Accident Analysis and Prevention

Prioritizing Road Safety Interventions: A Robust Multi-Model Ranking of Fatal Collision Risk Factors Using Canadian National Data

--Manuscript Draft--

Manuscript Number:	AAAP-D-26-00050
Article Type:	Full Length Article
Keywords:	Road safety; fatal collisions; collision risk factors; Policy Prioritization; consensus ranking; interpretable machine learning
Corresponding Author:	Shengkun Xie Toronto Metropolitan University Ted Rogers School of Management
First Author:	Shengkun Xie
Order of Authors:	Shengkun Xie
	Nguyen Luong
Abstract:	Identifying the key risk factors associated with fatal road collisions is fundamental for evidence-based road safety policy and targeted injury prevention. However, a persistent challenge in safety science is that different machine learning (ML) models often yield inconsistent feature-importance rankings, even when applied to identical datasets, which can undermine confidence in safety interventions. This study develops a robust, multi-method consensus framework to identify collision-related factors that are consistently associated with fatal outcomes, accounting for both model-specific bias and temporal heterogeneity. Using eight years of person-level data from the Canadian National Collision Database (2014--2021), we integrate feature-importance rankings from Random Forests, Extreme Gradient Boosting, and Neural Networks across all study years into a unified rank matrix. To extract the dominant ranking structure, we propose an Eigenvectors-Weighted Consensus Ranking (EWCR) approach based on principal component analysis and its variant, and systematically compare it with distance-to-ideal (i.e., TOPSIS-based) and Non-negative Matrix Factorization (NMF)-based consensus methods. Comprehensive agreement analyses—including rank concordance, rank-shift distributions, and stability diagnostics—demonstrate strong consistency in identifying a stable core set of fatal collision risk factors across models and time. By identifying risk factors that are consistently prioritized across models and reporting years, the proposed consensus-ranking framework provides a transparent and defensible basis for evidence-based road safety policy, supporting the prioritization and targeting of interventions aimed at reducing fatal collisions.

Dear Editor,

I am pleased to submit our original research article, **"Prioritizing Road Safety Interventions: A Robust Multi-Model Ranking of Fatal Collision Risk Factors Using Canadian National Data,"** for consideration for publication in *Accident Analysis & Prevention*.

A persistent challenge in road safety science is the "model-specific dependency" of risk factor identification. Different machine learning (ML) architectures often yield inconsistent feature-importance rankings, even when applied to the same collision datasets. This study addresses this bottleneck by developing a robust, multi-method consensus framework that accounts for both model-specific bias and temporal heterogeneity.

Utilizing eight years of person-level data (2014–2021) from the Canadian National Collision Database (NCDB), we integrate feature-importance rankings from Random Forests, Extreme Gradient Boosting (XGBoost), and Neural Networks. Our primary contribution is the proposal of an **Eigenvectors-Weighted Consensus Ranking (EWCR)** approach, which we systematically compare against TOPSIS and Non-negative Matrix Factorization (NMF)-based methods.

Our findings demonstrate strong consistency in identifying a stable core set of fatal collision risk factors, reducing reliance on single-model assumptions. We believe this work provides a transparent, defensible foundation for safety practitioners to reliably prioritize interventions and allocate safety investments.

This manuscript is original work and has not been published or submitted elsewhere. All authors have approved the submission. We believe it aligns closely with the scope of *Accident Analysis & Prevention* regarding the development of evidence-based road safety policy and injury prevention.

Thank you for your time and consideration.

Sincerely,

Shengkun Xie, Ph.D. Global Management Studies Toronto Metropolitan University

Prioritizing Road Safety Interventions: A Robust Multi-Model Ranking of Fatal Collision Risk Factors Using Canadian National Data

Shengkun Xie*, Nguyen Duy Anh Luong

Global Management Studies, Ted Rogers School of Management, Toronto Metropolitan University, 350 Victoria Street, Toronto, ON M5B 2K3,Canada

ARTICLE INFO

Keywords:
Road safety
fatal collisions
collision risk factors
Policy Prioritization
consensus ranking
interpretable machine learning

ABSTRACT

Identifying the key risk factors associated with fatal road collisions is fundamental for evidence-based road safety policy and targeted injury prevention. However, a persistent challenge in safety science is that different machine learning (ML) models often yield inconsistent feature-importance rankings, even when applied to identical datasets, which can undermine confidence in safety interventions. This study develops a robust, multi-method consensus framework to identify collision-related factors that are consistently associated with fatal outcomes, accounting for both model-specific bias and temporal heterogeneity. Using eight years of person-level data from the Canadian National Collision Database (2014–2021), we integrate feature-importance rankings from Random Forests, Extreme Gradient Boosting, and Neural Networks across all study years into a unified rank matrix. To extract the dominant ranking structure, we propose an Eigenvectors-Weighted Consensus Ranking (EWCR) approach based on principal component analysis and its variant, and systematically compare it with distance-to-ideal (i.e., TOPSIS-based) and Non-negative Matrix Factorization (NMF)-based consensus methods. Comprehensive agreement analyses—including rank concordance, rank-shift distributions, and stability diagnostics—demonstrate strong consistency in identifying a stable core set of fatal collision risk factors across models and time. By identifying risk factors that are consistently prioritized across models and reporting years, the proposed consensus-ranking framework provides a transparent and defensible basis for evidence-based road safety policy, supporting the prioritization and targeting of interventions aimed at reducing fatal collisions.

1. Introduction

Road traffic collisions remain a major global public health concern, responsible for substantial human and economic losses each year (World Health Organization, 2023). A large body of research has shown that crash outcomes arise from the interaction of human, environmental, vehicular, and systemic factors rather than from isolated causes (Huang, Zhou, Koelper, Li and Nie, 2020). Identifying the key factors associated with crash outcomes is therefore fundamental for accident prevention and road safety improvement. From a policy and intervention-planning perspective, inconsistent identification of high-risk factors can translate directly into unstable safety priorities, inefficient allocation of limited resources, and reduced effectiveness of prevention strategies. Ensuring that empirical evidence used to guide road safety policy is both reliable and robust is thus of critical importance.

With the increasing availability of large-scale administrative datasets such as the Canadian National Collision Database (NCDB), modern road safety research increasingly relies on computational approaches capable of capturing nonlinear patterns, high-order interactions, and complex data structures (Shuai and Kwon, 2025). Machine learning (ML) and deep learning (DL) models, including Random Forests (RF), Extreme Gradient Boosting (XGB), and Neural Networks (NN), have therefore become prominent tools for predicting crash severity and fatal injury outcomes (Ali, Hussain and Haque, 2024; Amiri, Afshari and Soltani, 2025; Hamdan and Sipos, 2025). While these models often demonstrate superior predictive performance compared with traditional regression-based approaches, particularly in the presence of nonlinearities, multicollinearity, and imbalanced outcomes (Iranitalab and Khattak, 2017; Obasi and Benson, 2023), their interpretation typically relies on post hoc feature-importance measures that are inherently model dependent. As a result, different ML architectures may yield substantially different rankings of influential risk factors even when trained on identical datasets.

shengkun.xie@torontomu.ca (S. Xie*)
ORCID(s): 0000-0002-9533-2096 (S. Xie*)

This lack of convergence across ML models represents a persistent challenge for road safety research and evidence-based policy making. Divergent importance rankings can obscure which risk factors are consistently associated with fatal outcomes, thereby complicating the prioritization of safety interventions and weakening confidence in data-driven decision support. The problem is further compounded when collision data span multiple reporting years, as temporal heterogeneity in traffic exposure, enforcement practices, vehicle technologies, and mobility patterns may induce additional instability in factor importance. Without an explicit framework to account for both model-specific variability and temporal effects, it becomes difficult to distinguish policy-relevant risk factors from artefacts driven by algorithmic assumptions or data characteristics.

To address this challenge, recent studies have increasingly adopted hybrid, ensemble, and consensus-based approaches to improve robustness in feature-importance assessment by synthesizing evidence across multiple modelling paradigms (Gyawali, Liu, Zou and He, 2022; Pes, 2020; Rengasamy, Rothwell and Figueredo, 2021b). Consensus-oriented frameworks are particularly valuable in the road safety context, as they facilitate the identification of stable, persistent risk factors while reducing sensitivity to individual model specifications (Rengasamy, Mase, Torres, Rothwell, Winkler and Figueredo, 2021a; Saeys, Abeel and Van de Peer, 2008). Indeed, prior empirical work has consistently highlighted factors such as crash configuration, vehicle involvement, driver age, roadway geometry, and weather conditions as key contributors to fatal collision outcomes (Ghandour, Hammoud and Al-Hajj, 2020; Luan, Jiang, Yang, Meng et al., 2024; Sun, Li, Zeng and Wang, 2025; Zhang, Lindsay, Clarke, Robbins and Mao, 2000), underscoring the importance of integrative analytical approaches that reconcile findings across models and time periods.

Against this background, the present study constructs a consensus-based ranking of factor-level importance for fatal collision outcomes using eight years of NCDB data (2014–2021). Rather than relying on a single predictive model, feature-importance information derived from RF, XGB, and NN models is aggregated across all study years into a unified rank matrix. A multi-method consensus-ranking framework is then developed to identify a stable core set of influential collision-related factors, with particular emphasis on convergence across models and temporal consistency. By explicitly addressing instability in feature-importance rankings, this approach provides a transparent and policy-relevant foundation for prioritizing fatal-collision risk factors in road safety analysis. The main contributions of this study are summarized as follows:

- Using eight years of person-level data from the Canadian National Collision Database (2014–2021), this study applies multiple machine learning models to predict fatal injury outcomes and systematically examines factor-level importance while accounting for cross-model and temporal variability.

