Data Science for Transportation (2025) 7:14
https://doi.org/10.1007/s42421-025-00128-2

RESEARCH

Roadway Crash Typology of Census Tracts Enables Targeted Interventions via Interpretable Machine Learning

Mohammed Abdalazeem[1] · **Jimi Oke**[1]

Received: 14 March 2025 / Revised: 17 June 2025 / Accepted: 19 June 2025 / Published online: 3 July 2025

Abstract
Roadway crashes are a leading cause of death and injury worldwide, necessitating targeted interventions for improved safety. While models exist to analyze spatial variations in specific risk factor effects, a comprehensive approach to identifying and characterizing distinct types of crash-prone spatial environments across large regions remains less explored. We address this gap by introducing a novel data-driven typology of census tracts based on crash indicators across Massachusetts, Connecticut, and Vermont using 2023 crash data, providing a framework for categorizing areas based on distinct crash patterns. We apply dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) to 25 crash characteristics derived from over 222,000 crash records and aggregated across 2,660 census tracts. This reveals six key latent dimensions that capture underlying variations in crash profiles. Gaussian Mixture Modeling (GMM) reveals five distinct crash types, driven by differences in speed conditions, roadway design, and vehicle involvement, which influence crash severity and frequency. We validate the typology using a Light Gradient Boosting Machines (LightGBM) classifier, which achieved 96% accuracy across 5 classes. Finally, we demonstrate how this typology enables targeted and evidence-based safety interventions. This integrated pipeline offers a novel tool for understanding regional safety patterns and guiding the development of context-specific, proactive safety interventions.

Keywords Roadway safety · Crash typology · UMAP · Gaussian mixture modeling · LightGBM · Targeted interventions

Introduction

Roadway crashes represent a persistent global crisis, claiming over a million lives annually and standing as a leading cause of death, particularly among young people aged 5–29 (WHO 2023). In the United States, approximately 2 million people were injured in motor vehicle crashes in 2013 alone with a crash fatality rate of 10.3 fatalities per 100,000, nearly twice the rate of the next highest country—New Zealand (Centers for Disease Control and Prevention 2016). The number of crash fatalities in the U.S. has continued to increase, with an estimated 43,000 people killed in roadway crashes in 2021, an increase of 10.5% from 2020 (US Department of Transportation 2023).

Crashes are complex events, influenced by driver behavior, vehicle characteristics, roadway design, and environmental conditions (FHWA 2010). Extensive research into these individual elements has identified key risk factors, including speeding, alcohol impairment, and distracted driving (FHWA 2010; Taylor and Rehm 2012). Effective interventions are known, spanning behavioral strategies, law enforcement, and roadway engineering (Fisa et al. 2022). However, despite these advances, there is still a need to better understand the spatial variability in how crash risk factors interact to form distinct crash patterns across different geographic contexts.

We address this need by developing a typology of census tracts based on crash indicators in Massachusetts, Connecticut, and Vermont. Unlike traditional hotspot analysis (which identifies high-frequency locations) or incident-based classifications (which categorize individual crashes), our approach finds a typology of the geographic areas themselves based on their aggregate profile of crash characteristics. This method

✉ Mohammed Abdalazeem
mamohammed@umass.edu

Jimi Oke
jboke@umass.edu

1 Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Rd, Amherst, MA 01003, USA

moves beyond analyzing individual crash incidents to identifying characteristic interactions of risk factors that define distinct types of crash-prone environments. While individual techniques like spatial analysis, clustering, dimensionality reduction (UMAP), and model interpretation (SHAP) have been applied in traffic safety, our primary contribution lies in the development and application of an integrated UMAP–GMM–LGBM pipeline for spatial crash typology. This involves standardizing multi-state crash data, using UMAP to uncover latent crash profile dimensions, applying GMM to identify distinct spatial types based on these dimensions, and employing a supervised model (LightGBM) with SHAP analysis for interpreting the characteristics that define and differentiate these resulting types. This comprehensive pipeline provides a transferable framework for obtaining interpretable data-driven crash types in a large number of potentially heterogeneous units, allowing a more nuanced understanding of spatial safety patterns to better inform targeted interventions.

Literature Review

Crash analysis has traditionally relied on statistical and behavioral approaches. Early safety studies applied regression models, including logistic, Poisson, and negative binomial regressions, as well as survival analysis, to understand crash occurrences and injury severity (Aziz et al. 2013; Schneider and Savolainen 2011). These models provided interpretable relationships between risk factors and crash outcomes, while behavioral research emphasized human factors. Studies have consistently shown that driver behaviors like speeding, distraction, and impairment account for a significant portion of crashes (Mekonnen et al. 2019). Traditional policy-driven interventions focus on enforcement (e.g., DUI laws, speed cameras) and infrastructure modifications (e.g., traffic calming measures), with programs such as Vision Zero utilizing crash data to guide multi-sector safety efforts (Kumar and Kalapurayil 2022; US Department of Transportation 2023). Yet, these strategies often rely on historical crash data and broad policy measures, which may overlook localized risk factors and evolving crash dynamics (Aguero-Valverde and Jovanis 2006).

With increasing data availability and computational capabilities, machine learning (ML) methods have been increasingly applied to crash analysis (Lin et al. 2018; Rella Riccardi et al. 2022; Ali et al. 2024; Bhowmik et al. 2019, 2021; Kim et al. 2006). Recent advances emphasize the integration of diverse data sources, including roadway infrastructure, environmental conditions, and human factors, to enhance predictive power and interpretability (Paikari et al. 2014; Eskandari Torbaghan et al. 2022; Imprialou and Quddus 2019). Sophisticated clustering techniques have been used to uncover hidden crash patterns, enabling more targeted safety interventions. Building on these developments, ML techniques such as decision trees, random forests (RF), support vector machines, and deep neural networks have demonstrated superior performance over traditional statistical models, particularly in capturing non-linear relationships (Behboudi et al. 2024). Hybrid approaches, which integrate ML with econometric models, have further improved predictive accuracy while preserving interpretability (Wang et al. 2025; Lin et al. 2018). For example, Jin et al. (2023) developed a two-stage framework that combines a logit model with random parameters with a deep neural network to capture unobserved heterogeneity and complex feature interactions in real-time crash risk prediction for highway tunnels.

Spatial crash analysis has also gained prominence, using clustering and GIS-based techniques to identify crash hotspots (Kamh et al. 2024; Mohammed et al. 2023). Methods such as Kernel Density Estimation (KDE) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering are widely used to detect high-risk locations (Kamh et al. 2024; Jia et al. 2018; Bíl et al. 2019). Network-constrained clustering methods, such as the Getis-Ord Gi* statistic, provide more refined spatial insights by adjusting for road network structures (Nie et al. 2015; Lu et al. 2013). Recent advances also integrated ML with spatial clustering to analyze spatiotemporal crash distributions, improving hotspot prediction accuracy (Kamh et al. 2024; Jia et al. 2018; Zhou et al. 2023; Nitsche et al. 2017; Sun et al. 2019; Uchida et al. 2010; Abdel-Aty and Haleem 2011; Bajada and Attard 2021; Guo et al. 2023; Bíl et al. 2019; Eckley and Curtin 2013; Prasannakumar et al. 2011; Nie et al. 2015; Abdel-Aty et al. 2005). Beyond hotspot identification, recent studies have integrated spatio-temporal context into crash prediction models. A study applied a sparse nonnegative matrix factorization with spatial constraints to urban crash data, revealing distinct crash patterns over space and time and identifying high-risk areas and periods for targeted interventions (Jin et al. 2024). Similarly, graph neural networks have been used to exploit the structure of the road network in the prediction of crashes (Gao et al. 2024). Another emerging direction is to model the spatial heterogeneity of crash risk factors. Spatially adaptive algorithms can reveal how the influence of various factors differs by location, which traditional global models might overlook (Zhang et al. 2024; Gu et al. 2023). However, such highly localized approaches can also make it challenging for practitioners to derive generalizable interventions that can be applied in broader geographic contexts (Li et al. 2021; Zhang et al. 2023). In addition, deep-reinforcement learning frameworks have been introduced to proactively manage crash risk through dynamic control strategies such as variable speed limits, particularly in high-risk environments such as highway tunnels (Jin et al. 2025). Furthermore, examination of pre-crash scenarios and the use of in-depth crash data have enabled the identification of critical factors and patterns that precede crashes, facilitating the development of proactive safety measures (Bajada and Attard 2021; Guo et al. 2023).

However, significant research gaps remain. Many studies focus on specific locations or types of crashes, limiting generalizability (Aziz et al. 2013). For example, injury severity models developed for a single urban area might not account for rural versus urban differences, and many analyses use fixed-effect models that cannot capture variation between sites (Aziz et al. 2013). Furthermore, while some research has examined multi-state or nationwide crash data, these have been relatively scarce, with challenges in data integration and consistency across jurisdictions persisting (Dadvar et al. 2020). The Highway Safety Information System (HSIS) provides multi-state datasets; however, differences in reporting, exposure, and road inventory often force most studies using these data to remain restricted to specific road types or regions (Dadvar et al. 2020).

Our study addresses these limitations by developing a spatial crash typology that integrates machine learning with multi-state crash data from Massachusetts, Connecticut, and Vermont. Using dimensionality reduction (UMAP) and Gaussian Mixture Modeling, we identify latent crash patterns at the census tract level. Unlike previous research, our approach aims to be transferable between different jurisdictions while preserving localized insights, offering a scalable framework for targeted safety interventions.

Methods

The methodological framework is shown in Fig. 1. First, we collected and integrated raw crash data from multiple sources. We then standardized and aggregated the data at the census tract level. Next, we applied dimensionality reduction techniques to identify latent factors in crash patterns. Using cluster analysis, we uncovered distinct crash types. Finally, we trained a classification model to validate the typology and highlight the relevant indicators.

Data

Sources and Initial Statistics

We sourced motor vehicle crash data for 2023 from the official data repositories of Massachusetts (MassDOT 2023), Connecticut (CTDOT 2023), and Vermont (VTrans 2023). While containing core information about crashes, each state's dataset exhibited variations in data structure, file organization, naming conventions, and the specific variables included. This required careful integration and standardization, as described below.

The primary data types we considered were: (1) detailed crash records from the respective state departments of transportation, (2) census tract boundary geometries from the U.S. Census Bureau, and (3) road network data, also from the U.S. Census Bureau, used for calculating road density. Table 1 provides a summary of key characteristics of the raw crash data for each state, providing an overview of the data landscape before integration.

Figure 2 displays the spatial distribution of crashes across Massachusetts, Connecticut, and Vermont. Figure 2a shows individual crash locations, while Fig. 2b presents the total number of crashes aggregated within each census tract in the region.

Data Aggregation and Feature Distributions

We integrated and prepared the crash data from the three states to create a unified dataset for analysis. Given the differences in state reporting standards and definitions, we obtained a standardized dataset by first identifying a set of shared or conceptually similar relevant variables across the

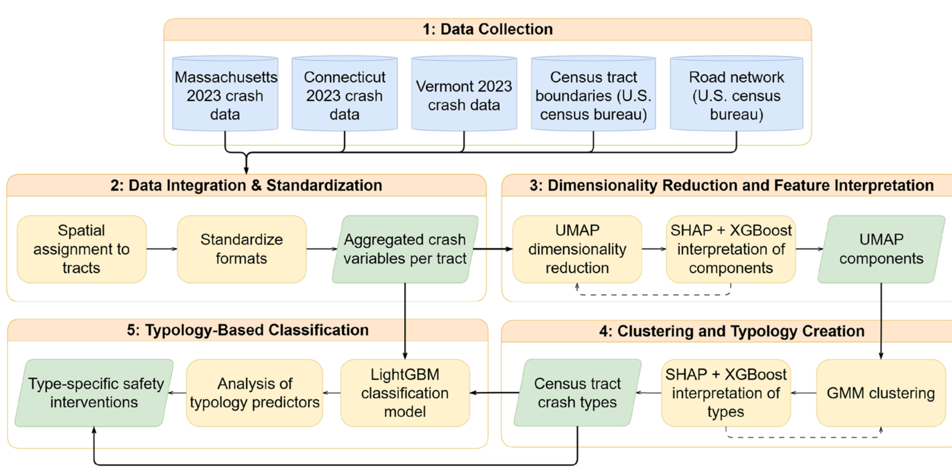


Fig. 1 Methodological flowchart

Table 1 Summary of raw crash data characteristics (2023)

State	Crash records	Initial columns	Cities/towns	Census tracts	Avg crashes per tract
Massachusetts	115,573	123	343	1,589	72
Connecticut	99,953	103	169	878	113
Vermont	7,010	37	233	193	35
Total	222,536	–	745	2,660	83

