffic accidents: An ensemble technique for imbalanced data. *BMC Public Health* **2022**, *22*, 1476. [CrossRef]
33. Wongvorachan, T.; He, S.; Bulut, O. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information* **2023**, *14*, 54. [CrossRef]
34. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
35. Toğaçar, M.; Ergen, B.; Çömert, Z.; Özyurt, F. A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. *IRBM* **2020**, *41*, 212–222. [CrossRef]

36. Ravikiran, H.K.; Deepak, R.; Deepak, H.A.; Prapulla, Kumar, M.S.; Sharath, S.; Yogeesh, G.H. A robust framework for Alzheimer's disease detection and staging: Incorporating multi-feature integration, MRMR feature selection, and Random Forest classification. *Multimed. Tools Appl.* **2024**. [CrossRef]
37. Wang, G.; Lauri, F.; Hassani, A.H.E. Feature selection by mRMR method for heart disease diagnosis. *IEEE Access* **2022**, *10*, 100786–100796. [CrossRef]
38. Rezvani, S.; Wang, X. A broad review on class imbalance learning techniques, *Appl. Soft Comput.* **2023**, *143*, 110415.
39. Kingsford, C.; Salzberg, S.L. What are decision trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013. [CrossRef]
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
41. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
42. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754. Available online: https://arxiv.org/abs/1603.02754 (accessed on 20 December 2024).
43. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6649.
44. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
45. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2013.