- We develop and compare a set of consensus-based ranking frameworks, including an Eigenvectors-Weighted Consensus Ranking (EWCR) derived from principal component analysis, a distance-to-ideal approach based on TOPSIS, and a Non-negative Matrix Factorization (NMF)–based weighted consensus ranking, to integrate feature-importance information across models and reporting years.

- The resulting consensus rankings provide a robust prioritization of fatal-collision determinants and offer a transparent, defensible decision-support tool for road safety practitioners, supporting evidence-based prioritization of safety investments and interventions.

The remainder of the paper is organized as follows. Section 2 reviews the existing literature on machine learning applications in road safety research. Section 3 describes the data and outlines the proposed consensus-ranking framework. Section 4 presents and discusses the empirical findings, with particular emphasis on their implications for road safety policy and practice. Finally, Section 5 concludes the paper and highlights directions for future research.

2. Literature Review

Road safety research has long sought to identify the factors associated with crash severity and fatal injury outcomes. Early studies relied predominantly on classical statistical approaches, including logistic regression, multinomial logit, and ordered probit models (Jermakian, 2011; Zhang et al., 2000). These methods provided transparent inference and established foundational insights into the effects of driver characteristics, roadway conditions, and environmental factors on crash outcomes (Asgarzadeh, Fischer, Verma, Courtney and Christiani, 2018; Li, Graham, Ding and Ren, 2019; Phun, Kato and Yai, 2018). However, their reliance on restrictive assumptions—such as linearity, independence, and predefined functional forms—limits their ability to capture complex nonlinear interactions and heterogeneous data structures that are common in modern collision datasets.

The growing availability of large-scale administrative crash databases has accelerated the adoption of machine learning (ML) techniques for crash severity analysis. Ensemble methods such as Random Forests and Extreme Gradient Boosting, along with neural network models, have demonstrated strong predictive performance across diverse traffic and geographic contexts (Elyassami, Hamid and Habuza, 2021; Islam, Abdel-Aty, Islam and Abdelraouf, 2024; Jamal, Zahid, Tauhidur Rahman, Al-Ahmadi, Almoshaogeh, Farooq and Ahmad, 2021). Systematic reviews confirm that ML-based approaches often outperform traditional regression models, particularly when handling high-dimensional feature spaces and imbalanced outcomes (Santos, Dias and Amado, 2022). Empirical studies further illustrate the utility of ML models for identifying influential risk factors, including vehicle type, roadway environment, and contextual conditions (Akter, Susilawati, Zubair and Chor, 2025; Gutiérrez-Rodríguez, Rojí and Cuadrado, 2025; Sorum and Pal, 2024).

Despite these advances, much of the ML literature in road safety remains focused on predictive accuracy, with comparatively less attention paid to the stability and consistency of feature-importance rankings. Most studies rely on model-specific importance measures—such as Gini importance in Random Forests or gain-based metrics in boosting algorithms—without assessing whether identified rankings are robust across alternative modeling paradigms (Jamal et al., 2021; Pathivada, Banerjee and Haleem, 2024). As a result, different ML models trained on the same data can yield divergent prioritizations of risk factors, creating ambiguity for interpretation and policy application.

This challenge is further compounded by temporal heterogeneity in collision data. Changes in traffic exposure, enforcement practices, vehicle technologies, and mobility patterns can lead to shifts in the apparent importance of risk factors over time. Such temporal instability has been particularly evident during disruptive periods such as the COVID-19 pandemic, which substantially altered travel behavior and crash dynamics (Saladié, Bustamante and Gutiérrez, 2020; Shaik and Ahmed, 2022; Vanlaar, Woods-Fry, Barrett, Lyon, Brown, Wicklund and Robertson, 2021). Analyses confined to single years or short time windows therefore risk emphasizing transient effects rather than identifying structurally persistent risk factors relevant for long-term road safety planning and regulation.

Hybrid and ensemble learning frameworks have been proposed to improve robustness and generalizability by combining multiple algorithms (Jamal et al., 2021; Nara, Sordi, Schaefer, Schreiber, Baierle, Sellitto and Furtado, 2019; Wang and Boukerche, 2020). While these approaches often enhance predictive performance, they typically do not provide a unified or interpretable assessment of factor importance across constituent models. This limitation has motivated the development of consensus-oriented frameworks aimed at identifying predictors that are consistently influential across models and analytical contexts (Abbas, Ying and Ayoubi, 2025; Gyawali et al., 2022; Pes, 2020; Rengasamy et al., 2021b). Related studies highlight the importance of accounting for temporal and contextual variability when assessing crash risk, including time-of-day effects, weather conditions, and behavioral factors (Golestani, Rezaei, Malekpour, Ahmadi, Ataei, Khosravi, Jafari, Shahraz and Farzadfar, 2025; Leoni, BahooToroody, Abaei, Cantini, BahooToroody and De Carlo, 2024; Pei, Wen and Pan, 2025; Yuan, Xiang, Huang and Gu, 2023; Zhang, Huang, Kuang, Yu, Zhu and Yang, 2025). Causal ML approaches have also been explored to address spurious associations in heterogeneous traffic datasets (Srivastava, Gohil and Ray, 2024).

Despite these methodological developments, formal comparisons of alternative consensus-ranking techniques remain limited, particularly in the context of multi-year collision data. Few studies systematically integrate feature-importance information across multiple ML models and reporting years to identify risk factors that are both stable and policy relevant. Consequently, a gap persists between advances in predictive crash severity modeling and the requirements of robust, evidence-based road safety policy. Addressing this gap calls for transparent, multi-method consensus-ranking frameworks that explicitly account for both model-specific and temporal variability in feature importance.

3. Materials and Methods

3.1. Data Source

This study uses data from the Canadian National Collision Database (NCDB), maintained by Transport Canada. The NCDB compiles police-reported collisions from all provinces and territories, providing detailed information on crash circumstances, involved vehicles, roadway conditions, and injury outcomes. Each row represents an individual involved in a collision. Our analysis covers the period from 2014 to 2021, capturing temporal variations in mobility, traffic exposure, and travel behavior that influenced collision characteristics and injury severity. The dataset includes variables at three hierarchical levels:

Table 1
Example of key risk factor variables, with variable code, name, and description.

Variable Code	Short Name	Description
P_ISEV	Injury Severity	Binary indicator of fatal or serious injury (target).
C_CONF	Collision Configuration	Type of crash (rear-end, head-on, angle, turning).
C_VEHS	Number of Vehicles	Count of vehicles involved in the collision.
C_HOUR	Collision Hour	Hour of day (0–23) when the crash occurred.
V_TYPE	Vehicle Type	Vehicle category (car, motorcycle, truck, bus, etc.).
C_WDAY	Day of Week	Identifies weekday versus weekend crashes.
V_YEAR	Vehicle Model Year	Proxy for vehicle safety features and technology level.
P_ID	Person Identifier	Role of the person in the collision (driver, passenger, pedestrian).
C_RALN	Road Alignment	Roadway alignment (straight, curve, hill).
C_RSUR	Road Surface	Surface condition at the time of the crash (dry, wet, snow, ice).

- Crash-level variables: Characteristics describing the collision event, including crash configuration, time of day, day of week, roadway geometry, and surface conditions.
- Vehicle-level variables: Details for each vehicle involved, such as vehicle type, model year, and movement characteristics.
- Person-level variables: Information about the individual involved, including role (driver, passenger, pedestrian) and injury severity.

The binary injury severity variable (ISEV) serves as the target, indicating whether the individual sustained a fatal or serious injury. Across modeling frameworks, certain predictors consistently demonstrated strong influence on injury severity. These factors, summarized in Table 1, encompass environmental, behavioral, and technological contributors to crash outcomes. To ensure comparability across years, all data underwent a standardized preprocessing step. This included harmonization of categorical codes, imputation of missing values, normalization of continuous variables, and one-hot encoding of categorical attributes. The resulting datasets provided a consistent foundation for descriptive analyses, model training, and stability assessments, both with and without conditioning on ISEV. This rigorous data preparation enhances the reliability and interpretability of model outputs, supporting evidence-based identification of priority risk factors and informing targeted road safety interventions and policy decisions.

3.2. Overall Framework

Figure 1 illustrates the comprehensive analytical framework developed to extract stable risk-factor insights from longitudinal collision data. The workflow is organized into four sequential phases: (1) Data Synthesis & Modeling, where eight years of NCDB person-level data (2014-2021) are processed through a multi-model ensemble (Random Forest, XGBoost, and Neural Networks) to generate a high-dimensional rank matrix; (2) Consensus Aggregation, which moves beyond simple mean-ranking to employ sophisticated information-fusion techniques, including the proposed Eigenvectors-Weighted Consensus Ranking (EWCR), TOPSIS, and NMF-based approaches; (3) Robustness Profiling, where the stability of these rankings is rigorously stress-tested through rank-shift distributions and overlap diagnostics to identify cross-method convergence; and (4) Sparsity-Driven Refinement, utilizing a Non-negative PCA (NPCA) extension to isolate the core 'stable' risk factors from secondary influences. This hierarchical framework ensures that the final synthesis of fatal-collision determinants is resilient to both model-specific artifacts and temporal fluctuations, providing a validated basis for targeted safety policy.