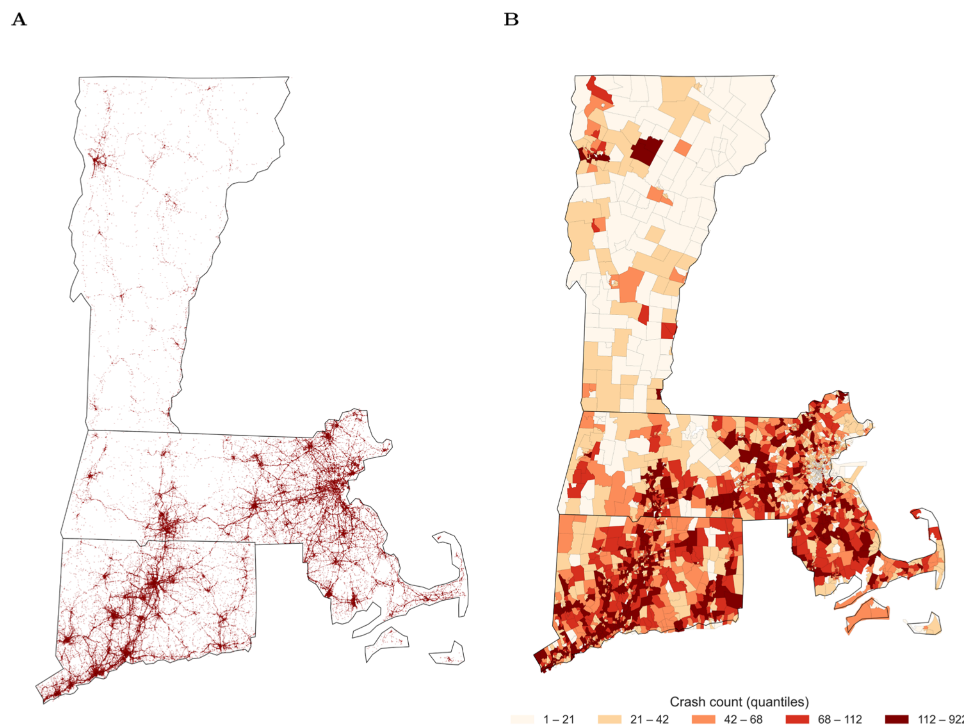


Fig. 2 Geographic distribution of crashes across Massachusetts, Connecticut, and Vermont in 2023: **a** Individual crash locations. and **b** Crashes aggregated by census tract

combined region, including information on crash, vehicle, and network information. We then created a uniform coding system for the core variables. For instance, we mapped different state-level crash severity classifications onto a common standardized scale and grouped disparate vehicle type codes into consistent categories, ensuring comparability across the datasets. This process yielded a dataset suitable for aggregation.

Spatial analysis of crash patterns requires aggregating data into a consistent geographic unit to allow for meaningful comparisons and pattern detection. We selected census tracts as the spatial unit for aggregation for several reasons. First, they provide standardized and relatively stable geographic boundaries across all three states, facilitating consistent regional analysis. Second, census tracts represent a common unit used in transportation planning and allow for potential future integration with demographic or socioeconomic data. Third, they offer a meso-level perspective, balancing the high granularity of individual road segments with broader regional divisions like counties, resulting in a manageable number of units ($N = 2,682$) for typology development.

We aggregated the individual crash records at the census tract level, which transformed the data into a summary of crash characteristics. For the categorical variables, including the *Intersection type* and *Crash severity,* we calculated the proportion of crashes within each tract belonging to each category. For the numerical variables, such as the *Speed limit*, we calculated the mean value within each census tract. The detailed procedures for standardizing the variable definitions and coding systems across the three states, and the subsequent data aggregation are provided in Appendix 1.

We selected the features based on several considerations. Firstly, we drew upon established traffic safety literature and domain expertise to identify characteristics known to influence crash patterns and outcomes (NHTSA 2021; Farmer 2019; Sullivan 2018; Flannagan et al. 2021). Secondly, we were constrained by data availability across all states. Thirdly, we aimed to provide a comprehensive representation of crash incidents within each census tract, with features such as crash severity proportions, typical vehicle involvement, intersection characteristics, environmental factors, and roadway context (e.g., mean speed limit, road density). This resulted in a final dataset with 2,660 rows (one for each census tract) and 25 columns representing the aggregated crash characteristics. Figure 3 shows the overall distributions of the aggregated features across all three states combined. While underlying raw data distributions may vary somewhat between states, as shown in Table 1, this study focuses on identifying common regional patterns emerging after data standardization and aggregation. The definitions of the features shown in the figure are provided in Table 2.

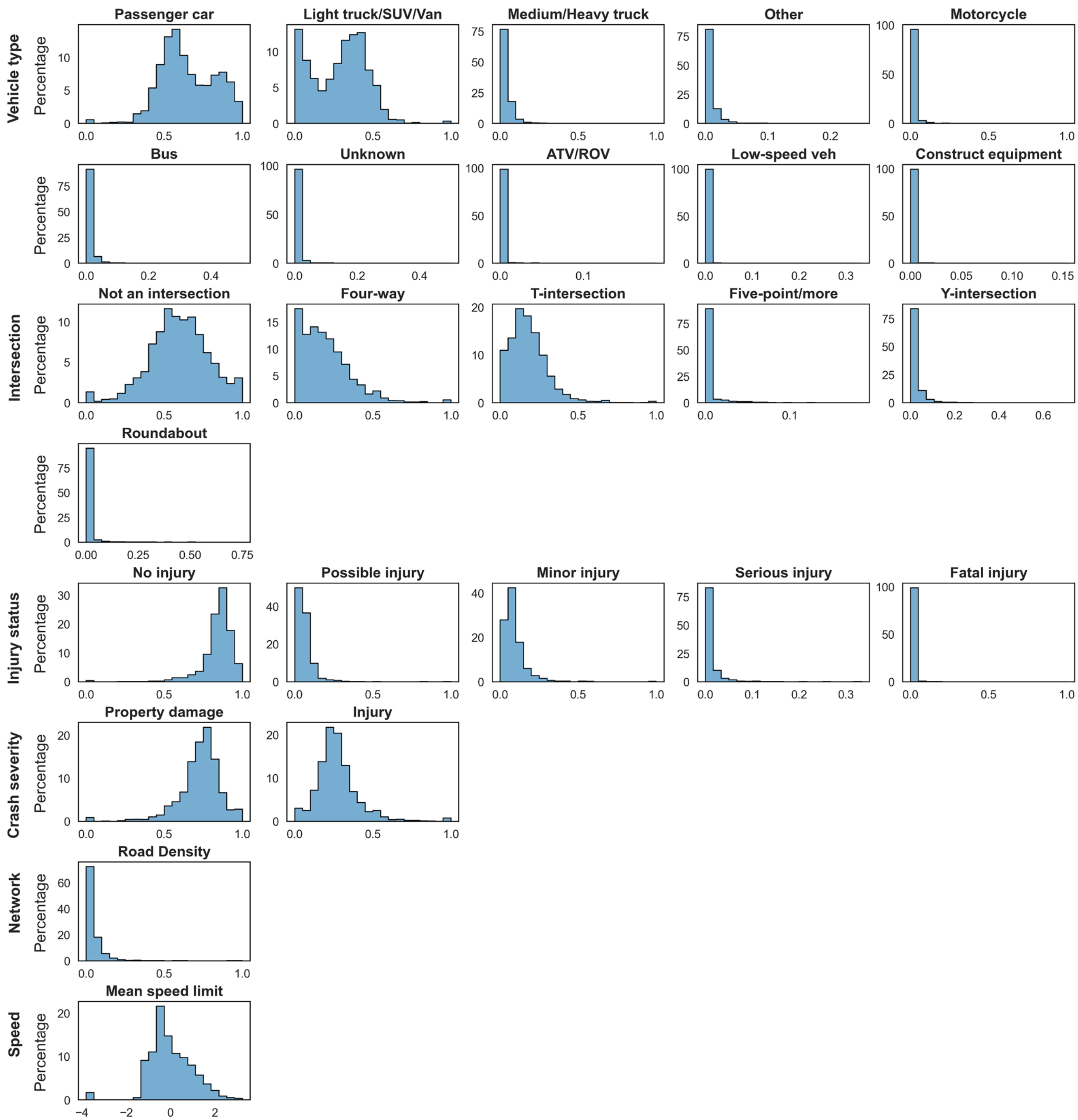


Fig. 3 Histograms of the 25 selected and aggregated crash characteristic variables, grouped by category

Table 2 Definition of the 25 aggregated crash characteristic features used as input for UMAP analysis within each census tract

Feature source	Feature value/category	Calculation method and/or unit
Vehicle Type Involved[a]	Passenger car	Proportion[b]
	Light truck/SUV/Van	
	Medium/Heavy truck	
	Other	
	Motorcycle	
	Bus	
	Unknown	
	ATV/ROV	
	Low-speed veh	
	Construct equipment	
Intersection Type[c]	Not an intersection	Proportion[b]
	Four-way	
	T-intersection	
	Five-point/more	
	Y-intersection	
	Roundabout	
Crash Severity Outcome	Property damage only	Proportion[b]
	Any Injury	
Highest Injury Status[d]	No injury	Proportion[b]
	Possible injury	
	Minor injury	
	Serious injury	
	Fatal injury	
Network	Road Density	Length per area (km/km^2)[e]
Crash Speed	Mean Speed Limit	Mean value (mph)[f]

[a]Based on types of vehicles involved in crashes within the tract

[b] Proportion of all crashes within the census tract matching the specified category/condition

[c] Based on the geographic coordinates of the crash location relative to intersections

[d] Based on the highest injury severity recorded for any person involved in a crash

[e] Calculated as total length of primary/secondary roads divided by tract area

[f] Mean of the posted speed limits associated with individual crash locations within the tract

Dimensionality Reduction

To identify the underlying latent factors driving the crash patterns, we applied dimensionality reduction (DR) to the 25 aggregated crash characteristic variables to reduce complexity while preserving the essential data structure and enabling effective clustering. We evaluated several approaches, namely Principal Component Analysis (PCA) (Hotelling 1933), Kernel PCA (KPCA) (Schölkopf et al. 1998), t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008), and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2020), to determine the optimal method for this task.

After tuning each DR method via grid search, we applied a temporary K-means clustering ($k = 8$) to the reduced dimensions and assessed cluster quality using three validity metrics: Silhouette Score, Davies–Bouldin Index (DB Index), and Calinski–Harabasz Index (CH Index). This step was performed for all DR methods to ensure a fair comparison of their ability to preserve meaningful structure in lower dimensions. We then calculated the average rank across the metrics to determine the best-performing model. Table 3

Table 3 Comparison of dimensionality reduction techniques

Method	Silhouette score	Davies–Bouldin index	Calinski–Harabasz index	Average rank
KPCA	0.272	1.062	2640.670	3.00
PCA	0.254	1.133	2420.980	4.00
t-SNE	*0.429*	0.725	3714.861	1.67
UMAP	0.414	*0.714*	*6130.536*	*1.33*

Italics indicate the best value for each metric presents a comparative analysis of the performance across the methods. UMAP achieved the best CH Index, best DB Index, and the best average rank of 1.33.

Based on this, we selected UMAP. At its core, UMAP constructs a high-dimensional graph representation of the data and then optimizes a low-dimensional layout to be as structurally similar as possible. This similarity is quantified using cross-entropy, aiming to minimize the divergence between the probability distributions representing connections in the high-dimensional (v_{ij}) and low-dimensional (w_{ij}) spaces, as shown in Equation 1:

$$\mathbb{H}(v, w) = \sum_{i \neq j} \left[v_{ij} \log \left(\frac{v_{ij}}{w_{ij}} \right) + (1 - v_{ij}) \log \left(\frac{1 - v_{ij}}{1 - w_{ij}} \right) \right] \quad (1)$$

where v_{ij} represents the probability of a connection between data points i and j in the high-dimensional space, and w_{ij} represents the corresponding probability in the low-dimensional space. UMAP aims to preserve the relationships between data points as faithfully as possible during the dimensionality reduction process by minimizing this cross-entropy (McInnes et al. 2020).

We optimized UMAP's hyperparameters through a grid search, assessing potential cluster quality on the reduced dimensions using the silhouette score. The parameter ranges explored and the results of this tuning process are summarized in Table 4 and visualized in Fig. 4. Applying UMAP with the selected optimized parameters transformed the original 25 aggregated features for each census tract into a new set of 6 numerical components. This resulting six-dimensional representation captures the core structure of the initial, high-dimensional data in a lower-dimensional space and served as the input for the subsequent clustering analysis ("Cluster Analysis" section).

Table 4 UMAP hyperparameter tuning: parameters and ranges

Parameter	Range explored
Distance metric	{Euclidean, Manhattan}
Minimum distance	[0.0, 0.5, 0.99]
No. neighbors	[3, 18]
No. components	[3, 10]

Latent Component Interpretation

The UMAP transformation reduces the dataset (the original 25 aggregated features) to a lower-dimensional space while preserving the important structure, resulting in a set of latent components that serve as inputs for clustering. To interpret the role of these components, we first fitted 6 eXtreme Gradient Boosting (XGBoost) models, each predicting one of the UMAP dimensions. XGBoost is a tree-based machine learning algorithm based on Gradient Boosting Machines (GBM) (Friedman 2000; Mason et al. 1999; Breiman 1997), which builds decision trees sequentially to correct previous errors and improve accuracy. XGBoost enhances the process via regularization and parallel computation, making it faster and more robust (Chen and Guestrin 2016). We then applied SHapley Additive exPlanations (SHAP) to quantify how each original crash feature contributes to the predicted UMAP dimensions (Lundberg and Lee 2017; Lundberg et al. 2020). This allows us to characterize the meaning of the latent space before using it for crash type analysis.

Given a UMAP component z_l, the trained XGBoost model estimates it as a function of the original crash features:

$$z_l = \hat{f}_l(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D) \quad (2)$$

where $\hat{f}_l(\cdot)$ is the predictive function learned by XGBoost, and $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D$ are the original crash characteristics. The SHAP value ϕ_d^l for feature $\boldsymbol{x}_d$ in predicting z_l is computed as:

$$\phi_d^l = \sum_{\mathcal{S} \subseteq \{\boldsymbol{x}_1, ..., \boldsymbol{x}_D\} \setminus \{\boldsymbol{x}_d\}} \frac{|\mathcal{S}|!(D - |\mathcal{S}| - 1)!}{D!} \left[\hat{f}_l(\mathcal{S} \cup \{\boldsymbol{x}_d\}) - \hat{f}_l(\mathcal{S}) \right] \quad (3)$$

where ϕ_d^l represents the marginal contribution of feature $\boldsymbol{x}_d$ across all possible subsets $\mathcal{S}$ of input features.

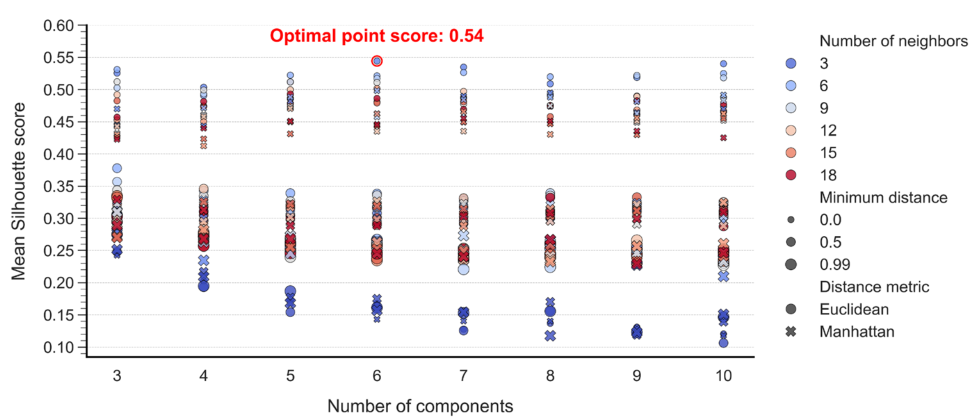


Fig. 4 UMAP hyperparameter tuning results: Silhouette score versus number of components, color-coded by number of neighbors

The SHAP values provide an additive decomposition of each model prediction, explaining how individual crash characteristics influence the UMAP representation. A positive ϕ_d^l indicates that feature $\boldsymbol{x}_d$ increases the predicted value of UMAP component z_l, whereas a negative value suggests a decreasing effect. The magnitude of ϕ_d^l reflects the strength of this influence, enabling a detailed interpretation of the latent space.

Cluster Analysis

We clustered census tracts in the reduced UMAP space using: K-means (MacQueen 1967), Gaussian Mixture Models (GMM) (Dempster et al. 1977), hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al. 2013), and Self-Organizing Maps (SOM) (Kohonen 1982). Table 5 presents a comparative analysis of the clustering performance across the methods. As with dimensionality reduction, we tuned each method using a grid search and then used three cluster quality metrics: the Silhouette Score, DB Index, and CH Index, to assess the quality of the clusters. GMM achieved the best average rank (1.33) and was selected as the optimal clustering method.

GMM assumes that the data points (in this case, UMAP component coordinates) are generated from a mixture of K Gaussian distributions, each characterized by its respective mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The model is specified as:

$$\hat{p}(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

where $\hat{p}(z)$ is the probability density of the data point z, K is the number of components, and π_k is the mixing coefficient for the k-th Gaussian component, representing the prior probability of a data point belonging to that component ($\sum_{k=1}^{K} \pi_k = 1$). Additionally, $\mathcal{N}(z \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian probability density function for the k-th component.

The parameters of the GMM (π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ for each component) were estimated using the Expectation–Maximization (EM) algorithm. The E step computes the probability that each data point z_n belongs to component or cluster k:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(z_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(z_n \mid \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \tag{5}$$

and in the M step, the parameters are updated as follows:

$$\pi_k = \frac{N_k}{N}, \quad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} z_n}{N_k}, \quad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (z_n - \boldsymbol{\mu}_k)(z_n - \boldsymbol{\mu}_k)^\top}{N_k} \tag{6}$$

where $N_k = \sum_{n=1}^{N} \gamma_{nk}$ is the effective number of points assigned to cluster k. These steps are repeated until convergence.

Table 6 summarizes the hyperparameters of the grid search used to optimize the GMM and their respective ranges. We examined the Bayesian Information Criterion (BIC), depicted in Fig. 5a, to assess model complexity penalties. Additionally, we used the Silhouette Score to evaluate different hyperparameter combinations, as shown in Fig. 5b. We selected $K = 5$ as the final number of clusters. This choice was based on the objectives of model fitness, parsimony, and interpretability. Specifically, $K = 5$ fell within the optimal range suggested by the BIC elbow (Fig. 5a), maintained a reasonably high average silhouette score (Fig. 5b), and yielded clusters with distinct profiles. We considered this level of granularity most ideal for distinguishing different dominant crash patterns across census tracts, potentially aiding the development of targeted safety strategies relevant to real-world conditions.

Applying the GMM algorithm with these optimized parameters to the UMAP-reduced data yielded 5 cluster assignments for each census tract. Figure 6 shows a pair plot of the UMAP components, with points colored according to their assigned GMM cluster (crash type).

Table 6 GMM hyperparameter tuning: parameters and ranges

Parameter	Range
Covariance matrix specification	{Full, Tied, Diagonal, Spherical}
No. components	[2, 14]
Parameter initialization method	{K-means, Random}
Regularization covariance	$\{10^{-6}, 10^{-5}, 10^{-4}\}$

Table 5 Comparison of clustering methods

Method	Silhouette	Calinski–Harabasz	Davies–Bouldin	Average rank
GMM	*0.467*	6013.262	*0.620*	*1.33*
HDBSCAN	0.311	633.718	1.832	4.00
K-means	0.449	*12539.947*	0.742	2.00
SOM	0.331	2561.390	0.693	2.67

Italics indicate the best value for each metric

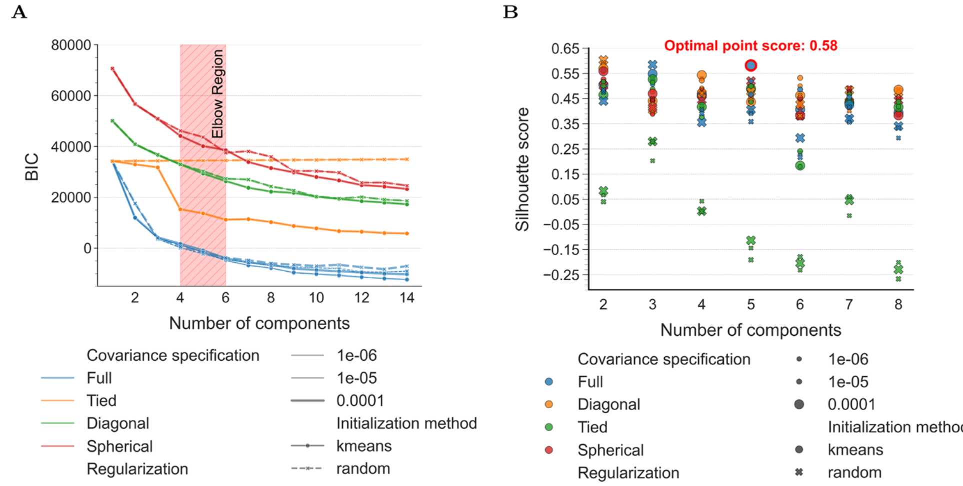


Fig. 5 Comparison of GMM hyperparameter tuning metrics: **a** Bayesian Information Criterion (BIC) values. The shaded red region indicates the elbow region, which suggests that a range of components between 4 and 6 balances model fit and complexity, and **b** The Silhouette score, color-coded by the covariance type, evaluated for number of components between 2 and 8

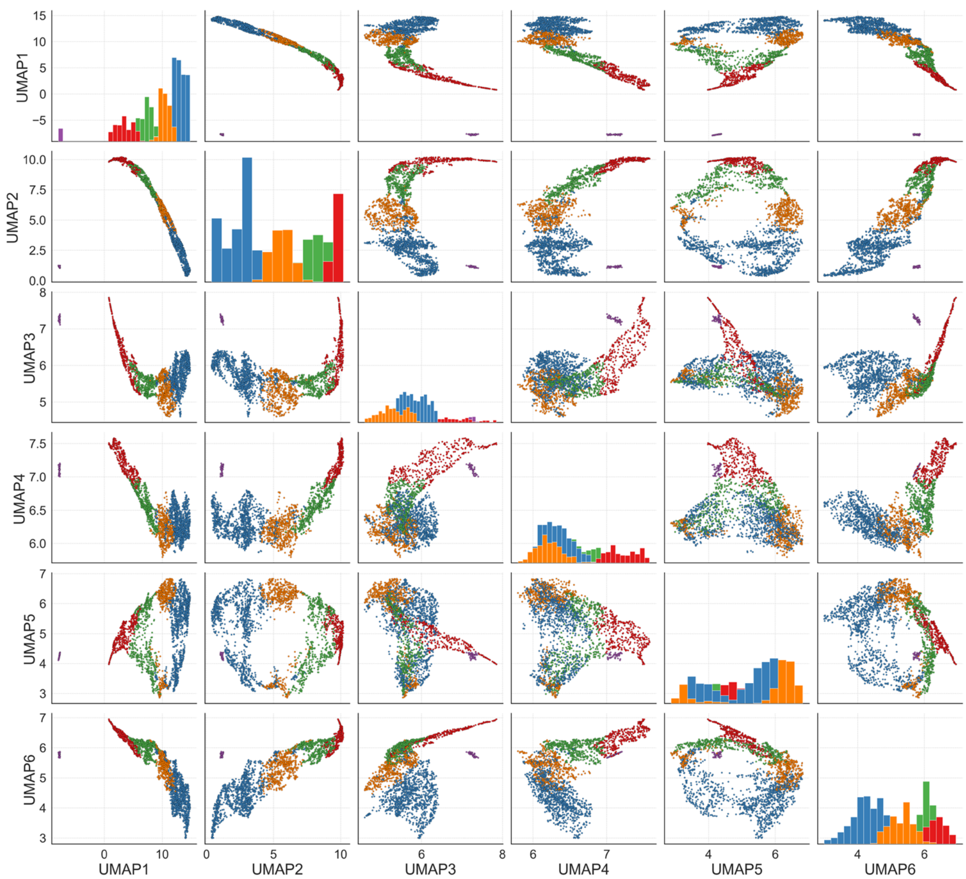


Fig. 6 Pairplot of UMAP components, colored by the crash type

Crash Typology Interpretation

After clustering, each census tract was assigned a crash type based on its position in the UMAP-reduced space. To understand how the latent UMAP components influence these assignments, we first fitted a single XGBoost model to predict crash type probabilities using the 6 UMAP dimensions (components) as input features. We then calculated SHAP values for each crash type. This analysis quantifies the contribution of each UMAP component to the probability that a census tract belongs to a specific type of crash.

Given a predicted crash type label $\hat{k}$, the XGBoost model estimates its probability as:

$$P(\hat{k} = k \mid z_1, z_2, ..., z_L) = \frac{e^{\hat{g}_k(z)}}{\sum_{k'=1}^{K} e^{\hat{g}_{k'}(z)}} \tag{7}$$

where $z_1, z_2, ..., z_L$ represent the UMAP components, K is the number of clusters, and $\hat{g}_k(\cdot)$ is the predictive function learned by XGBoost that maps the reduced feature space to class scores before applying a softmax transformation.

The SHAP value ψ_l^k for UMAP component z_l in predicting crash type k is computed as:

$$\psi_l^k = \sum_{\mathcal{S} \subseteq \{z_1,...,z_L\} \setminus \{z_l\}} \frac{|\mathcal{S}|!(L - |\mathcal{S}| - 1)!}{L!} \left[\hat{g}_k(\mathcal{S} \cup \{z_l\}) - \hat{g}_k(\mathcal{S})\right] \tag{8}$$

where ψ_l^k represents the marginal effect of component z_l across all possible subsets $\mathcal{S}$ of the input features.