3.3. Feature Importance Measures from Machine Learning Models

To account for potential variation in identified risk factors across analytical methods and time periods, feature-importance measures were derived from multiple machine learning models trained separately on each year of the NCDB data. For each reporting year, three supervised learning models, Random Forest (RF), Extreme Gradient Boosting (XGB), and a feedforward Neural Network (NN), were trained using the full set of cleaned variables to predict a binary collision-severity outcome (i.e., fatal versus non-fatal). Consistent training–testing splits, preprocessing procedures, and hyperparameter settings were applied across all years to ensure that differences in feature importance reflected underlying data patterns rather than modeling inconsistencies. This multi-model, multi-year design allows assessment

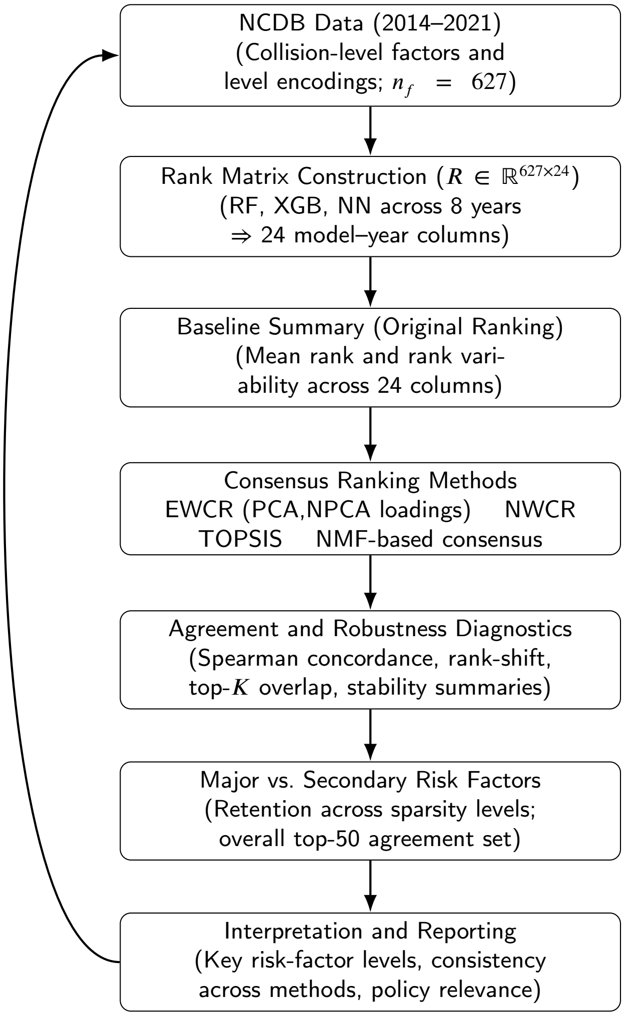


Figure 1: Overall framework for cross-model, cross-year consensus factor ranking using NCDB (2014–2021).

of the stability of factor level importance over time, which is essential for informing reliable road safety analysis and policy prioritization.

Feature importance was quantified using model-specific but widely adopted measures. For the RF model, importance was assessed using the mean decrease in Gini impurity. The XGB model was trained with a learning rate of 0.05, a maximum tree depth of 6, and 300 boosting iterations, with feature importance measured through split-based gain. The NN model consisted of three fully connected layers with ReLU activations and was optimized using the Adam algorithm. Because neural networks do not provide intrinsic importance scores, SHAP values were computed, and the mean absolute SHAP contribution of each factor was used to represent its importance. Together, these feature importance measures capture complementary perspectives on how collision-related factors contribute to severity outcomes across different modeling paradigms.

All feature-importance outputs were subsequently transformed into ranks and combined into a unified factor-by-model-year matrix. Let n_f denote the total number of unique collision-related factors (in this work, $n_f = 627$), and let n_{my} (in this work, it is 24) represent the number of model–year combinations arising from three models over eight years. The resulting rank matrix is defined as

$$R \in \mathbb{R}^{n_f \times n_{my}},$$

where each column corresponds to a specific model applied to a specific year of data. This structure provides a consistent representation of how each factor is prioritized across models and time, facilitating direct comparison of importance rankings. Because not all factors are retained or influential in every model–year combination, some rankings may be absent. To ensure a complete and comparable matrix, any factor not appearing in a given model–year output was assigned the lowest (least influential) rank observed in that column. Formally,

$$R_{ij} = \begin{cases} R_{i,j}, & \text{if available,} \\ \max_k R_{kj}, & \text{if missing.} \end{cases} \tag{1}$$

This conservative treatment preserves the relative ordering of observed ranks, while avoiding artificial inflation of importance for absent factors. The resulting matrix R provides a comprehensive and aligned representation of factor level importance across models and years, serving as the empirical basis for subsequent consensus-ranking analyses. By explicitly incorporating variability across modeling approaches and temporal contexts, this framework supports more robust identification of collision risk factors that are relevant for evidence-based road safety policy and intervention planning.

3.4. Simple Averaging of Original Rankings Approach

The raw ranking results are organized in a matrix $R \in \mathbb{R}^{n_f \times n_{my}}$, where each column corresponds to a specific machine learning model and data reporting year combination and each row represents a collision-related factor level. To obtain an overall measure of importance for factor level i, its rank can be averaged across all models and years, leading to the following baseline aggregation:

$$\text{OriginalRank}_i = \frac{1}{n_{my}} \sum_{j=1}^{n_{my}} R_{ij}, \tag{2}$$

where smaller values indicate factor levels that consistently rank higher in model-specific importance measures. This simple averaging approach implicitly assigns equal weight to all model–year combinations. In the absence of strong empirical or theoretical justification to prioritize particular models or time periods, this assumption provides a transparent and reproducible benchmark for synthesizing heterogeneous ranking outcomes. The resulting baseline ranking serves as a reference against which more sophisticated, data-driven aggregation methods are evaluated in subsequent sections.

3.5. Eigenvectors-Based Weighted Consensus Ranking (EWCR)

The weighted consensus ranking exploits the covariance structure of the rank matrix R. Let

$$\Sigma = \frac{1}{n_f} \bar{R}^\top \bar{R} \tag{3}$$

denote the sample covariance matrix among the model–year rankings, where $\bar{R}$ is the centralized version of R. Principal Component Analysis (PCA) is applied to Σ to extract its dominant modes of variation:

$$\Sigma = V \Lambda V^\top, \tag{4}$$

where $\Lambda = [\lambda_1, \ldots, \lambda_{n_{my}}]$ contains eigenvalues in decreasing order and V the associated eigenvectors. Empirically, the first principal component (PC1) captures most of the shared variation across model–year rankings, whereas later components tend to reflect noise, idiosyncratic model effects, or year-specific deviations. Accordingly, EWCR constructs interpretable, nonnegative weights from PC1 loadings:

$$w_j = \frac{|v_{1j}|}{\sum_{t=1}^{n_{my}} |v_{1t}|}, \qquad j = 1, \ldots, n_{my}, \tag{5}$$

ensuring $w_j \geq 0$ and $\sum_j w_j = 1$. The resulting consensus score for factor i is

$$\text{EWCR\_score}_i = \sum_{j=1}^{n_{my}} w_j R_{ij}. \tag{6}$$

When PC1 does not dominate the covariance structure (e.g., $\lambda_1 / \sum_{r=1}^{n_{my}} \lambda_r < 0.60$), it is preferable to incorporate the first M principal components that collectively explain a proportion $p \in (0, 1)$ of the total variance:

$$\frac{\sum_{m=1}^{M} \lambda_m}{\sum_{r=1}^{n_{my}} \lambda_r} \geq p. \tag{7}$$

Weights are then constructed by combining absolute loadings across the retained components:

$$w_j = \sum_{m=1}^{M} \left(\frac{\lambda_m}{\sum_{r=1}^{M} \lambda_r} \right) \left(\frac{|v_{mj}|}{\sum_{t=1}^{n_{my}} |v_{mt}|} \right), \qquad j = 1, \dots, n_{my}. \tag{8}$$

This multi-component extension preserves secondary but meaningful agreement patterns while downweighting noise-dominated variation.

3.6. Extension to Nonnegative PCA (NPCA)

While PCA provides a data-driven weighting scheme, its standard formulation allows for negative loadings, which can result in counter-intuitive 'subtractive' contributions from certain models. To ensure the consensus remains physically interpretable and that all computational architectures contribute constructively, we employ Non-negative Principal Component Analysis (NNPCA). This constraint yields loading vectors with strictly non-negative coefficients, ensuring the aggregate ranking is a true additive representation of model agreement. The first Non-negative Principal Component Analysis (NPCA) loading vector $v^{(1)} \in \mathbb{R}^{n_{my}}$ is defined as the solution to the following constrained optimization problem:

$$\begin{aligned} \underset{v \in \mathbb{R}^{n_{my}}}{\text{maximize}} \quad & v^\top \Sigma v \\ \text{subject to} \quad & \|v\|_2 = 1 \\ & v \geq 0 \end{aligned} \tag{9}$$

By imposing the non-negativity constraint $v \geq 0$, we ensure that the resulting principal component is a positive combination of the original model rankings. Unlike standard PCA, which may assign negative loadings that mathematically "cancel out" model agreement, NPCA ensures that each computational architecture contributes constructively to the consensus. This is particularly vital in safety research to maintain the physical interpretability of the weighted index. Additional non-negative components $v^{(m)}$ can be obtained via sequential deflation of the covariance matrix Σ. Let $v^{(1)}$ denote the first non-negative loading vector. The unnormalized weights are defined as:

$$\tilde{w}_j = v_j^{(1)}, \qquad j = 1, \dots, n_{my} \tag{10}$$

and normalization yields:

$$w_j = \frac{\tilde{w}_j}{\sum_{t=1}^{n_{my}} \tilde{w}_t} \tag{11}$$

Because $v^{(1)}$ is constrained to be non-negative, the weights w_j represent the proportional contribution of each model–year to the consensus, ensuring that no model diminishes the aggregate score through subtractive weighting. If multiple components are retained to capture a higher proportion of explained variance, let

$$\lambda_m = (v^{(m)})^\top \Sigma v^{(m)} \tag{12}$$

denote the variance explained by component m. The consensus weight is then determined following the logic of the standard EWCR formulation.

3.7. TOPSIS-Based Ranking

The Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) provides a complementary consensus ranking by evaluating how closely each factor resembles an ideal best ranking pattern while remaining distant from an ideal worst pattern. In this framework, each of the n_{my} model–year columns of the rank matrix R is treated as a decision criterion, and each collision-related factor represents an alternative to be ranked. Let $w = (w_1, w_2, \ldots, w_{n_{my}})$, denote the nonnegative, normalized column weights derived from the first principal component loading in the EWCR procedure:

$$w_j = \frac{|v_{1j}|}{\sum_{t=1}^{n_{my}} |v_{1t}|}. \tag{13}$$

These weights quantify the contribution of each model–year ranking to the dominant consensus pattern and ensure that TOPSIS incorporates the same PCA-informed structure as EWCR. To eliminate scale differences across model–year rankings, each criterion column is normalized using vector normalization:

$$\tilde{R}_{ij} = \frac{R_{ij}}{\sqrt{\sum_{k=1}^{n_f} R_{kj}^2}}, \qquad i = 1, \ldots, n_f, \quad j = 1, \ldots, n_{my}. \tag{14}$$

The weighted decision matrix is then constructed as

$$V_{ij} = w_j \, \tilde{R}_{ij}, \tag{15}$$

so that criteria more strongly aligned with the dominant PCA consensus pattern exert greater influence on the resulting ranking. Because lower ranks indicate greater importance, the ideal best and ideal worst values for each criterion are defined as

$$v_j^+ = \min_i V_{ij}, \qquad v_j^- = \max_i V_{ij}. \tag{16}$$

The ideal best corresponds to consistently low (highly important) normalized ranks, whereas the ideal worst corresponds to consistently high ranks. For each factor i, TOPSIS computes Euclidean distances to these reference profiles:

$$D_i^+ = \sqrt{\sum_{j=1}^{n_{my}} \left(V_{ij} - v_j^+\right)^2}, \qquad D_i^- = \sqrt{\sum_{j=1}^{n_{my}} \left(V_{ij} - v_j^-\right)^2}. \tag{17}$$

The TOPSIS score for factor i is defined as its relative closeness to the ideal best:

$$C_i = \frac{D_i^-}{D_i^- + D_i^+}, \tag{18}$$

with larger values indicating factors that are closer to the ideal (i.e., consistently highly ranked) across all model–year combinations. The final ranking is obtained as

$$\text{TOPSIS\_rank}_i = \text{rank}(C_i), \tag{19}$$

where rank 1 indicates the factor closest to the ideal consensus pattern. This TOPSIS formulation provides a multi-criteria perspective on rank aggregation that is directly aligned with the PC1-based consensus structure used in EWCR, thereby offering a complementary yet coherent measure of global factor importance.

3.8. Normalized Weighted Consensus Ranking (NWCR)

The Normalized Weighted Consensus Ranking (NWCR) provides a third method for aggregating the importance of each collision-related factor across all model–year combinations. In contrast to the PCA-based weighting used in EWCR or the multi-criteria distance evaluation of TOPSIS, NWCR is based on a direct normalization of raw ranks followed by an simple averaging method. This approach emphasizes simplicity and interpretability while preserving the central tendency of feature importance patterns across models and time periods.

Let $R \in \mathbb{R}^{n_f \times n_{my}}$ denote the rank matrix. To place all columns on a common scale, each column is normalized using min–max scaling:

$$R_{ij}^{\text{norm}} = \frac{R_{ij} - \min_k R_{kj}}{\max_k R_{kj} - \min_k R_{kj}}, \qquad i = 1, \dots, n_f, \quad j = 1, \dots, n_{my}. \tag{20}$$

This transformation maps all rank values to the interval $[0, 1]$, where lower normalized values indicate factors that appear near the top of a model's ranking and higher values correspond to factors judged less important. Normalization ensures comparability across columns even when the underlying rank distributions differ. For each factor i, the NWCR score is computed as the arithmetic mean of its normalized ranks across all model–year combinations:

$$\text{NWCR\_score}_i = \frac{1}{n_{my}} \sum_{j=1}^{n_{my}} R_{ij}^{\text{norm}}. \tag{21}$$

Lower values of this score correspond to factors that consistently receive strong importance rankings across multiple models and years, whereas higher values indicate factors that are less influential on average. The final NWCR ranking is obtained by ordering factors in ascending order of their consensus scores:

$$\text{NWCR\_rank}_i = \text{rank}\left(\text{NWCR\_score}_i\right). \tag{22}$$

Although it is methodologically simpler than both EWCR and TOPSIS approachers, NWCR provides a stable,interpretable baseline that reflects the average normalized importance of each factor. Its role as a contrast to the PCA-guided and distance-based consensus formulations helps illuminate structural patterns that persist across the different aggregation approaches.

3.9. NMF-Based Weighted Consensus Ranking

In addition to the normalized averaging procedure used in NWCR, a more expressive consensus ranking method was constructed using Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000). This approach decomposes the rank matrix into a set of latent components that summarize recurring patterns of factor importance across the full collection of model–year rankings. These components are then weighted to produce a structured global consensus. To prepare the data for factorization, each column of the rank matrix R is rescaled using columnwise max normalization:

$$R_{ij}^{\text{scaled}} = \frac{R_{ij}}{\max_k R_{kj}}. \tag{23}$$

This is to ensure that all entries lie within $[0, 1]$, and the differences in absolute rank ranges do not influence the decomposition. The algorithm used for NMF then seeks a low-rank, non-negative approximation of the matrix R, represented as follows

$$R^{\text{scaled}} \approx WH, \tag{24}$$

where $W \in \mathbb{R}_+^{n_f \times K}$ contains factor loadings and $H \in \mathbb{R}_+^{K \times n_{my}}$ contains the contribution of each latent component to each of the model–year rankings. In this study, $K = 3$ components were selected to balance interpretability, reconstruction accuracy, and computational stability. The model was initialized using the Non-negative Double Singular Value Decomposition (NNDSVD), with reconstruction error assessed using the Frobenius norm $\|R^{\text{scaled}} - WH\|_F$.

The columns of W represent distinct latent patterns of factor relevance that recur across models and years, while each row of H quantifies how strongly a given latent component influences each model–year ranking. Components that exhibit larger total mass in W are interpreted as contributing more strongly to the global structure. Accordingly, the raw weight of component k is defined as the sum of its loadings,

$$\alpha_k = \sum_{i=1}^{n_f} W_{ik}, \tag{25}$$

and normalized to produce a set of nonnegative component weights,

$$\tilde{\alpha}_k = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}, \qquad k = 1, \dots, K. \tag{26}$$

The final NMF-based consensus score for factor i is then computed as the weighted combination of its loadings across the K latent components:

$$\text{NMF\_score}_i = \sum_{k=1}^{K} W_{ik}\,\tilde{\alpha}_k. \tag{27}$$

Factors with larger scores exert greater influence across the latent structures and are therefore assigned higher consensus importance. Sorting these scores in descending order produces the final NMF-based consensus ranking. This method complements EWCR and TOPSIS by capturing deeper, parts-based structure in the multi-model, multi-year ranking matrix. Whereas EWCR emphasizes shared linear structure and TOPSIS emphasizes geometric idealness, NMF identifies interpretable latent patterns across the ensemble of models and years.

3.10. Sparsity-Controlled Consensus Ranking

While PCA-based EWCR provides a global consensus by leveraging shared variance across all model–year rankings, it assigns nonzero weights to every column and therefore does not explicitly distinguish between strongly informative and weak or redundant model–year contributions. To further refine the interpretation of factor importance, we extend the EWCR framework using a sparsity-controlled non-negative consensus approach. This procedure explicitly limits the number of model–year rankings contributing to the final consensus, thereby enabling a principled separation between major risk factors and those whose importance depends on broader aggregation. Let Σ denote the covariance matrix of the model–year rank columns, and let $v \in \mathbb{R}^{n_{my}}_{\geq 0}$ represent a non-negative consensus loading vector obtained by iterative variance maximization under non-negativity constraints. Sparsity is imposed through a top-k truncation operator, which retains only the k largest entries of v and sets all remaining entries to zero. For a specified sparsity proportion $\rho \in (0, 1]$, the number of retained model–year rankings is defined as $k = \lceil \rho \cdot n_{my} \rceil$. The sparse loading vector is constructed as

$$v_j^{(\rho)} = \begin{cases} v_j, & j \in \text{top-}k(v); \\ 0, & \text{otherwise,} \end{cases}$$

where top-$k(v)$ denotes the indices of the k largest elements of v. The resulting vector is normalized to obtain valid consensus weights

$$w_j^{(\rho)} = \frac{v_j^{(\rho)}}{\sum_{t=1}^{n_{my}} v_t^{(\rho)}}, \qquad \sum_{j=1}^{n_{my}} w_j^{(\rho)} = 1.$$

This formulation provides direct and interpretable control over sparsity. For instance, a 30% setting retains approximately 30% of model–year rankings, while 20% and 10% settings enforce progressively stronger sparsity by concentrating the consensus on fewer agreement sources. As sparsity increases, a larger proportion of weights become exactly zero, forcing the consensus ranking to rely only on the strongest and most stable cross-model and cross-year signals. Using these sparse weights, factor-level consensus scores are computed as

$$\text{EWCR}_i^{(\rho)} = \sum_{j=1}^{n_{my}} R_{ij}\, w_j^{(\rho)},$$

and factors are re-ranked accordingly. Factors that remain highly ranked under increasing sparsity constraints can therefore be interpreted as core determinants of fatal collision severity.