The SHAP values provide an interpretable breakdown of how the reduced-dimensional representation influences crash type assignment. A positive ψ_l^k indicates that z_l increases the likelihood of assignment to crash type k, while a negative value suggests a suppressing effect.

Assessing Typology Separability and Robustness

We developed a supervised classification modeling approach to assess whether the five crash types derived from the unsupervised UMAP–GMM pipeline ("Cluster Analysis" section) represented statistically separable and recoverable patterns based on the original crash characteristics. Thus, we trained models to predict the GMM-assigned cluster label (crash type 1–5) for each census tract using the original 25 aggregated crash characteristic features (Table 2) as inputs, instead of the 6 UMAP dimensions. The goal was to determine whether these original features were sufficient to reliably distinguish between the clusters identified in the latent space, thus confirming the robustness and distinctiveness of the typology. This approach also provides a model capable of assigning new census tracts to a crash type based on their features, without re-running the UMAP–GMM pipeline.

We evaluated several classification algorithms, including light gradient boosting machines (LightGBM) (Ke et al. 2017), random forests (Breiman 2001), logistic regression (Cox 1958), artificial neural networks (ANN) (Rosenblatt 1958), and XGBoost (Chen and Guestrin 2016). We conducted a grid search with 3-fold cross-validation for each model to optimize hyperparameters, aiming to maximize prediction accuracy.

Table 7 provides a detailed comparison of the evaluation metrics for the top models on the test set. LightGBM (Ke et al. 2017), a variant of GBM designed for efficiency using histogram-based learning to speed up training, achieved the strongest performance, with the best scores across all metrics considered, including an overall accuracy of 96.1%.

Table 7 Evaluation metrics for the top classification models

Metric	ANN	LightGBM	Logistic	Random	XGBoost
Accuracy	0.908	*0.961*	0.879	0.947	0.949
F1 Score (Macro)	0.914	*0.966*	0.892	0.930	0.929
Log Loss	0.211	*0.099*	0.268	0.324	0.120
Precision (Macro)	0.917	*0.967*	0.894	0.957	0.922
Recall (Macro)	0.912	*0.966*	0.890	0.909	0.938
ROC AUC (Macro)	0.992	*0.999*	0.986	0.994	0.998

Italics indicate the best-performing model (LightGBM) across all metrics

Formally, LightGBM minimizes the following objective function:

$$\mathcal{L} = \sum_{n=1}^{N} \ell(y_n, \hat{y}_n) + \sum_{t=1}^{T} \Omega(\hat{f}_t), \tag{9}$$

where $\ell(y_n, \hat{y}_n)$ is the loss function comparing the true label y_n with the predicted value $\hat{y}_n$, N is the total number of training samples, $\hat{f}_t$ represents the prediction function of the t-th decision tree, T is the total number of trees in the boosting process, and $\Omega(\hat{f}_t)$ is a regularization term that controls the complexity of each tree.

The optimization is performed using a second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{n=1}^{N} \left[g_n \hat{f}_t(\boldsymbol{x}_n) + \frac{1}{2} h_n \hat{f}_t^2(\boldsymbol{x}_n)\right] + \Omega(\hat{f}_t), \tag{10}$$

where $\mathcal{L}^{(t)}$ is the approximated objective function at iteration t, $\hat{f}_t(\boldsymbol{x}_n)$ is the output of the t-th tree for input $\boldsymbol{x}_n$, and g_n and h_n are the first and second derivatives (gradient and Hessian) of the loss function with respect to the model output at sample $\boldsymbol{x}_n$.

Results and Discussion

Latent Crash Components are Characterized by Speed, Network, and Vehicle Type

We analyzed SHAP values from six trained XGBoost models, each corresponding to one UMAP component. Figure 7 presents SHAP value plots for the top 5 most important features of each UMAP component, sorted by mean absolute SHAP value.

Analysis of the SHAP plots (Fig. 7) revealed that the six UMAP components represent distinct combinations of speed, location/intersection type, and vehicle type characteristics. *Hi-Speed Dense-Road Truck* (Fig. 7a)

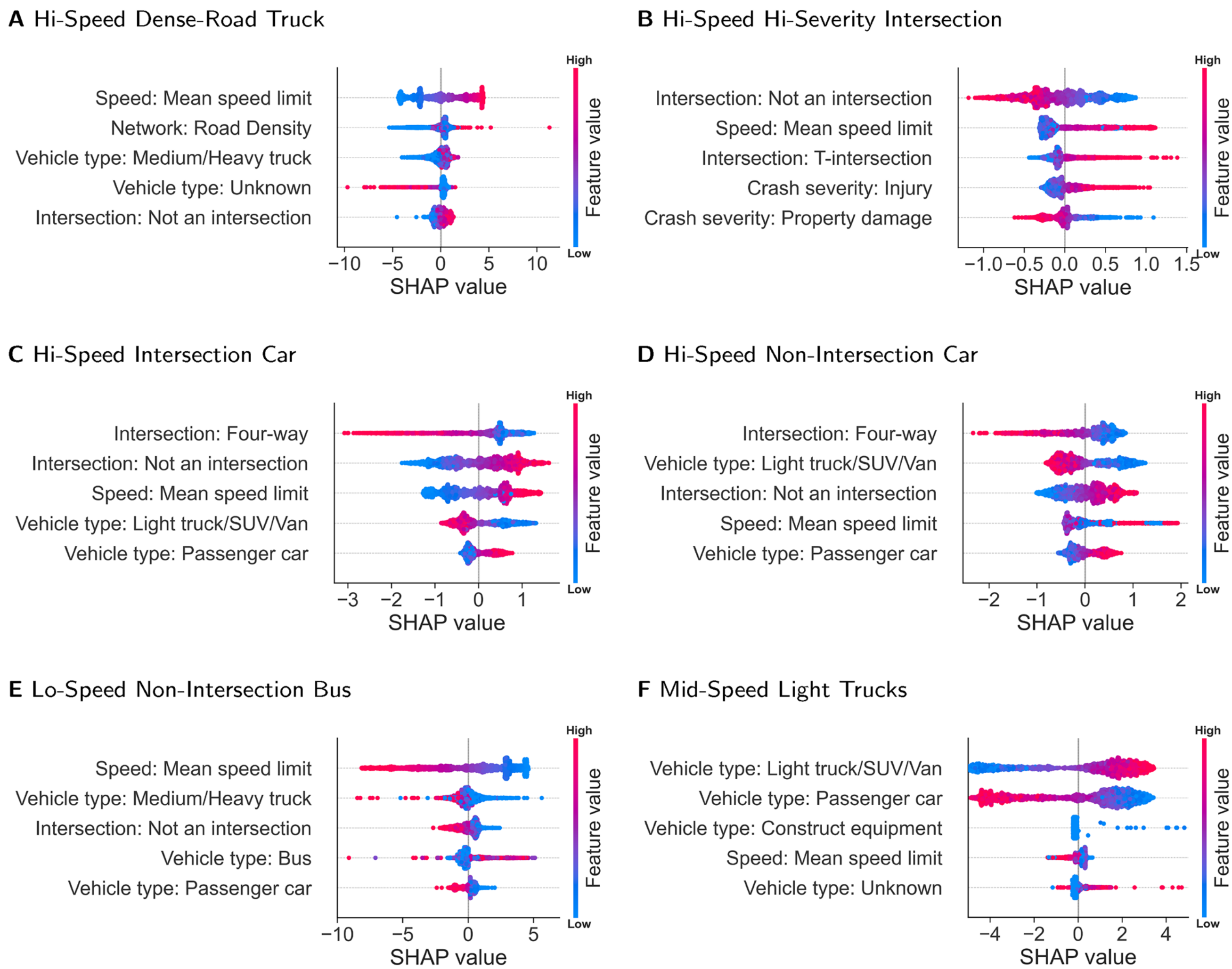


Fig. 7 SHAP value plots showing feature influence on each UMAP component. Each point in the plot represents a single census tract and each subplot shows the SHAP values for the top 5 features influencing the corresponding UMAP component

captures high-speed roadways, higher road density, and crashes involving trucks. *Hi-Speed Hi-Severity Intersection* (Fig. 7b) represents higher-severity crashes (injuries) occurring at intersections, particularly T-intersections, and at higher speeds. *Hi-Speed Intersection Car* (Fig. 7c) highlights higher-speed crashes at intersections (but not four-way intersections), and more passenger cars involved. *Hi-Speed Non-Intersection Car* (Fig. 7d) relates to crashes away from four-way intersections, often involving passenger cars at higher speeds. *Lo-Speed Non-Intersection Bus* (Fig. 7e) is associated with lower average speeds, non-intersection crashes, and bus involvement. Finally, *Mid-Speed Light Trucks* (Fig. 7f) primarily distinguishes between crashes involving light trucks/SUVs/vans versus passenger cars. Based on these interpretations, we assigned concise, descriptive names to each UMAP component, as summarized in Table 8.

Crash Types are Defined by Spatial Patterns, Roadway Design, and Vehicle Interaction

The typology analysis revealed five distinct crash types. The spatial distribution of these types across the study area is visualized in Fig. 8, which reveals geographical patterns ranging from dense urban cores to more dispersed suburban and rural settings. Radar plots of component averages in each type and the type-specific SHAP plots from the fitted XGBoost model are shown in Fig. 9.

In naming the crash types, we adopted a binomial approach. The first term refers to the speed characteristics (*Fast*, *Mixed*, or *Slow*) based on the predominant speed-related UMAP component score averages. The second term indicates the spatial context (*Rural*, *Suburban*, or *Urban*) based on observed geographic distribution patterns

Table 8 Description of UMAP components

Name	Key characteristics
Hi-Speed Dense-Road Truck	High-speed roadways with higher road density and a significant proportion of truck involvement
Hi-Speed Hi-Severity Intersection	Higher-severity (injury) crashes, often at intersections (especially T-intersections) and associated with higher speeds
Hi-Speed Intersection Car	Higher-speed crashes at intersections other than four-way intersections, with a higher proportion of passenger cars involved
Hi-Speed Non-Intersection Car	Fewer crashes at four-way intersections, a higher proportion of non-intersection and passenger car involvement
Lo-Speed Non-Intersection Bus	Lower-speed, non-intersection crashes, with a greater proportion of bus involvement
Mid-Speed Light Trucks	Primarily distinguishes between crashes involving light trucks/SUVs/vans (higher values) and passenger cars (lower values)

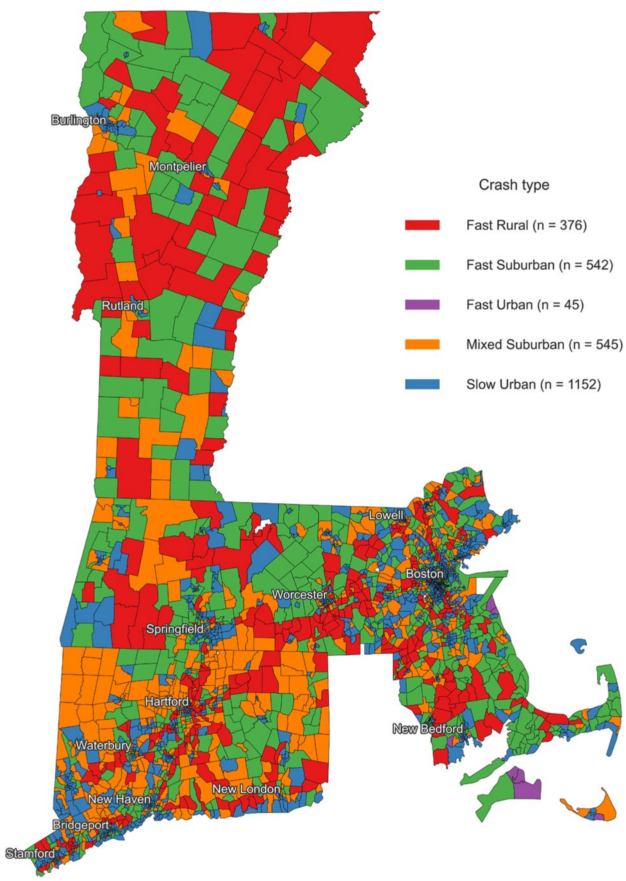


Fig. 8 Spatial distribution of the five crash types in Massachusetts, Connecticut, and Vermont across census tracts. Each color represents a different crash type, with the number of census tracts shown in parentheses in the legend

and infrastructure characteristics, such as road density and intersection type.