Throughout the empirical analysis, the Eigenvectors-Weighted Consensus Ranking is treated as the primary basis for substantive interpretation and policy-relevant conclusions. The alternative aggregation approaches (TOPSIS, NWCR, and NMF-based rankings) are used to assess robustness and stability of factor prioritization across different methodological perspectives, rather than as independent bases for inference.

4. Results

4.1. Descriptive Importance Patterns under the Simple Average Ranking Method

To jointly examine baseline feature importance and stability, Figure 2 plots each factor's mean rank against the standard deviation of its rank across the 24 model–year combinations. This representation provides a direct

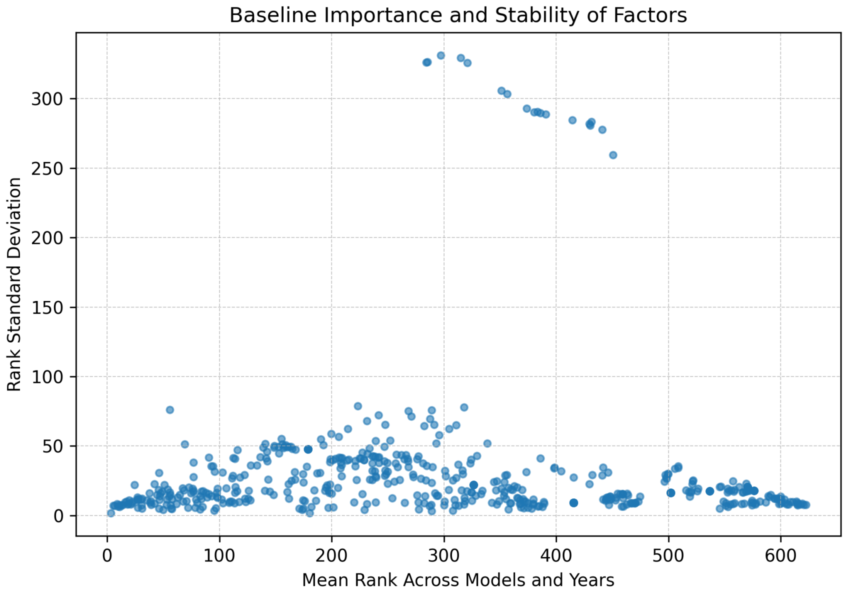


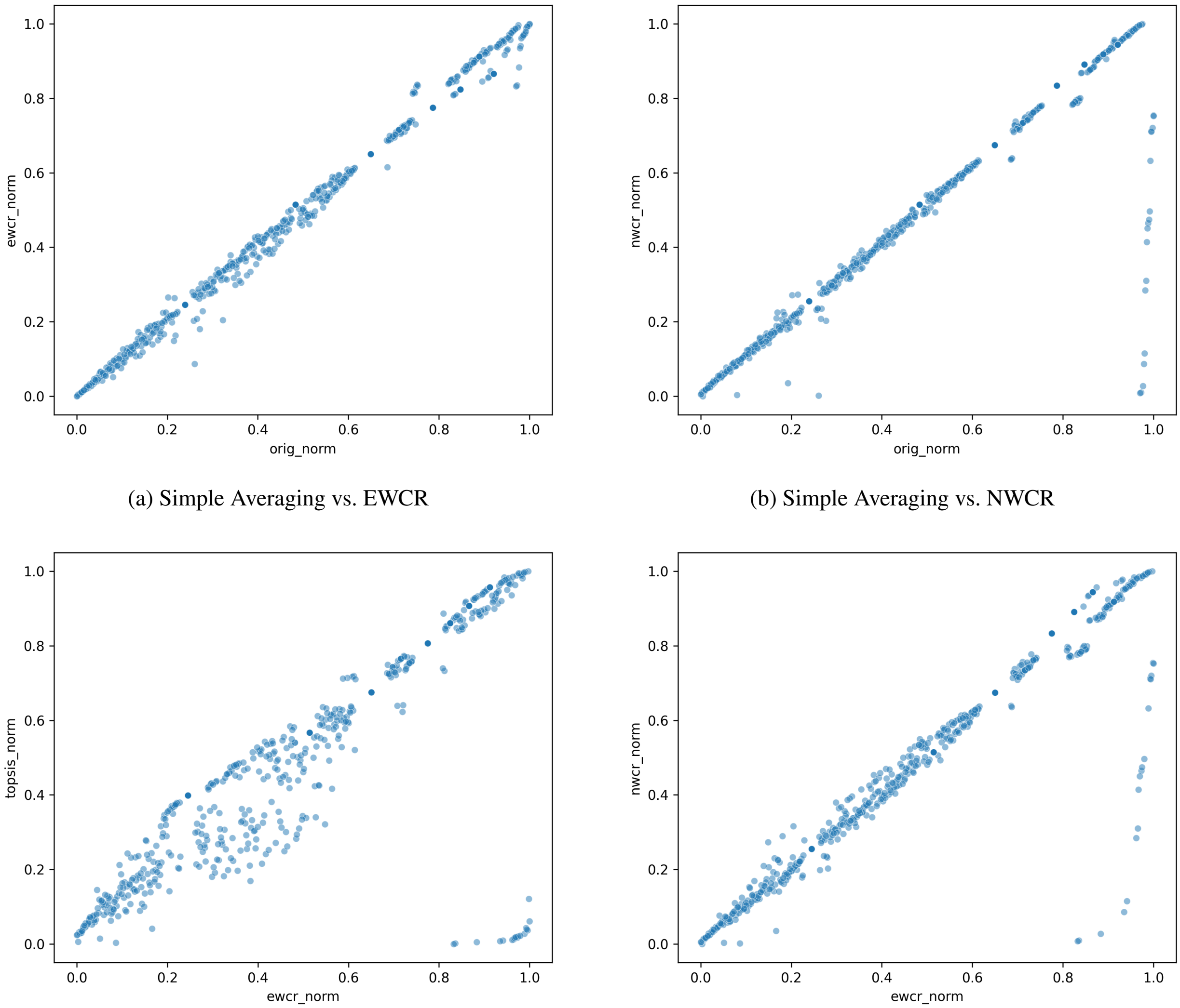
Figure 2: Baseline feature importance and stability of collision-related factors. Each point represents a factor level, plotted by its mean rank across all model–year combinations and the standard deviation of its rank. Lower mean ranks and lower variability indicate factors that are both highly influential and stable across models and years.

visualization of how consistently each factor is prioritized across different modelling approaches and time periods. The distribution of mean ranks reveals a strongly right-skewed structure in which a relatively small subset of factors demonstrates persistently high importance across multiple models and years, while a long tail of factors exhibits consistently lower relevance. In Figure 2, the highly influential variables appear in the lower-left region of the plot, characterized by both low mean rank and low rank variability. Many of these stable, high-importance factors correspond to broad crash descriptors such as collision configuration categories, vehicle type classes, and injury severity indicators. Their position in this region reflects strong and consistent predictive signals shared across tree-based models (RF and XGB) and neural networks. In contrast, a substantial portion of factors exhibits markedly higher variability. In Figure 2, the variables populate the central and upper regions of the plot, indicating moderate-to-high mean ranks coupled with large dispersion. This instability is particularly pronounced for neural network importance estimates, where SHAP-based attribution often spreads importance more diffusely across correlated predictors,leading to elevated rank volatility.

4.2. Cross-Method Concordance

This section evaluates the extent to which the four ranking procedures, the simple averaging method, EWCR, TOPSIS, and NWCR, produce consistent assessments of factor importance. All methods are placed on a common scale that facilitates direct comparison of their behaviour.The analysis proceeds by examining several complementary indicators of agreement. Figure 3 presents four pairwise scatterplots illustrating how the normalized rankings relate to one another after the correction of the EWCR weights. The comparison between the simple averaging method and EWCR (Panel a) shows a very tight diagonal structure, reflecting the fact that PC1 loadings produce a weighting scheme that preserves most of the global rank ordering while moderating the influence of noisy model–year columns. Only a small number of mid-ranked factors deviate substantially from the diagonal, indicating that EWCR now acts as a variance-informed refinement of the simple averaging approach.

The simple averaging versus NWCR comparison (Panel b) continues to exhibit the strongest concordance among all methods, forming an almost perfect line. This behaviour is expected because NWCR is an average of min–max normalized ranks and therefore represents a smoothed version of the original ranking. The near-perfect agreement indicates that NWCR primarily compresses the rank scale without altering the underlying ordering of factors. The EWCR versus TOPSIS plot (Panel c) displays visibly greater dispersion, particularly in the mid-range. TOPSIS relies



(a) Simple Averaging vs. EWCR

(b) Simple Averaging vs. NWCR

(c) EWCR vs. TOPSIS

(d) EWCR vs. NWCR

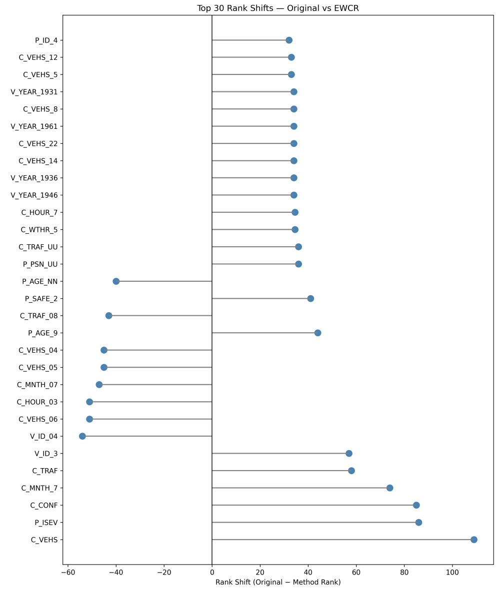
Figure 3: Pairwise scatterplots comparing normalized rankings across the four importance estimation methods. Strong diagonal patterns indicate high concordance, while dispersion reflects methodological differences in how structural or geometric information is extracted from the rank matrix.

on geometric distance to idealized profiles, which amplifies differences among factors whose normalized rankings place them near equidistant positions between best and worst-case patterns. As a result, TOPSIS applies more aggressive reordering of mid-ranked factors compared to EWCR, even when the extremes remain highly concordant.

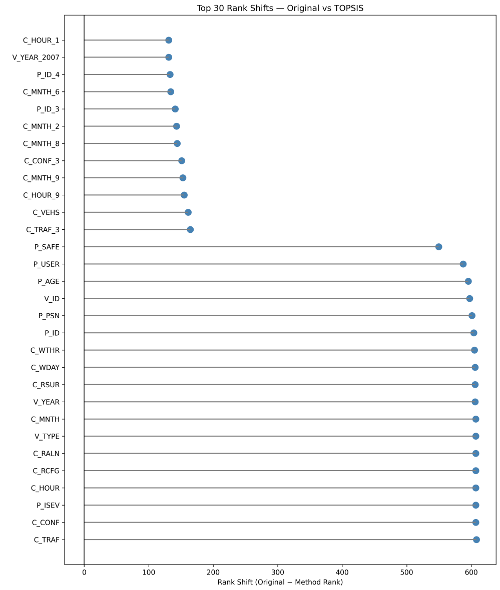
Finally, the EWCR versus NWCR comparison (Panel d) demonstrates moderate linear agreement with noticeable curvature and clustering. This pattern highlights a fundamental distinction between the two approaches: NWCR reflects an average normalized level of importance, whereas EWCR emphasizes shared variance across the 24 model–year rankings. Thus, factors whose importance varies considerably across models or time tend to show the largest divergence between EWCR and NWCR.

4.3. Rank-Shift Diagnostics and Methodological Divergence

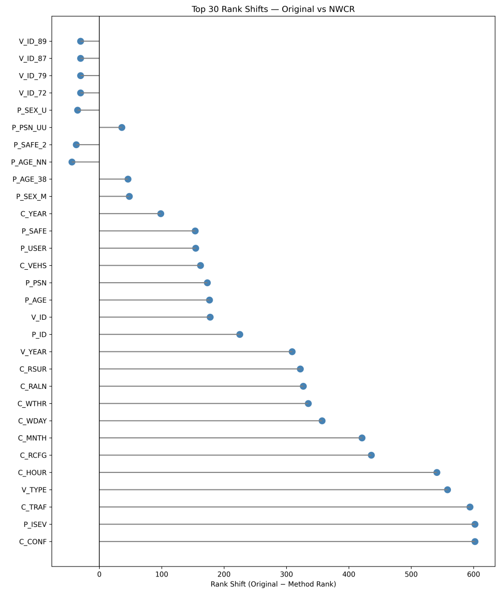
To quantify how each consensus ranking method reshapes the baseline ordering implied by the simple averaging procedure, a rank-shift statistic was computed for every factor level. For a given method m and factor level i, the shift is defined as $\Delta_i^{(\mathrm{simp},m)} = \mathrm{rank}_{i,\mathrm{simp}} - \mathrm{rank}_{i,m}$, where positive values indicate that method m assigns a worse (higher)



(a) Top 30 Rank Shifts—Simple Averaging vs. EWCR



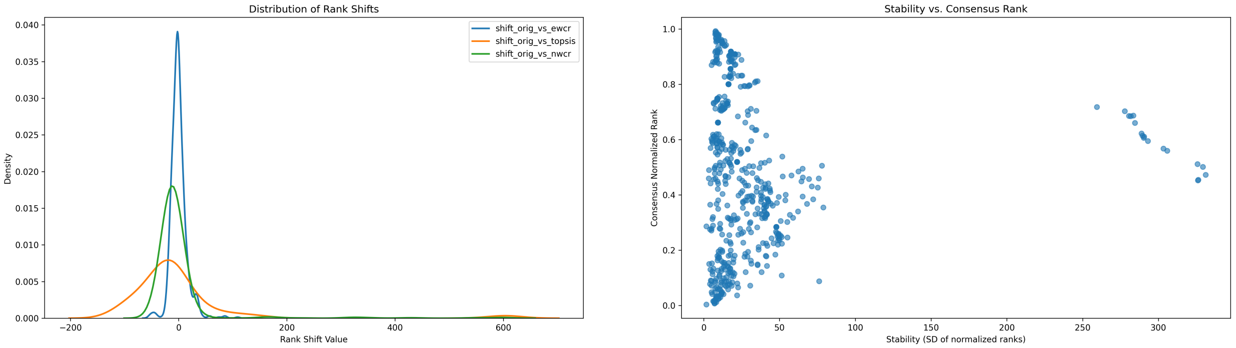
(b) Top 30 Rank Shifts—Simple Averaging vs. TOPSIS



(c) Top 30 Rank Shifts—Simple Averaging vs. NWCR

Figure 4: Largest absolute rank shifts between the original ranking and each consensus method. Positive values indicate demotion; negative values indicate promotion relative to the original ranking.

rank to factor i than the simple averaging approach, and negative values indicate an upward promotion. This measure identifies the factors whose importance assessments are most sensitive to methodological choices. Figure 4 presents the thirty largest absolute shifts for the comparisons. The shifts between the simple average ranking and EWCR are generally modest. Most factors move only slightly, reflecting that EWCR primarily smooths noise-dominated variation while preserving the global structure. Larger deviations occur for factors associated with model–year rankings that



(a) Kernel density estimates of rank–shift distributions.

(b) Stability (standard deviation across methods) vs. consensus normalized rank.

Figure 5: Rank–shift behaviour (left) and stability dynamics (right) across consensus methods.

load strongly on secondary principal components, which EWCR down-weights through its PC1-based weighting mechanism.

In contrast, the simple averaging versus TOPSIS comparison exhibits substantially larger and more irregular shifts. This behaviour reflects TOPSIS's geometric foundation. Factors located near the midpoint between the ideal-best and ideal-worst profiles may experience amplified reordering when Euclidean distances dominate over raw rank magnitudes. As a result, TOPSIS reshapes mid-range factors more aggressively than either EWCR or NWCR. The shifts produced by NWCR are the smallest overall. Because NWCR is simply the average of min–max normalized ranks, it acts as a smoothed version of the simple averaging rank procedure rather than a structural reweighting or distance-based transformation. Consequently, NWCR exhibits minimal reordering, with nearly all shifts confined to small adjustments around the simple averaging ranking.

Figure 5 summarizes how the three consensus ranking methods diverge from the simple average consensus ranking and how these deviations relate to cross-method stability. The left panel displays kernel density estimates of the rank–shift distributions, defined as $\Delta_i^{(\text{simp},k)} = \text{rank}_{i,\text{simp}} - \text{rank}_{i,k}$, for each consensus ranking method $k \in \{\text{EWCR}, \text{TOPSIS}, \text{NWCR}\}$. NWCR exhibits the sharpest, most concentrated peak around zero, confirming that it behaves as a smoothed, min–max–normalized analogue of the original mean ranking with minimal reordering. EWCR shows a slightly wider distribution, reflecting the influence of PCA loading weights that down-weight noisy model–year columns and modestly reshape mid-ranked factors. TOPSIS displays the broadest spread, with long tails extending well beyond ± 200, indicating substantial reordering introduced by its distance-to-ideal formulation, particularly for factors positioned near geometric midpoints between best and worst ranking profiles. The right panel plots factor stability, quantified as the standard deviation of normalized ranks across all four methods, against the consensus normalized rank. A clear triangular pattern emerges. Some highly influential factors cluster tightly at low variability, demonstrating strong cross-method agreement, whereas less important variables exhibit increasingly large dispersion. This widening band reflects inconsistent treatment across methods, typically arising from long-tailed categorical variables, sparse factor levels, or predictors whose importance is sensitive to model-specific non-linearities. A small number of extreme instability outliers appear at the far right, corresponding to factors that differ markedly in how tree-based, neural network, PCA-weighted, and distance-based methods interpret their relevance.

4.4. Latent Consensus Structure Revealed by NMF

NMF was applied to the max scaled rank matrix R_{scaled} to produce its low rank approximation by retaining three latent components. The Frobenius reconstruction error was small, indicating that three latent components capture the dominant cross models and cross years ranking structure. In Table 2, the first component (Comp. 1) loads heavily on a set of crash configuration and vehicle-type indicators (e.g., `C_RCFG_02`, `V_TYPE_08`, `C_VEHS_01`), reflecting a dominant pattern of global severity relevance common across nearly all model–year combinations. The second component (Comp. 2) captures a coherent block of person-level identifiers (e.g., `P_ID_32`–`P_ID_37`) together with the general vehicle count variable `C_VEHS`, indicating a secondary structure associated with demographic and contextual attributes. The third component (Comp. 3) highlights temporal and roadway-related predictors (`C_HOUR_16`,

Table 2
Top six highest-loading factors for each NMF component (from matrix W).