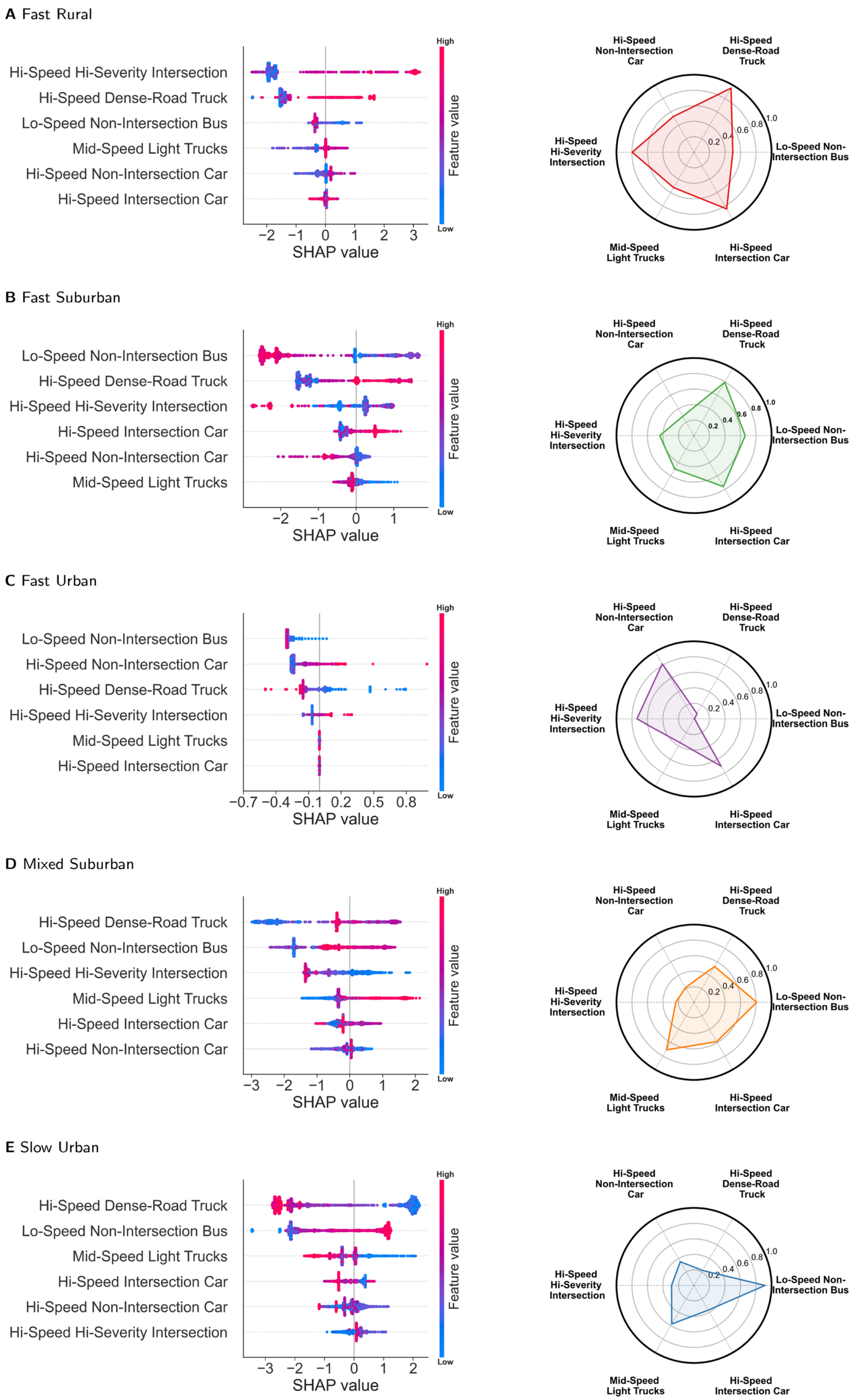
The figures within Fig. 9 jointly explain how the six latent components contribute and help interpret the five crash types. The radar plots in Fig. 9 show the average magnitude of each component within a type, while the SHAP-value figures in Fig. 9 show the directional influence of high-versus-low values on the probability that a tract belongs to that type. A crash type is therefore defined by the unique combination of components that push membership probability up (positive SHAP values) or down (negative SHAP values). Furthermore, Fig. 11 provides a visual summary mapping the key features contributing to each crash type classification. Table 9 summarizes the key spatial and characteristic features of each crash type. We describe the types in detail below.

Fast Rural

Distributed across all three states in areas surrounding urban centers and extending into less densely populated regions, this type is most strongly defined by *Hi-Speed Hi-Severity Intersection*, with a secondary, but still substantial, association with *Hi-Speed Dense-Road Truck*. This type is also less common in areas with high *Lo-Speed Non-Intersection Bus* values. This profile points to injury-producing crashes at intersections on higher-speed roadways, often involving trucks, which is common in suburban–rural transition zones.

Mixed Suburban

This type is widespread across the three states, but less concentrated in dense urban cores. While there is some influence from *Lo-Speed Non-Intersection Bus*, this type is prominently associated with *Hi-Speed Dense-Road Truck* and *Mid-Speed Light Trucks*, indicating a strong association with crashes on higher-speed roadways, with higher road density, and substantial truck involvement, typical of major suburban arterial roads.


A Fast Rural
Hi-Speed Hi-Severity Intersection
Hi-Speed Dense-Road Truck
Lo-Speed Non-Intersection Bus
Mid-Speed Light Trucks
Hi-Speed Non-Intersection Car
Hi-Speed Intersection Car
−2 −1 0 1 2 3
SHAP value
High
Feature value
Low
B Fast Suburban
Lo-Speed Non-Intersection Bus
Hi-Speed Dense-Road Truck
Hi-Speed Hi-Severity Intersection
Hi-Speed Intersection Car
Hi-Speed Non-Intersection Car
Mid-Speed Light Trucks
−2 −1 0 1
SHAP value
C Fast Urban
Lo-Speed Non-Intersection Bus
Hi-Speed Non-Intersection Car
Hi-Speed Dense-Road Truck
Hi-Speed Hi-Severity Intersection
Mid-Speed Light Trucks
Hi-Speed Intersection Car
−0.7 −0.4 −0.1 0.2 0.5 0.8
SHAP value
D Mixed Suburban
Hi-Speed Dense-Road Truck
Lo-Speed Non-Intersection Bus
Hi-Speed Hi-Severity Intersection
Mid-Speed Light Trucks
Hi-Speed Intersection Car
Hi-Speed Non-Intersection Car
−3 −2 −1 0 1 2
SHAP value
E Slow Urban
Hi-Speed Dense-Road Truck
Lo-Speed Non-Intersection Bus
Mid-Speed Light Trucks
Hi-Speed Intersection Car
Hi-Speed Non-Intersection Car
Hi-Speed Hi-Severity Intersection
−3 −2 −1 0 1 2
SHAP value
Hi-Speed Non-Intersection Car
Hi-Speed Dense-Road Truck
Lo-Speed Non-Intersection Bus
Hi-Speed Intersection Car
Mid-Speed Light Trucks
Hi-Speed Hi-Severity Intersection
0.2 0.4 0.6 0.8 1.0

◂**Fig. 9** Comparison of SHAP value plots (left) and radar plots (right) for the five identified crash types. SHAP plots show the impact of each UMAP component on the model's crash type classification, with features sorted by their mean absolute SHAP value. Radar plots show the mean normalized distribution of UMAP components across each crash type

Fast Urban

This type is highly concentrated in the densest urban cores of the case study area. It is also characterized by crashes in urban cores that happen at high speeds, away from traditional intersections, and involve injuries, which can be inferred from high values of *Hi-Speed Non-Intersection Car*, *Hi-Speed Hi-Severity Intersection*, and *Hi-Speed Intersection Car*, as well as being less frequent in areas with high *Hi-Speed Dense-Road Truck* values.

Slow Urban

Concentrated in and around urban centers, this type presents a contrasting profile. It is clearly linked to *Lo-Speed Non-Intersection Bus* and occurs less frequently in locations with high *Hi-Speed Dense-Road Truck* and *Mid-Speed Light Trucks*. This suggests crashes occurring in areas with lower-speeds, with a notable presence of buses, and significantly less truck and light truck/SUV/van traffic, which are characteristic of urban and inner-suburban local road networks.

Fast Suburban

With a broad distribution across all three states (similar to *Mixed Suburban* but more dispersed), census tracts of this type exhibit high values on both *Hi-Speed Dense-Road Truck* and *Hi-Speed Intersection Car*, while being less associated with *Lo-Speed Non-Intersection Bus* and *Mid-Speed Light Trucks*. This combination indicates crashes on higher-speed roadways with a mix of vehicle types, but less frequent involvement of buses and light trucks/SUVs/vans, a pattern that aligns with suburban highways carrying a blend of commuting and through traffic.

Classification Performance Confirms Typology Separability and Reveals Key Predictors

The fitted LightGBM classification model predicted crash types with an accuracy of 96.06% on the test set. Figure 10a shows the confusion matrix, and Fig. 10b shows the receiver operating characteristics (ROC) curves. The high Area Under the Curve (AUC) scores (0.993–1.000) and strong precision, recall, and F1-scores (0.96–0.97 macro-averaged, Table 7) indicate excellent discrimination between the five types based on the original 25 features. This high level of predictive performance confirms that the five crash types identified via the unsupervised UMAP–GMM process are robustly separable and represent distinct patterns based on the original aggregated crash characteristics.

Figure 10c shows the feature importance. The mean speed limit within a census tract was the strongest predictor. This indicates that the prevailing roadway design

Table 9 Summary of crash type spatial distribution and characteristics

Type	No. tracts	Spatial distribution	Key crash characteristics
Fast Rural	376	Distributed across all three states, in areas surrounding urban centers and extending into less densely populated regions	Injury at intersections on higher-speed roadways with truck involvement, typical of suburban–rural transitions
Fast Suburban	542	Broad distribution across all three states but encompassing a wider range of suburban/exurban settings	Higher-speed roadways, often involving a mix of vehicles, but less likely to involve buses or light trucks, common on suburban highways
Fast Urban	45	Highly concentrated in the densest urban cores of the study area	At high speeds away from traditional intersections in the urban core
Mixed Suburban	545	Widespread across the three states, but less concentrated in dense urban cores. More prevalent in suburban/exurban areas	Higher-speed roadways in suburban/exurban settings with high truck involvement
Slow Urban	1152	Concentrated in and around urban centers	Lower-speed roads in urban/inner suburban areas, higher bus involvement, and lower truck and light-truck involvement

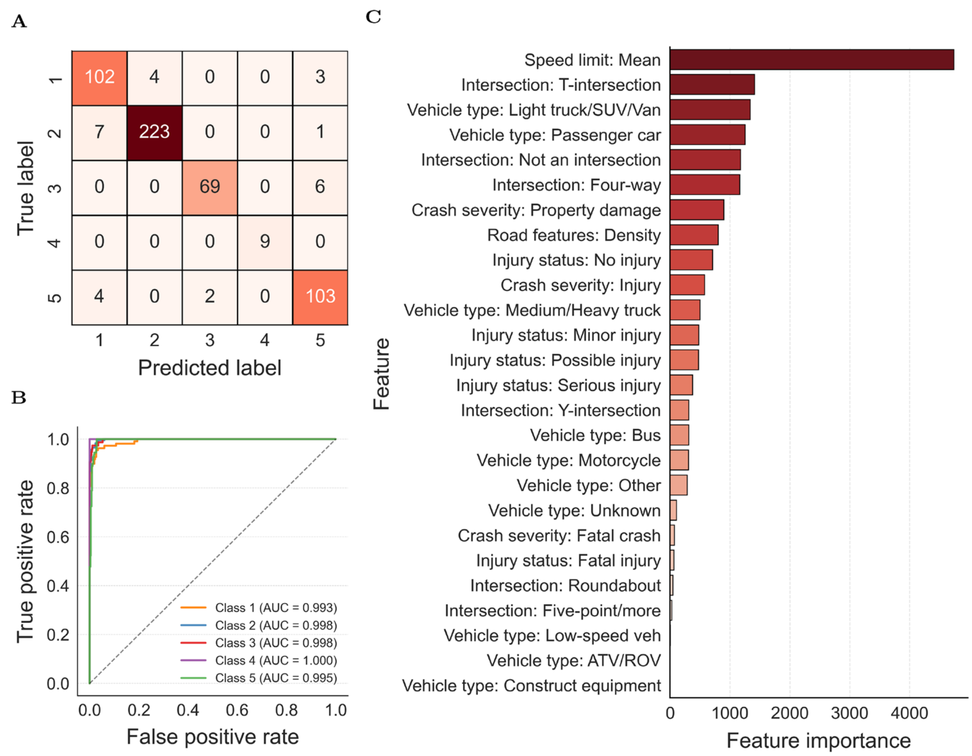


Fig. 10 LightGBM classifier validation test results: **a** Confusion matrix, **b** ROC curves for each crash type, and **c** Feature importance

and traffic flow characteristics are fundamental in influencing the types of crashes that occur within an area. Higher speed limits generally indicate a greater presence of higher-capacity roads, designed for faster and potentially higher-volume traffic, which can alter crash risk and severity.

The next most important features relate to the network characteristics, such as the road density or the intersection type (T-intersections, four-way intersections, and non-intersections). The prevalence of a type of intersection within a census tract was a key differentiator of the crash types, indicating that the model successfully captured the influence of roadway geometry and traffic control on crash patterns.

The presence of different vehicle types (Light truck/SUV/Van, Passenger car) was also a major factor. This suggests that the mix of vehicles using the roadways within a census tract, which are typically influenced by factors such as population size, land use, demographics, and commuting patterns, contributes to the specific crash characteristics observed. Finally, the model also relied on crash severity and injury status. This suggests that the risk of specific crash outcomes varies across census tracts.

Crash Typology Enables Targeted Safety Interventions

The crash typology provides a framework for developing targeted safety interventions. Understanding the dominant characteristics derived from the latent UMAP components and the spatial distributions of the crash types allows transportation planners and policymakers to tailor mitigation strategies to the specific risk factors present in different contexts. This targeted approach supports more efficient resource allocation and potentially greater impacts on crash reduction.

To demonstrate this, we identified six categories of safety interventions from the literature, ranging from geometric design changes to traffic control and enforcement. Figure 11 visualizes the connections established through our analysis, mapping the influence of the key UMAP components onto the different crash types, and subsequently linking those crash types to the relevant high-level intervention categories commonly employed. The width of the flows in the diagram indicates the strength of the influence of the feature (positive or negative SHAP value) or the number of specific interventions applied within a category for a given type.

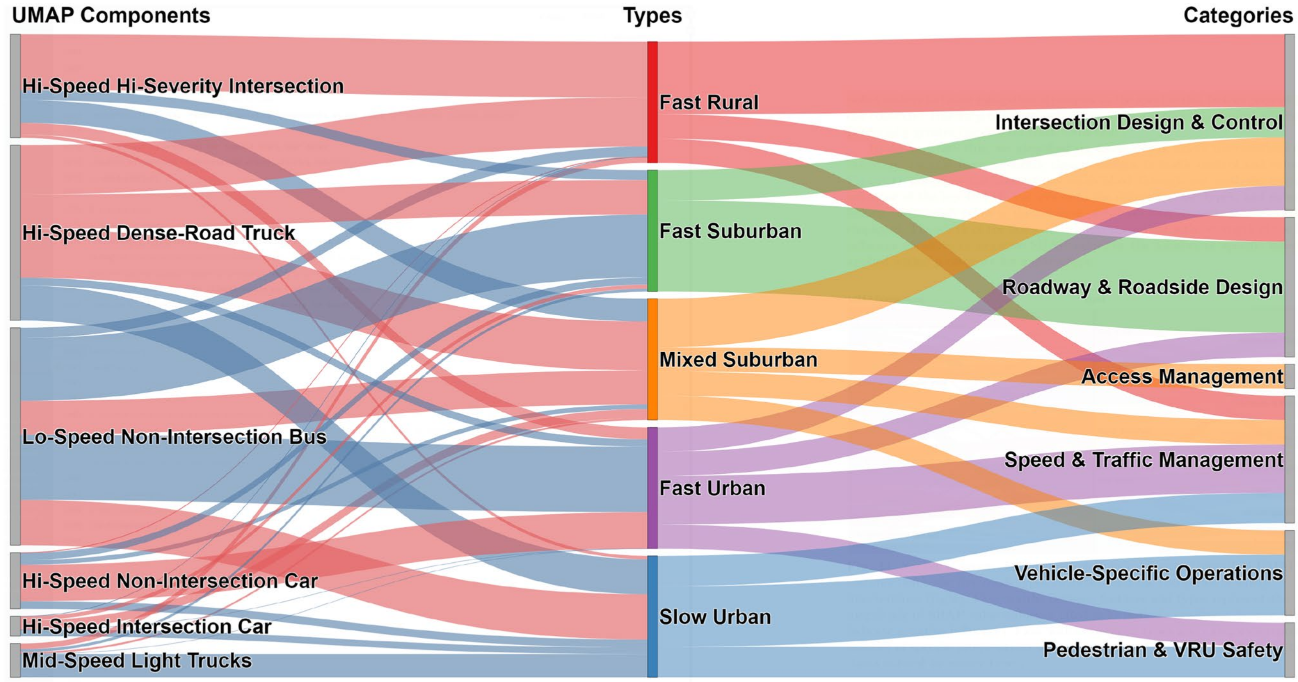


Fig. 11 Sankey diagram showing the relationship between identified UMAP components (left), the five crash types (middle), and the categories of safety interventions (right). Flow widths from components to types represent the magnitude of SHAP value influence (red for positive, blue for negative). Flows from types to intervention categories indicate the number of recommended interventions (colored by crash type)

To provide actionable detail beyond these categories, we then examined 33 specific examples of evidence-based interventions tailored to the characteristics of each crash type. Table 10 summarizes the interventions by type. For example, we specify countermeasures such as modern roundabout conversions for *Fast Rural* areas influenced by high-speed intersection risks, or improved pedestrian crossing facilities for *Slow Urban* areas characterized by bus and pedestrian activity. Each specific intervention is accompanied by a description of the mechanism and risks it addresses, linking to the underlying crash characteristics, and is supported by citations to established safety resources, such as FHWA guides, technical reports, and the Highway Safety Manual (HSM) (FHWA 2025; AASHTO 2010). The risk factors addressed are related to the typology characteristics identified in "Latent Crash Components are Characterized by Speed, Network, and Vehicle Type" section and "Crash Types are Defined by Spatial Patterns, Roadway Design, and Vehicle Interaction" section. Sources provide evidence and implementation guidance for the specific intervention. Using this typology, agencies can develop context-specific safety plans informed by the unique crash profile of different areas within their jurisdiction.

Limitations

The aggregation of crash data to the census tract level, while necessary to develop area-based types, is subject to the Modifiable Areal Unit Problem (MAUP) (Openshaw 1979). This means that the specific boundaries (zoning effect) and the scale (scale effect) of the census tracts can influence the resulting patterns and typology. Consequently, significant variation in crash characteristics within a single census tract may be masked. Future research should explore the sensitivity of these findings to different spatial aggregation units, such as uniform grid cells or functionally defined road segments, to better understand the impact of MAUP and intra-unit heterogeneity.

Another limitation of this study is its reliance on crash data for a single year (2023). Although this provides a valuable snapshot, it does not capture potential temporal dynamics in crash patterns, which can evolve due to changes in infrastructure, traffic volumes, travel behavior, or policy interventions over time. Acquiring and standardizing comprehensive high-quality crash data across multiple jurisdictions and years presented significant challenges, which constrained the temporal scope of this study. Future research, however,

Table 10 Detailed interventions, associated mechanisms and risks addressed, and sources for the five identified crash types

Type	Intervention	Mechanism and risks addressed
Fast Rural	(1) Modern roundabout conversion (FHWA 2025; U.S. Department of Transportation 2017; Wang and Cicchino 2022)	(a) High speeds and severe angle crashes at intersections (b) Elimination of head-on/right-angle conflicts (c) Truck accommodation with proper design
	(2) Restricted Crossing U-Turn (RCUT) (Hummer et al. 2014; Lindley 2008)	(a) Severe crossing conflicts on high-speed roads (b) Elimination of direct crossing/left-turn conflicts (c) Simplified driver decisions crossing fast traffic
	(3) Addition of left/right turn lanes (Lindley 2008; AASHTO 2010; Bahar et al. 2008)	(a) Rear-end crash reduction on high-speed roads (b) Separation of slower turning vehicles from through traffic
	(4) Enhanced intersection signage (oversized, LEDs) (FHWA 2015)	(a) Improved intersection awareness/yielding
	(5) Intersection Conflict Warning System (ICWS) (Tian et al. 2022)	(a) Real-time conflict warnings
	(6) Road lighting (FHWA 2015)	(a) Improved visibility day/night
	(7) Shoulder/centerline rumble strips (FHWA 2025, 2021)	(a) Run-off-road and head-on crash reduction (b) Inattentive driver alerts
	(8) High Friction Surface Treatment (HFST) (FHWA 2025, 2014)	(a) Increased pavement grip on curves/approaches
Fast Suburban	(9) Shoulder and centerline rumble strips (FHWA 2025, 2021)	(a) Run-off-road and head-on crash reduction (b) Inattentive driver alerts (lane departure)
	(10) Addition of median barriers (esp. cable) (Marzougui et al. 2012)	(a) Prevention of cross-median head-on collisions
	(11) Clear zones (FHWA 2011)	(a) Run-off-road crash severity mitigation
	(12) High-Friction Surface Treatment (HFST) (on curves/ramps) (EDC 2024)	(a) Loss-of-control crash reduction on curves/ramps (b) Increased pavement grip, especially in wet conditions
	(13) Intersection control upgrades (roundabout, signalization w/ warning, RCUT) (FHWA 2024, 2022)	(a) High-speed intersection crash mitigation b) Forced speed reduction or clear right-of-way assignment
Fast Urban	(14) Road diets/lane narrowing (FHWA 2025, 2021, 2017a)	(a) Reduced excessive speeds on urban arterials
		(b) Fewer multi-lane conflicts (sideswipes, lane changes)
		(c) Visual narrowing encourages slower speeds
	(15) Speed tables/raised crosswalks (FHWA 2017b)	(a) Forced slower speeds at intersection, mid-block
		(b) Reduced pedestrian collision severity
	(16) Pedestrian refuge islands (NACTO 2025)	(a) Mid-crossing safety on wide, fast streets
	(17) Leading Pedestrian Intervals (LPIs) (NACTO 2025)	(a) Temporal separation of pedestrians and turning vehicles
	(18) Automated speed enforcement (speed cameras) (AAA et al. 2021; Hu and McCartt 2016)	(a) Direct targeting and deterrence of excessive speeding
		(b) Reduction in high-severity crashes tied to speed
	(19) Intersection simplification/control upgrade (roundabout, signalization at odd junctions) (FHWA 2025; National Academy 2007)	(a) Crash mitigation at non-traditional/complex locations
		(b) Conflict point reduction
		(c) Order/speed control at complex merges/junctions
Mixed Suburban	(20) Raised median/restricted left turns (Dixon et al. 2016)	(a) Reduced mid-block conflict points on high-density arterials (b) Elimination of uncontrolled left turns across high-speed traffic
	(21) Dedicated turn lanes (AASHTO 2010; Simpson and Troy 2015)	(a) Rear-end/spillback reduction on high-speed arterials
	(22) Protected signal phasing (left turns) (AASHTO 2010; Simpson and Troy 2015)	(a) Separation of turning vehicles from opposing traffic (b) Near-elimination of left-turn angle crashes

Table 10 (continued)

Type	Intervention	Mechanism and risks addressed
	(23) Optimized signal timing (FHWA 2008)	(a) Dilemma zone issue reduction (b) Reduced red-light running
	(24) Red-light cameras (Hu and Cicchino 2017)	(a) Violation deterrence, reducing angle crashes (b) Reduced red-light running
	(25) Context-appropriate speed limits (FHWA 2024)	(a) Management of excessive speeds on suburban arterials
	(26) Road diets (Persaud and Lyon 2010)	(a) Traffic calming, reduced multi-lane conflicts
	(27) Geometric design for trucks (AASHTO 2018)	(a) Adequate turning radii/space (b) Mitigation of physical risks from heavy truck presence
	(28) Designated truck routes/delivery zones (FHWA 2012)	(a) Conflict minimization via routing
Slow Urban	(29) Bus stop design (far-side stops, bus bulbs) (National Research Council 1994)	(a) Minimized conflicts from bus stopping maneuvers (b) Reduced sightline obstructions and awkward merges
	(30) Dedicated bus lanes (Matters 2025)	(a) Separation of buses from general traffic, reducing conflicts
	(31) Transit signal priority (TSP) (Smith et al. 2005; Matters 2025)	(a) Smoother bus operations, fewer abrupt maneuvers
	(32) Pedestrian crossing enhancements at bus stops (FHWA 2025)	(a) Safe crossing provision for transit users near buses (b) Improved visibility and driver yielding behavior
	(33) Curb management/parking enforcement (National Association of City Transportation Officials 2013; Agency SFMT 2013)	(a) Conflict reduction from illegal/double parking near bus routes (b) Maintained clear sightlines and access for buses

should aim to incorporate multi-year datasets to explore the temporal stability of the identified crash types and investigate how crash profiles within census tracts change over time.

Conclusion

In this research, we developed a novel typology of census tracts based on their crash profiles, using three New England states as a case study. This typology builds on conventional incident-focused analysis or hotspot identification by clustering geographic areas based on their multivariate crash profiles. We identified five distinct crash types present across the region through an unsupervised learning pipeline that integrates UMAP and GMM. These types are characterized by unique combinations of factors, such as the mean of posted speed limits associated with each crash within the tract, the prevalence of intersections, vehicle participation, and the resulting outcomes of severity of the crash, which reveal predictable spatial safety patterns.

We found that the crash types are conceptually distinct and statistically separable and predictable ("Classification Performance Confirms Typology Separability and Reveals Key Predictors" section). The LightGBM classification model achieved 96% accuracy predicting the typology membership from original features, which proved the value of the integrated UMAP-GMM-classifier pipeline for extracting meaningful geographic safety patterns from complex crash data. Our approach provides a valuable framework for transportation safety analysis by offering a complementary perspective to traditional methods. Instead of focusing solely on high-risk locations or individual incidents, it characterizes different types of crash-prone environments based on their underlying risk factor interactions. The multi-state scope facilitates the discovery of potentially transferable regional patterns, while the SHAP-based interpretation ensures that the insights derived from the machine learning models are explainable and actionable for safety practitioners.

We used the proposed typology to identify high-risk areas based on their predicted crash type, allowing us to prioritize safety investments where they were most needed. More importantly, we demonstrated how the typology can guide the development of 33 evidence-based targeted interventions. Instead of applying generic safety measures, we linked each crash type to specific risk factors and recommended interventions tailored to those patterns.

There are several promising avenues for future work. First, investigating the transferability of this typology to other regions would be valuable. Although the specific crash types and their characteristics may vary, the underlying principle that area-level factors drive spatial patterns in crash risk is likely to be generalizable. Second, incorporating additional data sources, such as detailed roadway inventory data, traffic volume information, and land use patterns, could further refine the typology and improve the precision of risk prediction. Third, evaluating the specific intervention strategies tailored to each crash type is a possible next step. This could involve collaborations with transportation agencies to implement pilot projects and assess their effectiveness.

This research provides a powerful novel tool to understand and address road safety. By shifting the focus from individual crashes to area-level patterns, the typology offers a more holistic and actionable framework to improve safety outcomes and develop targeted interventions that can save lives and reduce injuries on our roadways.

Appendix: Data Standardization and Aggregation Details

Data Integration and Preparation

To construct a unified dataset suitable for cross-state analysis, crash data from Massachusetts, Connecticut, and Vermont for the year 2023 underwent a rigorous integration and preparation process. This process involved several key steps designed to maximize consistency and comparability, addressing potential variations in state-specific reporting practices, field definitions, and coding schemes.

1. *Initial state-level processing and column standardization:* For each state's raw dataset, we first performed initial cleaning, including handling missing values where appropriate and addressing inconsistencies in column naming conventions (e.g., case sensitivity, special characters). We identified conceptually equivalent variables across the three datasets, even when field names differed. A common schema was established, and state-specific column names were systematically mapped to these standardized names. For instance, variations like "Town Name" (CT), "CITY_TOWN_NAME" (MA), and "City or Town" (VT) were all standardized to "city_town". Similarly, date and time fields were standardized (e.g., to "crash_datetime"). Common temporal variables, such as "hour_of_day", "day_of_week", and "month", were derived consistently from the standardized timestamp information.
2. *Core variable selection:* Based on the requirements of the crash typology analysis and data availability across all three states, we selected a core set of variables. This set included identifiers, spatio-temporal context, crash outcomes, environmental conditions, vehicle information, roadway characteristics, and basic contributing factors. The focus was on variables that could be reliably standardized and were directly relevant for aggregate spatial analysis, rather than including all fields present in the original state reports (e.g., person-specific details not used in the aggregation). The key selected variable categories included:
   - Crash context: Unique crash identifier, date, time, hour, day of week, month, latitude, longitude, city/town identifier.
   - Crash outcome: Crash severity, highest injury level.
   - Environmental conditions: Light conditions, weather conditions.
   - Vehicle information: Vehicle type(s) involved.
   - Roadway characteristics: Relation to intersection, intersection type, traffic control type, posted speed limit.
3. *Cross-state value standardization:* This step addressed differing coding systems and definitions within the selected variables. We developed and applied explicit mapping functions based on reviewing state data dictionaries to standardize categorical values.
   - *Crash severity:* State-specific codes (e.g., numerical codes in MA/VT, descriptive terms in CT) representing different injury levels were mapped onto a unified scale based on KABCO principles (Flannagan et al. 2021) (Fatal Injury, Suspected Serious Injury, Suspected Minor Injury, Possible Injury, Property Damage Only). This ensured consistent interpretation of severity across states before aggregation into features such as "Crash Severity: Fatal crash proportion" and "Crash Severity: Any Injury crash proportion".
   - *Vehicle type:* Diverse state codes and descriptions for vehicle types (e.g., "Passenger car", "Automobile", "Sedan", "Station Wagon", various truck classifications) were consolidated into standardized, broader categories relevant for safety analysis (e.g., "Passenger Car", "Light Truck/SUV/Van", "Medium/Heavy Truck", "Motorcycle", "Bus", "Other/Unknown"). This involved creating comprehensive mapping dictionaries.
   - *Intersection relation and type:* Terms describing crash location relative to junctions (e.g., "Intersection", "Intersection Related", "Non-Intersection") and specific junction types (e.g., "Four-way intersection", "T-Intersection", "Roundabout", various signal/sign control combinations) were mapped to consistent categories (e.g., "Intersection Type: Four-way", "Intersection Type: T-intersection", "Intersection Relation: Non-intersection").
   - *Other categorical variables:* Similar mapping procedures were applied to other selected categorical

fields to ensure uniform definitions before aggregation.

4. *Data merging:* Following the standardization of column names and value codes, the processed and standardized datasets from the three states were concatenated into a single unified dataset containing individual crash records ready for the aggregation stage.

Data aggregation

To prepare the data for typology analysis, we aggregated information to the census tract level. This involved summarizing the characteristics of individual crash records within each tract and incorporating spatial context features derived from external data sources.

For categorical variables derived from the standardized crash records (e.g., “Crash severity” categories, “Vehicle type” categories, “Intersection type”), we calculated the proportion of crashes belonging to each category within each census tract. For instance, for the “Vehicle type” category "Passenger car," we calculated the proportion of all crashes within a tract that involved at least one passenger car. For numerical variables derived from crash records (specifically, the posted “Speed limit” associated with crashes), we calculated the mean value within each census tract.

We also incorporated a measure of roadway infrastructure intensity using external data. Census tract boundaries for all three states were obtained from the U.S. Census Bureau’s TIGER/Line shapefiles (2023). Road network data (primary and secondary roads) were also sourced from these TIGER/Line files. Road density was calculated for each census tract by summing the length of these roads within the tract boundary (using an appropriate projected coordinate system for accurate measurement) and dividing by the tract’s area.

This aggregation process resulted in the final dataset used for dimensionality reduction and clustering. It comprises 2,660 census tracts (rows) and 25 features (columns). These features include the aggregated crash characteristics (proportions and means calculated above) and the calculated “Road features: Density” metric representing network infrastructure intensity. Each feature represents an aggregated characteristic of the crashes occurring within, or a spatial property of, that census tract during 2023. The specific definitions of all 25 features are provided in Table 2.

Acknowledgements The authors would like to acknowledge the support of the New England University Transportation Center (NEUTC) for funding this research project.

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: JO; data collection: MA; algorithm development and coding: MA; analysis and interpretation of results: MA, JO; draft manuscript preparation: MA, JO. All authors reviewed the results and approved the final version of the manuscript.

Funding This research was funded by the US Department of Transportation via the New England University Transportation Center (Grant Number: 69A35523483)

Availability of data and materials The crash data used in this study were sourced from the official repositories of Massachusetts, Connecticut, and Vermont. These datasets are publicly accessible through the respective state portals. For reproducibility and further analysis, the processed data and associated code used in this research are available in a dedicated GitHub repository at https://github.com/narslab/spatial-crash-typology.

Declarations

Ethical Approval This study did not require ethical approval as it did not involve any human or animal subjects.

Conflict of Interest The authors declare that they have no conflict of interest.

References

AAA, Advocates for Highway and Auto Safety, Association, G.H.S., Insurance Institute for Highway Safety, Council, N.S.: Automated Enforcement Program Checklist: For Red Light Cameras and Automated Speed Enforcement (2021)

AASHTO: Highway Safety Manual (2010) https://www.highwaysafetymanual.org/Pages/default.aspx

AASHTO (2018) A policy on geometric design of highways and streets, 7th edn. AASHTO, Washington, DC

Abdel-Aty M, Haleem K (2011) Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accident Anal Prevent 43(1):461–470. https://doi.org/10.1016/j.aap.2010.10.002

Abdel-Aty M, Keller J, Brady PA (2005) Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transp Res Rec 1908(1):37–45. https://doi.org/10.1177/0361198105190800105

Agency SFMT (2013) Transit-Only Lane Enforcement. Technical report

Aguero-Valverde J, Jovanis PP (2006) Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Anal Prevent 38(3):618–625. https://doi.org/10.1016/j.aap.2005.12.006

Ali Y, Hussain F, Haque MM (2024) Advances, challenges, and future research needs in machine learning-based crash prediction models: a systematic review. Accident Anal Prevent 194:107378. https://doi.org/10.1016/j.aap.2023.107378

Aziz HMA, Ukkusuri SV, Hasan S (2013) Exploring the determinants of pedestrian-vehicle crash severity in New York City. Accident Anal Prevent 50:1298–1309. https://doi.org/10.1016/j.aap.2012.09.034

Bahar GB, Masliah M, Wolff R, Park P (2008) United States. Department of Transportation. Federal highway administration. Office of safety: desktop reference for crash reduction factors, Technical Report FHWA-SA-08-011

Bajada T, Attard M (2021) A typological and spatial analysis of pedestrian fatalities and injuries in Malta. Res Transp Econ 86:101023. https://doi.org/10.1016/j.retrec.2020.101023

Behboudi N, Moosavi S, Ramnath R (2024) Recent advances in traffic accident analysis and prediction: a comprehensive review of machine learning techniques
Bhowmik T, Yasmin S, Eluru N (2019) Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Anal Methods Accident Res 24:100107. https://doi.org/10.1016/j.amar.2019.100107
Bhowmik T, Yasmin S, Eluru N (2021) A new econometric approach for modeling several count variables: a case study of crash frequency analysis by crash type and severity. Transp Res Part B: Methodol 153:172–203. https://doi.org/10.1016/j.trb.2021.09.008
Bíl M, Andrášik R, Sedoník J (2019) A detailed spatiotemporal analysis of traffic crash hotspots. Appl Geogr 107:82–90. https://doi.org/10.1016/j.apgeog.2019.04.008
Breiman L (1997) Arcing the edge
Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
Campello RJGB, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G (eds) Advances in knowledge discovery and data mining. Springer, Berlin, Heidelberg, pp 160–172
Centers for Disease Control and Prevention: Motor Vehicle Crash Deaths—How is the US doing? MMWR (2016) Morbidity and Mortality Weekly Report, Vol. 71
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system, In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD ’16, pp. 785–794. Association for Computing Machinery, New York, NY. https://doi.org/10.1145/2939672.2939785
Cox DR (1958) The regression analysis of binary sequences. J Roy Stat Soc: Ser B (Methodol) 20(2):215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
CTDOT: Connecticut Crash Data Repository (2023) available at: https://www.ctcrash.uconn.edu/
Dadvar S, Lee Y-J, Shin H-S, Khodaparasti H (2020) Integrated highway safety information system (HSIS) datasets by means of a roadway safety data integrator (RSDI) tool. Data Brief 32:106154. https://doi.org/10.1016/j.dib.2020.106154
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc: Ser B (Methodol) 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
U.S. Department of Transportation: Roundabouts: An Informational Guide (2017)
Dixon K, Layton R, Butorac M, Ryus P, Gattis JL, Brown L, Huntington D (2016) Access management application guidelines. The National Academies Press, Washington, DC
Eckley DC, Curtin KM (2013) Evaluating the spatiotemporal clustering of traffic incidents. Comput Environ Urban Syst 37:70–81. https://doi.org/10.1016/j.compenvurbsys.2012.06.004
EDC F (2024) Innovation spotlight: high friction surface treatment
Eskandari Torbaghan M, Sasidharan M, Reardon L, Muchanga-Hvelplund LCW (2022) Understanding the potential of emerging digital technologies for improving road safety. Accident Anal Prevent 166:106543. https://doi.org/10.1016/j.aap.2021.106543
Farmer CM (2019) The effects of higher speed limits on traffic fatalities in the United States, 1993–2017
FHWA: Freight and Land Use Handbook (2012) available at: https://ops.fhwa.dot.gov/publications/fhwahop12006/
FHWA: High Friction Surface Treatments: Project Case Study. Technical report (2014)
FHWA: Longitudinal Rumble Strips and Stripes on Two-Lane Roads | FHWA (2021) https://highways.dot.gov/safety/proven-safety-countermeasures/longitudinal-rumble-strips-and-stripes-two-lane-roads
FHWA: Intersection Safety | FHWA (2022) https://highways.dot.gov/safety/intersection-safety
FHWA: Highway Safety Manual | FHWA (2024) https://highways.dot.gov/safety/data-analysis-tools/highway-safety-manual
FHWA: Intersection Control Evaluation | FHWA (2025) available at: https://highways.dot.gov/safety/intersection-safety/ice
FHWA: Proven Safety Countermeasures | FHWA (2025) available at: https://highways.dot.gov/safety/proven-safety-countermeasures
FHWA: Road Diets (Roadway Reconfiguration) | FHWA (2021) https://highways.dot.gov/safety/proven-safety-countermeasures/road-diets-roadway-reconfiguration
FHWA: Roadside Design Guide, 4th Edition | FHWA (2011) https://highways.dot.gov/safety/pedestrian-bicyclist/safety-tools/51-52-roadside-design-guide-4th-edition
FHWA: Safe Transportation for Every Pedestrian (STEP) | FHWA (2025) available at: https://highways.dot.gov/safety/pedestrian-bicyclist/step
FHWA: Speed Management | FHWA (2024) available at: https://highways.dot.gov/safety/speed-management
FHWA: Summary Report: Development of a Speeding-Related Crash Typology - FHWA-HRT-10-039 (2010) available at: https://www.fhwa.dot.gov/publications/research/safety/10039/
FHWA: Traffic Calming ePrimer | FHWA (2017) https://highways.dot.gov/safety/speed-management/traffic-calming-eprimer
FHWA: Traffic Signal Timing Manual (2008) available at: https://ops.fhwa.dot.gov/publications/fhwahop08024/chapter5.htm
FHWA: Unsignalized Intersection Improvement Guide (UIIG) | FHWA (2015) available at: https://highways.dot.gov/safety/data-analysis-tools/rsdp/rsdp-tools/unsignalized-intersection-improvement-guide-uiig
FHWA: Vision Zero in New York City | FHWA (2017) https://highways.dot.gov/safety/zero-deaths/vision-zero-new-york-city
Fisa R, Musukuma M, Sampa M, Musonda P, Young T (2022) Effects of interventions for preventing road traffic crashes: an overview of systematic reviews. BMC Public Health 22(1):513. https://doi.org/10.1186/s12889-021-12253-y
Flannagan CA, Rupp JD, Mann NC (2021) Development of a comprehensive approach for serious traffic crash injury measurement and reporting systems. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/26305
Friedman J (2000) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232. https://doi.org/10.1214/aos/1013203451
Gao X, Jiang X, Haworth J, Zhuang D, Wang S, Chen H, Law S (2024) Uncertainty-aware probabilistic graph neural networks for road-level traffic crash prediction. Accident Anal Prevent 208:107801. https://doi.org/10.1016/j.aap.2024.107801
Gu Y, Liu D, Arvin R, Khattak AJ, Han LD (2023) Predicting intersection crash frequency using connected vehicle data: A framework for geographical random forest. Accident Anal Prevent 179:106880. https://doi.org/10.1016/j.aap.2022.106880
Guo H, Wang Z, Sherony R, Bao S (2023) A data-driven framework of crash scenario typology development for child vulnerable road users in the U.S. SAE Technical Paper 2023-01-0787, SAE International, Warrendale, PA. https://doi.org/10.4271/2023-01-0787
Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417–441. https://doi.org/10.1037/h0071325
Hu W, Cicchino J (2017) Effects of Turning on and off Red Light Cameras on Fatal Crashes in Large U.S. Cities. available at: https://www.iihs.org/topics/bibliography/ref/2121
Hu W, McCartt A (2016) Effects of Automated Speed Enforcement in Montgomery County, Maryland, on Vehicle Speeds, Public

Opinion, and Crashes. available at: https://www.iihs.org/topics/bibliography/ref/2097
Hummer J, Ray B, Daleiden A, Jenior P, Knudsen J, Kittelson Associates I (2014) Restricted crossing u-turn: informational guide. Technical Report FHWA-SA-14-070
Imprialou M, Quddus M (2019) Crash data quality for road safety research: current state and future directions. Accident Anal Prevent 130:84–90. https://doi.org/10.1016/j.aap.2017.02.022
Jia R, Khadka A, Kim I (2018) Traffic crash analysis with point-of-interest spatial clustering. Accident Anal Prevent 121:223–230. https://doi.org/10.1016/j.aap.2018.09.018
Jin J, Huang H, Yuan C, Li Y, Zou G, Xue H (2023) Real-time crash risk prediction in freeway tunnels considering features interaction and unobserved heterogeneity: a two-stage deep learning modeling framework. Anal Methods Accident Res 40:100306. https://doi.org/10.1016/j.amar.2023.100306
Jin J, Liu P, Huang H, Dong Y (2024) Analyzing urban traffic crash patterns through spatio-temporal data: a city-level study using a sparse non-negative matrix factorization model with spatial constraints approach. Appl Geogr 172:103402. https://doi.org/10.1016/j.apgeog.2024.103402
Jin J, Huang H, Li Y, Dong Y, Zhang G, Chen J (2025) Variable speed limit control strategy for freeway tunnels based on a multi-objective deep reinforcement learning framework with safety perception. Expert Syst Appl 267:126277. https://doi.org/10.1016/j.eswa.2024.126277
Kamh H, Alyami S, Khattak A, Alyami M, Almujibah H (2024) Exploring road traffic accidents hotspots using clustering algorithms and GIS-based spatial analysis. IEEE Access 13:1. https://doi.org/10.1109/ACCESS.2024.3443245
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3147
Kim D-G, Washington S, Oh J (2006) Modeling crash types: new insights into the effects of covariates on crashes at rural intersections. J Transp Eng 132(4):282–292. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:4(282)
Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43(1):59–69. https://doi.org/10.1007/BF00337288
Kumar A, Kalapurayil HKM (2022) Modeling crash severity and collision types using machine learning, Publications
Li X, Liu J, Zhang Z, Parrish A, Jones S (2021) A spatiotemporal analysis of motorcyclist injury severity: Findings from 20 years of crash data from Pennsylvania. Accident Anal Prevent 151:105952. https://doi.org/10.1016/j.aap.2020.105952
Lindley J (2008) Guidance memorandum on consideration and implementation of proven safety countermeasures | FHWA. available at: https://highways.dot.gov/safety/proven-safety-countermeasures/guidance-memorandum-consideration-and-implementation-proven
Lin L, Wang Q, Sadek AW (2018) State University of New York at Buffalo, Department of Civil, S. and Environmental Engineering: Novel Machine Learning Methods for Accident Data Analysis. Technical report
Lu J, Gan A, Haleem K, Wu W (2013) Clustering-based roadway segment division for the identification of high-crash locations. J Transp Saf Secur 5(3):224–239. https://doi.org/10.1080/19439962.2012.730118
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions, In: Advances in neural information processing systems, vol. 30, pp. 4765–4774. Curran Associates, Inc., Long Beach, California
Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intel 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9
Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605
MacQueen J (1967) Some methods for classification and analysis of multivariate observations, In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1: statistics 5.1, pp. 281–298
Marzougui D, Mahadevaiah U, Tahan F, Kan CDS, McGinnis R, Powers R (2012) Guidance for the selection, use, and maintenance of cable barrier systems. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/22717
Mason L, Baxter J, Bartlett P, Frean M (1999) Boosting algorithms as gradient descent, In: Advances in neural information processing systems, vol. 12, pp. 512–518. MIT Press, Denver, Colorado
MassDOT: MassDOT: Crash Data Portal (2023) available at: https://apps.impact.dot.state.ma.us/cdp/home
Matters TfL|EJ (2025) Bus safety. available at: https://www.tfl.gov.uk/corporate/publications-and-reports/bus-safety-data
McInnes L, Healy J, Melville J (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv
Mekonnen TH, Tesfaye YA, Moges HG, Gebremedin RB (2019) Factors associated with risky driving behaviors for road traffic crashes among professional car drivers in Bahirdar city, northwest Ethiopia, 2016: a cross-sectional study. Environ Health Prevent Med 24:17. https://doi.org/10.1186/s12199-019-0772-1
Mohammed S, Alkhereibi AH, Abulibdeh A, Jawarneh RN, Balakrishnan P (2023) GIS-based spatiotemporal analysis for road traffic crashes; in support of sustainable transportation planning. Transp Res Interdiscip Persp 20:100836. https://doi.org/10.1016/j.trip.2023.100836
NACTO: Urban Street Design Guide (2025)
National Academy (2007) Roundabouts in the United States. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/23216
National Association of City Transportation Officials: Urban Street Design Guide (2013)
National Research Council (ed.) (1994) Transportation environmental issues: air, noise, water, mitigation processes and alternative fuels. Transportation Research Record, vol. 1444. National Academy Press, Washington, DC
NHTSA: 2021 Data: Passenger Vehicles (2021)
Nie K, Wang Z, Du Q, Ren F, Tian Q (2015) A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: a case study from Wuhan, China. Sustainability 7(3):2662–2677. https://doi.org/10.3390/su7032662
Nitsche P, Thomas P, Stuetz R, Welsh R (2017) Pre-crash scenarios at road junctions: a clustering method for car crash data. Accident Anal Prevent 107:137–151. https://doi.org/10.1016/j.aap.2017.07.011
Openshaw S (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem
Paikari E, Moshirpour M, Alhajj R, Far BH (2014) Data integration and clustering for real time crash prediction, In: Proceedings of the 2014 IEEE 15th international conference on information reuse and integration (IEEE IRI 2014), pp 537–544. https://doi.org/10.1109/IRI.2014.7051936
Persaud B, Lyon C (2010) Evaluation of lane reduction "road diet" measures on crashes - FHWA-HRT-10-053. available at: https://www.fhwa.dot.gov/publications/research/safety/10053/
Prasannakumar V, Vijith H, Charutha R, Geetha N (2011) Spatio-temporal clustering of road accidents: GIS based analysis and

assessment. Procedia - Soc Behav Sci 21:317–325. https://doi.org/10.1016/j.sbspro.2011.07.020

Rella Riccardi M, Galante F, Scarano A, Montella A (2022) Econometric and machine learning methods to identify pedestrian crash patterns. Sustainability 14(22):15471. https://doi.org/10.3390/su142215471

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408. https://doi.org/10.1037/h0042519

Schneider WH, Savolainen PT (2011) Comparison of severity of motorcyclist injury by crash types. Transp Res Rec 2265(1):70–80. https://doi.org/10.3141/2265-08

Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319. https://doi.org/10.1162/089976698300017467

Simpson CL, Troy SA (2015) Safety effectiveness of flashing yellow arrow: evaluation of 222 signalized intersections in North Carolina. Transp Res Rec: J Transp Res Board 2492(1):46–56. https://doi.org/10.3141/2492-05

Smith HR, Hemily B, Ivanovic M, Fleming G (2005) Transit signal priority (TSP), a planning and implementation handbook

Sullivan S (2018) Roadway intersection type

Sun M, Sun X, Shan D (2019) Pedestrian crash analysis with latent class clustering method. Accident Anal Prevent 124:50–57. https://doi.org/10.1016/j.aap.2018.12.016

Taylor B, Rehm J (2012) The relationship between alcohol consumption and fatal motor vehicle injury: high risk at low alcohol levels. Alcohol Clin Exp Res 36(10):1827–1834. https://doi.org/10.1111/j.1530-0277.2012.01785.x

Tian D, Gerberich SG, Kim H, Ryan AD, Erickson DJ, Morris NL (2022) Evaluation of the efficacy of an intersection conflict warning system at two-way stop-controlled rural intersections: difference-in-differences and triple-difference analytical approaches. Injury Prevent 28(3):204–210

Uchida N, Kawakoshi M, Tagawa T, Mochida T (2010) An investigation of factors contributing to major crash types in Japan based on naturalistic driving data. IATSS Res 34(1):22–30. https://doi.org/10.1016/j.iatssr.2010.07.002

US Department of Transportation: The Roadway Safety Problem | US Department of Transportation (2023) available at: https://www.transportation.gov/NRSS/SafetyProblem

VTrans: Vermont Agency of Transportation: Public Crash Query Tool (2023) available at: https://apps.vtrans.vermont.gov/crashpublicquerytool/

Wang M, Lee W-C, Liu N, Fu Q, Fujan W, Ge Y (2025) A data-driven deep learning framework for prediction of traffic crashes at road intersections. Appl Sci 15:752

Wang J, Cicchino J (2022) Safety effects of roundabout conversions in Carmel, Indiana, the Roundabout City. available at: https://www.iihs.org/topics/bibliography/ref/2235

WHO: Global Status Report on Road Safety (2023) available at: https://www.who.int/publications/i/item/9789240086517

Zhang Z, Liu J, Li X, Fu X, Yang C, Jones S (2023) Localizing safety performance functions for two-way STOP-controlled (TWST) three-leg intersections on rural two-lane two-way (TLTW) roadways in Alabama: A geospatial modeling approach with clustering analysis. Accident Anal Prevent 179:106896. https://doi.org/10.1016/j.aap.2022.106896

Zhang Z, Xu N, Liu J, Jones S (2024) Exploring spatial heterogeneity in factors associated with injury severity in speeding-related crashes: an integrated machine learning and spatial modeling approach. Accident Anal Prevent 206:107697. https://doi.org/10.1016/j.aap.2024.107697

Zhou R, Huang H, Lee J, Huang X, Chen J, Zhou H (2023) Identifying typical pre-crash scenarios based on in-depth crash data with deep embedded clustering for autonomous vehicle safety testing. Accident Anal Prevent 191:107218. https://doi.org/10.1016/j.aap.2023.107218