Factor	Comp. 1	Comp. 2	Comp. 3
C_RCFG_02	0.253965	–	0.506654
V_TYPE_08	0.253397	–	0.506439
C_VEHS_01	0.253374	–	0.506352
C_MNTH_03	0.250693	–	0.503504
P_SAFE_02	0.250500	–	0.503097
V_TYPE_07	0.250075	–	0.502875
C_VEHS	–	0.360897	–
P_ID_32	–	0.322916	0.430696
P_ID_33	–	0.322916	0.430696
P_ID_35	–	0.322916	0.430696
P_ID_36	–	0.322916	0.430696
P_ID_37	–	0.322916	0.430696
C_HOUR_16	0.064819	–	0.514034
C_RALN_2	0.090330	–	0.512330
C_WDAY_6	0.064055	–	0.511222
C_CONF_31	0.015475	–	0.507885
C_CONF_21	0.000000	–	0.507781
P_ISEV_3	0.000000	–	0.507645

C_RALN_2, C_WDAY_6) along with selected configuration indicators (C_CONF_31, C_CONF_21). These patterns confirm that NMF successfully separates a dominant global severity axis from secondary axes representing model-dependent and temporal-context effects.

The model–year loadings contained in the H matrix, whose partial results are displayed in Table 3, describe how strongly each latent component contributes to the factor level ranking of each model–year combination. Table 3 provides an excerpt of these values for representative neural network (NN), random forest (RF), and XGBoost (XGB) models. These three components exhibit distinct activation patterns across methods and years. Component 1 is highly sparse, with many zero entries, and shows large but irregular spikes in selected NN, RF, and XGB years. This behaviour indicates that the first latent component captures a set of narrow, model-specific effects rather than a global ranking structure. Component 2 exhibits moderate magnitudes across nearly all model–year combinations, with values typically between 0.45 and 3.25. Its consistent expression across NN, RF, and XGB suggests that this component represents a broad shared pattern in the feature importance rankings. Although not the dominant structure, Component 2 serves as a stable mid-level contributor that reflects agreement across methods. Component 3 displays the most pronounced and systematic behaviour. For nearly all RF and XGB years, its loadings cluster tightly around values of 1.95–1.97, forming a strong and coherent block of activation. NN models also load heavily onto this component in earlier years (2014–2016), but the strength of this activation declines sharply in later years for the NN approach, where values approach zero. This contrast shows that Component 3 captures the dominant global structure in the feature importance patterns, particularly for tree-based models.

4.5. Effect of Sparsity on Identification of Major Risk Factors

When risk factors exhibit weak or negligible associations with fatal collisions, their relative ranking becomes less consequential. Consequently, imposing sparsity on the rank matrix facilitates the identification of the most influential risk factors by filtering out those with marginal effects. Table 4 summarizes the selection outcomes of the ten most important risk factors identified in the overall consensus ranking. For each factor level, the table indicates whether it appears among the top-ranked set under the baseline PCA-based EWCR and under each sparsity-controlled configuration (30%, 20%, and 10%). This comparison illustrates how increasing sparsity progressively filters out factors whose importance depends on broader aggregation, while retaining those supported by the strongest agreement structure.

From Table 4, we observe that a core group of collision-configuration variables, including C_CASE and multiple C_CONF levels, remains selected across all sparsity controls. Their persistence even under the strictest sparsity constraint

Table 3
Excerpt of the NMF H matrix for selected model–year combinations.

Model–Year	Comp. 1	Comp. 2	Comp. 3
NN 2014	0.00349	0.46751	1.96414
NN 2015	0.01470	0.46516	1.96197
NN 2016	0.04767	0.48057	1.94848
⋮	⋮	⋮	⋮
RF 2014	0.00000	0.45129	1.97221
RF 2015	0.00000	0.45141	1.97165
RF 2016	0.02196	0.46718	1.96075
⋮	⋮	⋮	⋮
XGB 2014	0.00000	0.45696	1.96946
XGB 2015	0.00060	0.47863	1.96082
XGB 2016	0.02590	0.47404	1.95648

Table 4
Selection of top-ranked risk factors under increasing sparsity constraints. Checkmarks indicate that a factor level appears among the top-ranked set for the corresponding consensus method.

Factor Level	**PCA**	**NSPCA 30%**	**NSPCA 20%**	**NSPCA 10%**
C_CASE	✓	✓	✓	✓
C_CONF	✓	✓	✓	✓
C_CONF_06	✓	✓	✓	✓
C_CONF_21	✓	✓	✓	✓
C_CONF_22	✓	✓	✓	✓
C_HOUR_15	✓	✓		
C_TRAF_18	✓	✓		
P_ISEV_3	✓	✓		
C_RSUR_2	✓			
P_USER_2	✓			

indicates that these variables represent major risk factors whose importance is independent of specific model–year contributions and reflects fundamental aspects of crash. In contrast, temporal, traffic-control, and injury-related variables such as `C_HOUR_15`, `C_TRAF_18`, and `P_ISEV_3` are retained only under moderate sparsity and are excluded as sparsity increases. These variables can be interpreted as secondary risk factors, whose influence is supported by multiple models but diminishes when the consensus is restricted to the strongest agreement features.

4.6. Summary of Key Risk Factors and Implication on Road Safety

To identify collision-related variables that demonstrate robust and method-independent importance, we examine overall agreement across all four ranking procedures, including the simple averaging method, EWCR, TOPSIS, and NWCR. Using the combined normalized ranking table, we construct a consensus set of the top 50 factors based on their mean rank across methods. This set represents variables that remain highly ranked regardless of the specific aggregation strategy applied, and therefore constitutes the most stable and reliable group of predictors in the analysis.

Across this top-50 consensus ranking group, strong agreement is observed among the four methods, indicating that multiple, fundamentally different ranking frameworks converge on a common set of influential collision factors. This result demonstrates that combining model diversity, temporal replication, and multiple consensus ranking methods produce a promising and robust identification of key risk factors. Within this broader agreement set, Table 5 presents a summary of the ten highest-ranked predictors based on the mean of the Original, EWCR, TOPSIS, and NWCR rankings. These variables represent the most influential factors within the overall consensus ranking and exhibit both high importance and strong cross-method stability.

The results reveal a coherent set of behavioural, collision-configuration, and environment-related predictors. Collision characteristics such as `C_CASE`, `C_CONF_21`, and `C_CONF_31` consistently rank among the most critical

Table 5
Top 10 most important risk factors based on mean consensus ranking across Original, EWCR, TOPSIS, and NWCR.

Factor Level	Simp.	EWCR	TOPSIS	NWCR	Mean	Std
C_CASE	4	3	5	1	3.25	1.71
P_ISEV_3	1	1	16	4	5.50	7.14
C_CONF_21	2	2	17	5	6.50	7.14
C_CONF_31	3	4	21	8	9.00	8.29
C_TRAF_18	5	5	19	10	9.75	6.60
C_HOUR_15	6	6	22	9	10.75	7.63
P_ISEV_2	7	7	20	11	11.25	6.13
P_USER_2	8	8	23	12	12.75	7.09
C_RALN_U	9	9	25	13	14.00	7.57
C_RSUR_2	10	10	30	14	16.00	9.52

factors, reflecting their direct relationship to crash mechanics and impact severity. Injury severity indicators (P_ISEV_2, P_ISEV_3) and roadway or environmental conditions (e.g., C_RALN_U, C_RSUR_2) also appear prominently and display relatively low rank variability across methods. The standard deviation values for these factors are generally small, indicating high methodological concordance and reinforcing their reliability as key determinants of fatal collision risk.

The importance of collision configuration, injury severity, and roadway or environmental conditions identified in this study is consistent with recent Canadian road safety research, which similarly highlights crash mechanics, roadway alignment, and surface conditions as robust predictors of fatal outcomes (e.g., McCullogh, Macpherson, Hagel, Giles, Fuselli, Pike, Torres and Richmond (2023); McCullogh, Macpherson, Harrington, Pike, Hagel, Buchan, Fuselli and Richmond (2025)). Unlike prior studies that rely on single-model or single-period analyses, this work contributes methodologically by applying a consensus ranking framework that synthesizes evidence across models and time, yielding more stable and policy-relevant risk factor prioritization.

From a policy perspective, the consistent ranking of these factors supports prioritizing interventions that target high-risk crash configurations and hazardous infrastructure. Measures such as redesigning conflict-prone intersections, improving roadway alignment and surface conditions, and strengthening protections in severe-injury crash scenarios should form the core of evidence-based road safety strategies. The low variability in rankings further justifies their use in regulatory standards, infrastructure investment decisions, and targeted enforcement aimed at reducing fatal collisions.

5. Conclusions and Future Work

This study proposes a consensus-ranking framework designed to address the variability and model dependence commonly encountered in machine-learning analyses of fatal-collision risk factors. Using eight years of data from the Canadian National Collision Database (2014–2021) and an ensemble of Random Forest, XGBoost, and neural network models, the analysis moves beyond single-model importance measures toward a unified and transparent ranking of predictors. The proposed Eigenvectors-Weighted Consensus Ranking (EWCR), together with TOPSIS- and NPCA-based extensions, provides a principled approach for identifying risk factors that are consistently influential across models and over time. The results reveal a stable core of high-priority predictors, most notably collision configuration, traffic control, and roadway and lighting conditions that persist across modeling approaches and temporal disruptions, including the COVID-19 period. From a road safety perspective, this stability supports prioritizing systemic interventions over approaches focused solely on individual driver behaviour. In particular, the consistent importance of intersection-related and environmental factors reinforces the relevance of safe system and environmental risk mitigation as central strategies for reducing fatal injuries. Agreement and rank-stability analyses further demonstrate that the most influential predictors are robust to methodological choice, enhancing the inferential reliability of the findings for policy and planning purposes. The proposed framework offers a defensible and transparent basis for evidence-based regulation, infrastructure investment, and targeted safety interventions.

Future research should extend this framework by incorporating traffic exposure and enforcement data to better disentangle crash risk from injury severity, and by developing spatio-temporal consensus methods to capture regional heterogeneity in risk stability. As mobility systems evolve and new vehicle technologies are introduced, the proposed longitudinal consensus approach provides a valuable tool for monitoring the persistence of risk factors and evaluating the long-term effectiveness of road safety policies.

Acknowledgement

The authors acknowledge partial support from Natural Sciences and Engineering Research Council of Canada.

References

Abbas, G., Ying, Z., Ayoubi, M., 2025. Consensus-driven feature selection for transparent and robust loan default prediction. Scientific Reports .

Akter, R., Susilawati, S., Zubair, H., Chor, W.T., 2025. Analyzing feature importance for older pedestrian crash severity: A comparative study of dnn models, emphasizing road and vehicle types with shap interpretation. Multimodal Transportation 4, 100203.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention 34, 729–741.

Ali, Y., Hussain, F., Haque, M.M., 2024. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. Accident Analysis & Prevention 194, 107378.

Amiri, M.A., Afshari, S., Soltani, A., 2025. Machine learning approaches to traffic accident severity prediction: Addressing class imbalance. Machine Learning with Applications , 100792.

Asgarzadeh, M., Fischer, D., Verma, S.K., Courtney, T.K., Christiani, D.C., 2018. The impact of weather, road surface, time-of-day, and light conditions on severity of bicycle-motor vehicle crash injuries. American journal of industrial medicine 61, 556–565.

Elyassami, S., Hamid, Y., Habuza, T., 2021. Road crashes analysis and prediction using gradient boosted and random forest trees, in: 2020 6th IEEE Congress on Information Science and Technology (CiSt), IEEE. pp. 520–525.

Ghandour, A.J., Hammoud, H., Al-Hajj, S., 2020. Analyzing factors associated with fatal road crashes: a machine learning approach. International journal of environmental research and public health 17, 4111.

Golestani, A., Rezaei, N., Malekpour, M.R., Ahmadi, N., Ataei, S.M.N., Khosravi, S., Jafari, A., Shahraz, S., Farzadfar, F., 2025. Predicting errors in accident hotspots and investigating satiotemporal, weather, and behavioral factors using interpretable machine learning: An analysis of telematics big data. Plos one 20, e0326483.

Gutiérrez-Rodríguez, R., Rojí, E., Cuadrado, J., 2025. Identifying relevant patterns between injury crashes and road safety inspection deficiencies. Journal of Safety Research 93, 99–134.

Gyawali, P.K., Liu, X., Zou, J., He, Z., 2022. Ensembling improves stability and power of feature selection for deep learning models, in: Machine Learning in Computational Biology, PMLR. pp. 33–45.

Hamdan, N., Sipos, T., 2025. Advancements in machine learning for traffic accident severity prediction: A comprehensive review. Periodica Polytechnica Transportation Engineering 53, 347–355.

Huang, Y., Zhou, Q., Koelper, C., Li, Q., Nie, B., 2020. Are riders of electric two-wheelers safer than bicyclists in collisions with motor vehicles? Accident Analysis & Prevention 134, 105336.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention 108, 27–36.

Islam, M.R., Abdel-Aty, M., Islam, Z., Abdelraouf, A., 2024. Real-time framework to predict crash likelihood and cluster crash severity. Transportation research record 2678, 202–217.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D., Ahmad, M., 2021. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. International journal of injury control and safety promotion 28, 408–427.

Jermakian, J.S., 2011. Crash avoidance potential of four passenger vehicle technologies. Accident Analysis & Prevention 43, 732–740.

Lee, D., Seung, H.S., 2000. Algorithms for non-negative matrix factorization. Advances in neural information processing systems 13.

Leoni, L., BahooToroody, A., Abaei, M.M., Cantini, A., BahooToroody, F., De Carlo, F., 2024. Machine learning and deep learning for safety applications: Investigating the intellectual structure and the temporal evolution. Safety science 170, 106363.

Li, H., Graham, D.J., Ding, H., Ren, G., 2019. Comparison of empirical bayes and propensity score methods for road safety evaluation: a simulation study. Accident Analysis & Prevention 129, 148–155.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation research part A: policy and practice 44, 291–305.

Luan, S., Jiang, Z., Yang, X., Meng, F., et al., 2024. Investigating risk factors associated with injury severity in highway crashes: A hybrid approach integrating two-step cluster analysis and latent class ordered regression model with covariates. Accident Analysis & Prevention 208, 107805.

McCullogh, E., Macpherson, A., Hagel, B., Giles, A., Fuselli, P., Pike, I., Torres, J., Richmond, S.A., 2023. Road safety, health equity, and the built environment: perspectives of transport and injury prevention professionals in five canadian municipalities. BMC public health 23, 1211.

McCullogh, E., Macpherson, A., Harrington, D.W., Pike, I., Hagel, B., Buchan, C., Fuselli, P., Richmond, S.A., 2025. Built environment change for injury prevention: insights from public health professionals across public health units in ontario, canada. BMC Public Health 25, 31.

Nara, E.O.B., Sordi, D.C., Schaefer, J.L., Schreiber, J.N.C., Baierle, I.C., Sellitto, M.A., Furtado, J.C., 2019. Prioritization of ohs key performance indicators that affecting business competitiveness–a demonstration based on maut and neural networks. Safety science 118, 826–834.

Obasi, I.C., Benson, C., 2023. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. Heliyon 9.

Pathivada, B.K., Banerjee, A., Haleem, K., 2024. Impact of real-time weather conditions on crash injury severity in kentucky using the correlated random parameters logit model with heterogeneity in means. Accident Analysis & Prevention 196, 107453.

Pei, Y., Wen, Y., Pan, S., 2025. Traffic accident severity prediction based on interpretable deep learning model. Transportation Letters 17, 895–909.

Pes, B., 2020. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Computing and Applications 32, 5951–5973.

Phun, V.K., Kato, H., Yai, T., 2018. Traffic risk perception and behavioral intentions of paratransit users in phnom penh. Transportation research part F: traffic psychology and behaviour 55, 175–187.

Rengasamy, D., Mase, J.M., Torres, M.T., Rothwell, B., Winkler, D.A., Figueredo, G.P., 2021a. Mechanistic interpretation of machine learning inference: A fuzzy feature importance fusion approach. arXiv preprint arXiv:2110.11713 .

Rengasamy, D., Rothwell, B.C., Figueredo, G.P., 2021b. Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. Applied Sciences 11, 11854.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques, in: Joint European conference on machine learning and knowledge discovery in databases, Springer. pp. 313–325.

Saladié, Ò., Bustamante, E., Gutiérrez, A., 2020. Covid-19 lockdown and reduction of traffic accidents in tarragona province, spain. Transportation research interdisciplinary perspectives 8, 100218.

Santos, K., Dias, J.P., Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research 80, 254–269.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis & Prevention 43, 1666–1676.

Shaik, M.E., Ahmed, S., 2022. An overview of the impact of covid-19 on road traffic safety and travel behavior. Transportation Engineering 9, 100119.

Shuai, Z., Kwon, T.J., 2025. Analyzing winter crash dynamics using spatial analysis and crash frequency prediction models with shap interpretability. Future Transportation 5, 17.

Slikboer, R., Muir, S.D., Silva, S., Meyer, D., 2020. A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data. Systematic reviews 9, 220.

Sorum, N.G., Pal, D., 2024. Identification of the best machine learning model for the prediction of driver injury severity. International journal of injury control and safety promotion 31, 360–375.

Srivastava, N., Gohil, B.N., Ray, S., 2024. Road traffic accident severity prediction using causal inference and machine learning, in: Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD), pp. 292–300.

Sun, K., Li, Y., Zeng, Q., Wang, X., 2025. Exploring the factors affecting the injury severity of crashes on mountainous freeways with high proportions of heavy vehicles considering their heterogeneous and interactive effects. Sustainability 17, 1624.

Vanlaar, W.G., Woods-Fry, H., Barrett, H., Lyon, C., Brown, S., Wicklund, C., Robertson, R.D., 2021. The impact of covid-19 on road safety in canada and the united states. Accident Analysis & Prevention 160, 106324.

Wang, J., Boukerche, A., 2020. The scalability analysis of machine learning based models in road traffic flow prediction, in: ICC 2020-2020 IEEE International Conference on Communications (ICC), IEEE. pp. 1–6.

World Health Organization, 2023. Global Status Report on Road Safety 2023. World Health Organization, Geneva, Switzerland. URL: https://www.who.int/publications/i/item/9789240086517.

Yuan, R., Xiang, Q., Huang, Y., Gu, X., 2023. Investigating the difference in factors influencing the injury severity between daytime and nighttime speeding-related crashes. Canadian Journal of Civil Engineering 51, 60–72.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in ontario. Accident Analysis & Prevention 32, 117–125.

Zhang, L., Huang, Z., Kuang, A., Yu, J., Zhu, L., Yang, S., 2025. Exploring the temporal correlations of factors affecting traffic safety on mountain freeways: Through new crash frequency modelling methods. PloS one 20, e0319831.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: