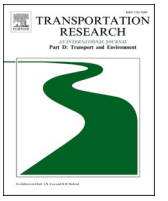
Transportation Research Part D 139 (2025) 104538

Contents lists available at ScienceDirect

Transportation Research Part D

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



Revealing equity gaps in pedestrian crash data through explainable artificial intelligence clustering

Jinli Liu [a,*], Gian Antariksa [b], Shriyank Somvanshi [b], Subasish Das [b]

[a] Geography and Environmental Studies, Texas State University, 601 University Drive, San Marcos, TX 78666, United States
[b] Ingram School of Engineering, Texas State University, 601 University Drive, San Marcos, TX 78666, United States

ARTICLE INFO

Keywords:
Pedestrian crash
CatBoost
Crash severity
Explainable AI
Hierarchical clustering
Social disparity

ABSTRACT

Pedestrian crashes represent a critical traffic safety issue, often resulting in fatal outcomes and raising significant equity concerns. This study analyzed detailed records of pedestrian-involved crashes in California from 2018 to 2021, employing a novel clustering framework enhanced by the SHapley Additive exPlanations approach. The proposed method significantly enhanced interpretability by effectively capturing complex non-linear relationships and interactions among features. The results indicate that impairment status and lighting conditions are pivotal in severe crash outcomes, while broader societal and demographic factors are more substantially associated with less severe cases. Non-injury pedestrian crashes tend to occur in less underserved, more resilient communities, whereas fatal crashes are more common in underserved communities with poor lighting and incomplete pedestrian infrastructure, particularly when pedestrians are under the influence of drugs or alcohol. The findings underscore the necessity for developing comprehensive safety measures that not only address situational risks but also consider broader societal conditions.

1. Introduction

Pedestrian safety is a pressing public health concern. According to the World Health Organization, 92 % of global road fatalities occur in low- and middle-income countries, which only possess 60 % of the world's vehicles (World Health Organization, 2024). These fatalities predominantly affect vulnerable road users such as pedestrians, cyclists, and motorcyclists, who account for over half of all road traffic deaths. In the U.S., data from the Centers for Disease Control and Prevention reveal that nearly 8,000 pedestrians were fatally injured in motor vehicle crashes in 2021 alone (CDC, 2023), with pedestrians comprising one-sixth of traffic-related deaths. Most pedestrian fatalities occur on high-capacity urban roads, and in 60 % of these cases in 2020, alcohol consumption by either the driver or the pedestrian was involved. Additionally, non-Hispanic American Indian and Alaska Native, as well as Black individuals, experienced the highest pedestrian fatality rates among all racial and ethnic groups in 2021. Addressing this issue, the Department of Transportation (DOT) is committed to maintaining a safe, efficient, and accessible transportation system nationwide (US Department of Transportation, 2024). This commitment includes bridging gaps in transportation infrastructure and public services to ensure equitable access for all communities.

Pedestrian safety continues to be a critical issue in urban transportation, as pedestrians face significant risks of injury and fatality in

* Corresponding author.
E-mail addresses: j_l848@txstate.edu (J. Liu), gph27@txstate.edu (G. Antariksa), shriyank@txstate.edu (S. Somvanshi), subasish@txstate.edu (S. Das).

https://doi.org/10.1016/j.trd.2024.104538
Received 28 June 2024; Received in revised form 8 November 2024; Accepted 27 November 2024

traffic crashes. Extensive research has explored various dimensions of pedestrian safety, including crash severity, crash frequency, and spatial crash patterns. One of the primary causes of pedestrian crashes is driver behavior, such as speeding, which has been repeatedly highlighted in numerous studies (Habibovic et al., 2013; Kemnitzer et al., 2019; Sheykhfard et al., 2021). Additional research has also found that factors like pedestrian age, crossing behavior, and alcohol consumption significantly contribute to crash risks (Henary et al., 2006; Hezaveh and Cherry, 2018; LaScala et al., 2001; Lasota et al., 2019; Niebuhr et al., 2016; Prange et al., 2010). Road conditions, such as poorly lit intersections and high-speed zones, further increase the danger for pedestrians, alongside temporal factors like time of day and day of the week. For instance, Toran Pour et al. (2018) observed a higher occurrence of male pedestrian crashes during late-night weekend hours, while crashes involving older pedestrians were more frequent during weekday off-peak times.

In pedestrian safety studies, a variety of analytical techniques have been employed to understand crash patterns and improve safety. Geospatial analysis, often used to map and analyze pedestrian crash hotspots, helps identify high-risk areas for targeted interventions (Hussain et al., 2023; Truong and Somenahalli, 2011). In addition, machine learning models such as decision trees and random forests have gained popularity for their ability to identify complex relationships between pedestrian demographics, traffic conditions, and crash risk factors (Li et al., 2017; Sivasankaran and Balasubramanian, 2021). Among these, pedestrian clustering pattern analysis, though relatively less explored, is crucial for identifying hidden patterns and trends in crash data that are not immediately apparent. By grouping similar crashes, clustering helps to reveal underlying factors that contribute to crashes, offering unique insights that can drive targeted safety interventions (Depaire et al., 2008). In the meantime, equity is an increasing concern in pedestrian safety, with several studies examining this issue (Braun et al., 2023; Chimba et al., 2018; Gu and Peng, 2021, 2021; Haddad et al., 2023; Wang et al., 2016).

Therefore, this research aims to enhance pedestrian safety by integrating rich data on environmental justice into the analysis and employing SHAP-based clustering to refine our understanding of the influences driving crash severities. This study utilizes comprehensive crash data from California from 2018 to 2021, which was gathered from the Highway Safety Information System (HSIS). It also incorporates environmental justice features sourced from the Equitable Transportation Community, providing a rich dataset that includes extensive socioeconomic, demographic, and build-environmental data at the Census Tract level. This rich dataset supports an in-depth analysis of transportation equity metrics and their influence on pedestrian crash severities. The dataset encompasses four tiers of information: unit level, crash level, roadway details, and zonal data, from which 77 variables were chosen to develop models predicting pedestrian crash severity across three categories: fatal, injury, and no injury. Advanced machine learning techniques, including CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting, were employed. The study utilizes the Tree SHAP algorithm to examine the influence of each variable on the predictions. Hierarchical clustering of SHAP values then reveals distinct patterns, enhancing our understanding of how various factors impact each severity category.

2. Literature review

Many studies have explored various aspects of pedestrian safety, including pedestrian crash severity analysis (Haleem et al., 2015; Pour-Rouholamin and Zhou, 2016), pedestrian crash frequency analysis (Chen and Zhou, 2016; Ding et al., 2018), and pedestrian clustering pattern analysis (Das et al., 2019; Kim and Yamashita, 2007; Sasidharan et al., 2015; Sun et al., 2019). Among these, pedestrian clustering pattern analysis, though relatively less explored, is crucial for identifying hidden patterns and trends in crash data that are not immediately apparent. By grouping similar crashes, clustering helps to reveal underlying factors that contribute to crashes, offering unique insights that can drive targeted safety interventions (Depaire et al., 2008). In traffic crash clustering pattern analysis, data are analyzed to discern groups of similar crash crashes based on various factors such as location (Anderson, 2009; Bíl et al., 2013; Kim and Yamashita, 2007), time (Kumar and Toshniwal, 2016), environmental conditions (Huang et al., 2023; Mehdizadeh et al., 2020), involved demographics (Fan et al., 2018), and crash characteristics (Kaplan and Prato, 2013; Weiss et al., 2016). Association rules mining (Das, 2021; Das et al., 2019) and correspondence analysis (Baireddy et al., 2018; Das and Sun, 2015), along with various clustering methods such as K-means (Anderson, 2009; Kim and Yamashita, 2007), hierarchical clustering (Song et al., 2020; Taamneh et al., 2017), latent class analysis (Behnood et al., 2014; Sasidharan et al., 2015; Sun et al., 2019) and density-based clustering (Xie and Yan, 2013; Zhang et al., 2018), are widely adopted to analyze patterns in pedestrian crash data.

Across various dimensions of pedestrian crash analysis-whether severity, frequency, or clustering-equity has emerged as a critical concern (Braun et al., 2023; Chimba et al., 2018; Gu and Peng, 2021, 2021; Haddad et al., 2023; Wang et al., 2016). Younes et al. (2023) analyzed factors affecting fatal pedestrian and bicyclist crashes in New Jersey using data at individual level and considered CBG level population density and income characteristics, finding that crashes occur disproportionately in low-income areas. Key factors influencing crash fatality include light conditions, non-motorist age, posted speed, and vehicle type. Haddad et al. (2023) investigated four categories of external factors influencing residential neighborhoods and their potential impact on pedestrian crashes and racial disparities. Their findings highlight that social resistance, transit stop density, and road design are the most significant determinants of pedestrian crashes, especially in CBGs where the majority of residents are Black. Li et al. (2022) evaluated transportation safety inequities by analyzing the spatial relationship between crash risks and the Social Vulnerability Index (SVI) developed by the Centers for Disease Control and Prevention. Their findings reveal a significant correlation between crash rates and social vulnerability, particularly in highly urbanized areas.

In recent years, advancements in study approaches have highlighted the promise of machine learning in many fields of analysis, accompanied by a growing focus on explainable AI to address the opacity of these models (Atakishiyev et al., 2021; Karim et al., 2022). The incorporation of explainable AI ensures that these models are not just black boxes; instead, it makes their decision-making processes accessible and interpretable. For example, Kong et al. (2023) adopted an interpretable machine learning framework using SHapley Additive exPlanations (SHAP) to understand the factors associated with critical pedestrian-involved near-crash events. Chang

et al. (2022) used XGBoost and SHAP to explore the effects of the built environment on fatal pedestrian crashes at location-specific level. Furthermore, Yue (2024) investigated the association between streetscape characteristics and pedestrian crashes at intersections. By incorporating the SHAP method with XGBoost, an interpretable framework was developed to examine the relationship between pedestrian crash occurrence and the surrounding streetscape environment. The findings are analyzed from global, local, and regional perspectives.

While many studies have used SHAP to explain crash contributing factors, few have explored clustering patterns based on SHAP values to reveal additional insights in pedestrian safety analysis. Although this method has been applied in other fields, its use in pedestrian safety remains relatively underexplored. For instance, Cilínio et al. (2023) used a supervised clustering approach with SHAP values to investigate fault detection systems in mobile networks. Similarly, Escriva et al. (2023) quantified attribute contributions in binary classification using explainable AI local attribution methods and aggregated local explanations for deeper insights. Their findings suggest that clusters based on SHAP perform better than those based on raw data, even with low accuracy models. Lu et al. (2022) found that SHAP analysis revealed patterns not easily discernible through original features identified by XGBoost in heart failure patients' EHRs. To our knowledge, no studies in pedestrian safety have adopted this approach in pedestrian safety analysis.

2.1. Research gaps and contribution

Despite the extensive research on pedestrian safety, two significant gaps remain. First, traditional clustering methods in pedestrian crash analysis generally rely on raw data and arithmetic means, making them sensitive to outliers. These methods also often assume linear relationships between data points, which limits their ability to capture the complex, non-linear interactions that may exist between features. While these techniques provide useful insights into data groupings based on spatial or simple metric similarities, they do not inherently explain why certain crashes are grouped together from the perspective of a predictive model. Although SHAP has been widely used to explain contributing factors in machine learning models, few studies have applied clustering based on SHAP values to pedestrian safety, leaving a gap in understanding how feature contributions can reveal deeper insights into crash patterns.

Second, there is a lack of research that integrates transportation equity into pedestrian safety analysis. Few studies consider how transportation disparities affect pedestrian crash outcomes, particularly in low-income or minority communities. The absence of an equity-focused perspective limits our understanding of how social vulnerabilities intersect with pedestrian safety. Comprehensive equitable transportation datasets, which can provide a more complete picture of these disparities, are not intensively applied to pedestrian safety research, leaving critical questions about the relationship between equity and safety underexplored.

This study aims to address these gaps by using SHAP values for hierarchical clustering in pedestrian safety analysis. Clustering based on SHAP values shifts the focus from grouping crashes by raw feature values to clustering based on the contribution of each feature to model predictions. This allows for a more detailed analysis of the factors driving pedestrian crashes, revealing patterns that

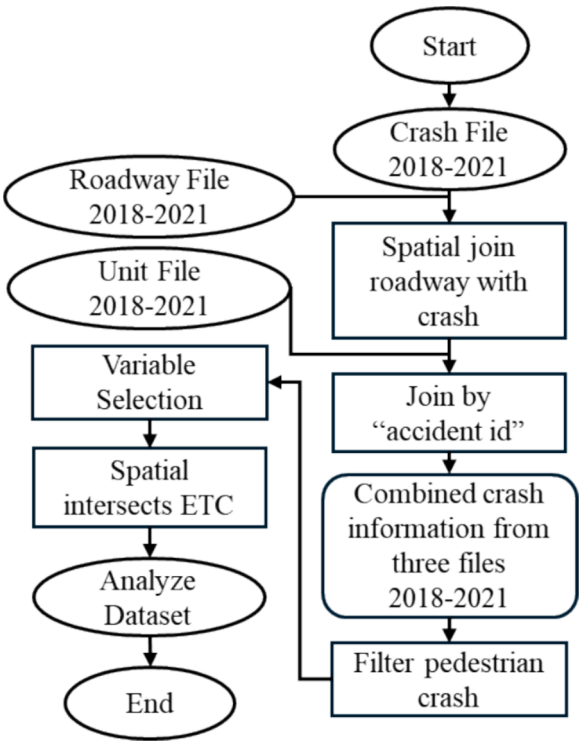


Fig. 1. Data Preparation Process.

may not be apparent from raw data alone. Furthermore, because SHAP values come from tree-based models, they are well-suited to capturing non-linear interactions between features, making the resulting clusters more reflective of the complex relationships that exist in real-world data. Calculating SHAP values acts as a preprocessing step that transforms raw data into a unified scale based on the output of the supervised prediction model. In addition, this study incorporates equitable transportation data to examine the relationship between transportation equity and pedestrian safety outcomes. By including this data, the analysis extends beyond traditional crash risk factors to explore how transportation inequities disproportionately impact vulnerable populations. This provides a more comprehensive view of pedestrian safety and helps to identify disparities in crash risks that affect underserved communities.

3. Data preparation

This study utilizes crash data collected from the HSIS, a comprehensive multistate database comprising detailed crash data, roadway inventories, and traffic volume information across selected states (United States Department of Transportation, 2024). This study covers four years, from 2018 to 2021, and focuses on California. It includes data on crash details, unit-level crash information, and roadway inventory shapefiles. Fig. 1 illustrates the data preparation process. Data processing was methodically carried out on a yearly basis. Taking 2018 as an illustrative example, the unit-level data for that year was merged with the corresponding crash data utilizing the crash ID as a key identifier. Subsequently, a spatial join was performed between the 2018 roadway file and the crash data, thus combining information from three distinct sources into a consolidated point-format file. The next step involved extracting pedestrian-involved crash instances, specifically where the 'party_group' was marked 'Yes'. Additionally, variable selection was conducted based on the data quality and relevance to the study's aims. To gain a macro-level perspective on crash locations, the cleaned pedestrian crash points were spatial intersected with the Equitable Transportation Community dataset (US Department of Transportation, 2024). This result dataset provides a detailed focus on transportation equity metrics, along with extensive socio-economic and demographic data at the Census Tract level, enabling a broader understanding of the factors contributing to pedestrian crash crashes.

In total, four levels of information were considered: unit level, crash level, roadway information, and zonal information. From these, 77 variables were selected to develop models predicting the severity of pedestrian crashes. Each record represented a unique individual-level crash, capturing the distinct characteristics and severity outcomes for each person involved in a crash. This individual-level approach acknowledges the complexity of crash events, which often involve multiple vehicles and pedestrians, each affected to varying extents. The final cleaned dataset comprises 5,446 instances of pedestrian-involved crashes. The original crash severity is categorized using the KABCO system, which classifies crashes into K (Killed), A (Incapacitating Injury), B (Non-Incapacitating Injury), C (Possible Injury), and O (Property Damage Only). For simplification purposes, pedestrian injury severity is categorized into three distinct classes: fatal, injury, and no injury. This is a common approach used in other pedestrian safety research (Rahim and Hassan, 2021; Khan et al., 2024). The distribution of these categories is as follows: injury accounts for 3,840 cases (71 %), fatal for 1,227 cases (23 %), and no injury for 379 cases (7 %). This indicates a significant data imbalance. To address this, SMOTE was applied to rebalance the dataset, resulting in 3,840 cases for each class. Previous pedestrian studies have examined variables associated with pedestrians (Baireddy et al., 2018; Mafi et al., 2018; Sasidharan et al., 2015; Sun et al., 2019), crash characteristics (Das et al., 2019; Hossain et al., 2022), road characteristics (Batouli et al., 2020; Mafi et al., 2018; Sasidharan et al., 2015), socio-economic and build-environmental factors (Haddad et al., 2023; Rahman et al., 2019). This study covers all four aspects. However, pedestrian demographic information such as age and gender are not included due to the lack of available data. Table 1 illustrates the descriptive statistics of the variables used in the models. Pedestrian sobriety impairment includes ten categories, with approximately 6 % of pedestrians involved found to be impaired by alcohol. Pedestrian actions include roadway, including shoulder (47 %), crossing at intersection crosswalks (26 %), crossing not at crosswalks (20 %), not in roadway (4 %), crossing at non-intersection crosswalks (2 %). Location types include highways (6 %), intersections (22 %), and ramps (18 %). Environmental justice variables include transportation insecurity, environmental burden, health vulnerability, social vulnerability, and climate and disaster risk burden (Table 2).

4. Methodology

The study begins with a prepared pedestrian crash dataset that includes data on 5,446 crashes, each characterized by 77 different attributes. Pedestrian injury severity is classified into three categories: 1,227 fatal cases, 3,840 injury cases, and 379 no-injury cases. Given the imbalance in sample sizes across these severity levels, Synthetic Minority Over-sampling Technique (SMOTE) is applied, creating a balanced dataset with 3,840 samples for each severity level. This approach strengthens the robustness of the analysis by ensuring equitable representation across categories. The next phase involves hyperparameter tuning and model selection for various tree-based machine learning models. The models considered in this study include CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. Through hyperparameter tuning, the study identifies the optimal settings for each model to ensure the best possible performance.

SHAP (SHapley Additive exPlanations) is a game-theory-based approach used to explain the output of machine learning models. It assigns each feature in the model a specific contribution to the prediction, helping to interpret how the model's inputs influence its decisions. SHAP values are calculated based on the Shapley value concept from cooperative game theory, ensuring that the feature attributions are fair and additive. Using the tuned CatBoost model, the study employs the Tree SHAP algorithm to analyze the influence of each variable on the model's predictions using the original dataset as input. This results in a SHAP value table for each severity level – fatal, injury, and no injury – with each table containing SHAP values for the corresponding 5,446 instances across 77 attributes. The study then proceeds to a global explanation phase, where it examines the overall importance of variables across different injury

Table 1
Descriptive Statistics of the Categorical Variables.

Variables	Fatal (1227)	Injury (3840)	NoInjury (379)	Variables	Fatal (1227)	Injury (3840)	NoInjury (379)
Pedestrian action (Ped_actn)				***Lighting Condition (Lghtng_cndtn)***			
Int. Xwalk	73 (5.95 %)	1253 (32.6 %)	95 (25.1 %)	Daylight	114 (9.29 %)	1618 (42.1 %)	158 (41.7 %)
Non-Int. Xwalk	10 (0.81 %)	90 (2.34 %)	1 (0.26 %)	Dusk/Dawn	41 (3.34 %)	120 (3.12 %)	17 (4.49 %)
Non-Xwalk	316 (25.8 %)	711 (18.5 %)	64 (16.9 %)	Lit Dark	503 (41.0 %)	1318 (34.3 %)	106 (28.0 %)
Off Road	28 (2.28 %)	198 (5.13 %)	16 (4.22 %)	Unlit Dark	569 (46.3 %)	773 (20.1 %)	98 (25.8 %)
Road w/Shoulder	800 (65.2 %)	1588 (41.4 %)	203 (53.6 %)	Not Stated	0 (0.00 %)	11 (0.29 %)	0 (0.00 %)
Party Sobriety (Sobriety)				***Location Type (Lctn_type)***			
Drugged	64 (5.22 %)	17 (0.44 %)	4 (1.06 %)	Highway	1028 (83.8 %)	2025 (52.7 %)	241 (63.6 %)
HBD Sober	14 (1.14 %)	38 (0.99 %)	0 (0.00 %)	Intersection	102 (8.31 %)	1007 (26.2 %)	78 (20.6 %)
HBD Unknown	61 (4.97 %)	139 (3.62 %)	3 (0.79 %)	Ramp	97 (7.91 %)	808 (21.0 %)	60 (15.8 %)
Influenced	139 (11.3 %)	166 (4.32 %)	12 (3.17 %)	***Roadway condition (Road_cndtn)***			
Other Impair	0 (0.00 %)	4 (0.10 %)	0 (0.00 %)	Construction Zone	42 (3.42 %)	93 (2.42 %)	17 (4.49 %)
Sober	145 (11.8 %)	2472 (64.4 %)	165 (43.5 %)	Flooded	2 (0.16 %)	3 (0.08 %)	2 (0.53 %)
Unknown	804 (65.5 %)	1003 (26.1 %)	195 (51.4 %)	Holes, Ruts	0 (0.00 %)	13 (0.34 %)	0 (0.00 %)
Party drug or physical impairment (Drug_physcl_imprmnt)				Loose Material	0 (0.00 %)	3 (0.08 %)	2 (0.53 %)
Driver Fatigue		0 (0.00 %)	2 (0.05 %)	0 (0.00 %)	0 (0.00 %)	9 (0.23 %)	1 (0.26 %)
Drugged	95 (7.74 %)	21 (0.55 %)	3 (0.79 %)	Normal	1153 (94.0 %)	3633 (94.6 %)	346 (91.3 %)
Physical Impair	0 (0.00 %)	2 (0.05 %)	1 (0.26 %)	Not Stated	1 (0.08 %)	20 (0.52 %)	1 (0.26 %)
Unstated	1132 (92.3 %)	3815 (99.3 %)	375 (98.9 %)	Obstruction	24 (1.96 %)	43 (1.12 %)	7 (1.85 %)
Primary Collision Factor (Prmry_cllsn_fctr)				Other	5 (0.41 %)	23 (0.60 %)	3 (0.79 %)
Alcohol Influence		49 (3.99 %)	174 (4.53 %)	13 (3.43 %)			
Improper Driving	1 (0.08 %)	5 (0.13 %)	2 (0.53 %)	Dry	1131 (92.2 %)	3430 (89.3 %)	359 (94.7 %)
Improper Turn	49 (3.99 %)	397 (10.3 %)	18 (4.75 %)	Not Stated	0 (0.00 %)	18 (0.47 %)	0 (0.00 %)
Non-Driver	53 (4.32 %)	89 (2.32 %)	23 (6.07 %)	Slippery	1 (0.08 %)	1 (0.03 %)	0 (0.00 %)
Other Violations	308 (25.1 %)	971 (25.3 %)	90 (23.7 %)	Snow, Icy	4 (0.33 %)	47 (1.22 %)	1 (0.26 %)
Speeding	119 (9.70 %)	510 (13.3 %)	39 (10.3 %)	Wet	91 (7.42 %)	344 (8.96 %)	19 (5.01 %)
Tailgating	0 (0.00 %)	1 (0.03 %)	1 (0.26 %)	***Traffic Operation (Traffic_operation)***			
Yield Failure	632 (51.5 %)	1603 (41.7 %)	183 (48.3 %)	Control Off	3 (0.24 %)	6 (0.16 %)	0 (0.00 %)
Unknown	16 (1.30 %)	90 (2.34 %)	10 (2.64 %)	Control On	172 (14.0 %)	1536 (40.0 %)	123 (32.5 %)
Crash Type (Crash_type)				Controls Obscured	1 (0.08 %)	3 (0.08 %)	0 (0.00 %)
Auto-Pedestrian	1089 (88.8 %)	2980 (77.6 %)	220 (58.0 %)	No Controls	1050 (85.6 %)	2289 (59.6 %)	256 (67.5 %)
Broadside	30 (2.44 %)	139 (3.62 %)	11 (2.90 %)	Not Stated	1 (0.08 %)	6 (0.16 %)	0 (0.00 %)
Head-On	4 (0.33 %)	78 (2.03 %)	3 (0.79 %)	***Weather Condition (Weather_condition)***			
Hit-Object	13 (1.06 %)	88 (2.29 %)	42 (11.1 %)	Clear	991 (80.8 %)	3210 (83.6 %)	326 (86.0 %)
Not-Stated	2 (0.16 %)	21 (0.55 %)	2 (0.53 %)	Cloudy	181 (14.8 %)	420 (10.9 %)	40 (10.6 %)
Other	10 (0.81 %)	23 (0.60 %)	19 (5.01 %)	Fog	13 (1.06 %)	13 (0.34 %)	2 (0.53 %)
Overturn	0 (0.00 %)	9 (0.23 %)	3 (0.79 %)	Not Stated	0 (0.00 %)	6 (0.16 %)	1 (0.26 %)
Rear-End	49 (3.99 %)	331 (8.62 %)	48 (12.7 %)	Other	3 (0.24 %)	3 (0.08 %)	0 (0.00 %)
Sideswipe	30 (2.44 %)	171 (4.45 %)	31 (8.18 %)	Raining	36 (2.93 %)	165 (4.30 %)	9 (2.37 %)
Number of Vehicles (Num_vhcls)				Snowing	1 (0.08 %)	21 (0.55 %)	1 (0.26 %)

(*continued on next page*)

Table 1 (*continued*)

Variables	Fatal (1227)	Injury (3840)	NoInjury (379)	Variables	Fatal (1227)	Injury (3840)	NoInjury (379)
1 Veh	929 (75.7 %)	2949 (76.8 %)	274 (72.3 %)	Wind	2 (0.16 %)	2 (0.05 %)	0 (0.00 %)
2–4 Vehs	284 (23.1 %)	873 (22.7 %)	103 (27.2 %)	***Access Control (Access_control)***			
5–7 Vehs	14 (1.14 %)	18 (0.47 %)	2 (0.53 %)	Conventional	371 (30.2 %)	1819 (47.4 %)	144 (38.0 %)
Median Variance (Median_variance)				Expressway	41 (3.34 %)	103 (2.68 %)	15 (3.96 %)
No Variance	1002 (81.7 %)	3134 (81.6 %)	310 (81.8 %)	Freeway	812 (66.2 %)	1888 (49.2 %)	217 (57.3 %)
Variable	193 (15.7 %)	615 (16.0 %)	58 (15.3 %)	One-Way Street	3 (0.24 %)	30 (0.78 %)	3 (0.79 %)
Wide No Var.	32 (2.61 %)	91 (2.37 %)	11 (2.90 %)	***Median Type (Median_type)***			
Number of Lanes (No_lanes)				2-Way Left Turn	82 (6.68 %)	377 (9.82 %)	33 (8.71 %)
1 Lane	0 (0.00 %)	1 (0.03 %)	0 (0.00 %)	Bus Lanes	0 (0.00 %)	1 (0.03 %)	0 (0.00 %)
10 or more	234 (19.1 %)	613 (16.0 %)	65 (17.2 %)	Cont. Left Turn	76 (6.19 %)	463 (12.1 %)	26 (6.86 %)
2–4 Lanes	504 (41.1 %)	1751 (45.6 %)	183 (48.3 %)	Grades Wall	1 (0.08 %)	3 (0.08 %)	0 (0.00 %)
6–8 Lanes	489 (39.9 %)	1475 (38.4 %)	131 (34.6 %)	Occ. Lane	2 (0.16 %)	13 (0.34 %)	2 (0.53 %)
Land Use (Land_use)				Other	10 (0.81 %)	68 (1.77 %)	7 (1.85 %)
Rural	185 (15.1 %)	445 (11.6 %)	57 (15.0 %)	Paved Median	503 (41.0 %)	1393 (36.3 %)	144 (38.0 %)
Urban	94 (7.66 %)	320 (8.33 %)	41 (10.8 %)	Rail & Bus	0 (0.00 %)	5 (0.13 %)	0 (0.00 %)
Urbanized	948 (77.3 %)	3075 (80.1 %)	281 (74.1 %)	Railroad	15 (1.22 %)	55 (1.43 %)	9 (2.37 %)
Terrain (Terrain)				Rev. Lane	1 (0.08 %)	3 (0.08 %)	0 (0.00 %)
Flat	901 (73.4 %)	2721 (70.9 %)	261 (68.9 %)	Sawtooth	1 (0.08 %)	4 (0.10 %)	1 (0.26 %)
Mountainous	46 (3.75 %)	204 (5.31 %)	23 (6.07 %)	Separate Grades	25 (2.04 %)	52 (1.35 %)	5 (1.32 %)
Rolling	280 (22.8 %)	915 (23.8 %)	95 (25.1 %)	Separate Structure	28 (2.28 %)	64 (1.67 %)	6 (1.58 %)
Design Speed (Design_speed)				Striped	138 (11.2 %)	587 (15.3 %)	54 (14.2 %)
25–35	40 (3.26 %)	300 (7.81 %)	27 (7.12 %)	Unpaved	345 (28.1 %)	752 (19.6 %)	92 (24.3 %)
40–55	183 (14.9 %)	1012 (26.4 %)	75 (19.8 %)	***Highway Group (Highway_group)***			
60–70	1004 (81.8 %)	2528 (65.8 %)	277 (73.1 %)	Divided Hwy	10 (0.81 %)	45 (1.17 %)	5 (1.32 %)
Roadway Classification (Roadway_clssfctn)				Ind. Right Align	1079 (87.9 %)	3208 (83.5 %)	320 (84.4 %)
2 Lane Roads		112 (9.13 %)	459 (12.0 %)	53 (14.0 %)	138 (11.2 %)	587 (15.3 %)	54 (14.2 %)
Freeways	853 (69.5 %)	1991 (51.8 %)	232 (61.2 %)	***Divided Roadway (Divided_roadway)***			
Multilane Roads	262 (21.4 %)	1390 (36.2 %)	94 (24.8 %)	No	138 (11.2 %)	587 (15.3 %)	54 (14.2 %)
				Yes	1089 (88.8 %)	3253 (84.7 %)	325 (85.8 %)

severity levels, providing insights into which factors are most influential in fatal, injury, and no-injury crash outcomes.

Next, the study selects SHAP values by class labels, narrowing down the analysis to instances within each severity category – 1227 fatal, 3840 injuries, and 379 no injuries – each detailed with the top 20 important attributes. In the subsequent phase, hierarchical clustering is applied to the selected SHAP values for each severity category. This analysis reveals distinct clustering patterns, identifying three clusters within fatal incidents, and two clusters each within injury and no-injury incidents. These patterns indicate the presence of subgroups within each injury severity category, characterized by their unique variable importance profiles as determined by the Tree SHAP algorithm. Fig. 2 presents the study flowchart.

Table 2
Descriptive Statistics of the Continuous Variables.

Variables	Variable code	Mean	STD	Min	Max
Roadway characteristics (Segment level information)					
Annual average daily traffic	AADT	102307.38	91884.36	260	719,463
Average lane width	Lane Width	11.86	1.46	0	25
Left shoulder width (outside)	Lft shldr wdth (otsd)	7.87	3.38	0	33
Lef shoulder width (inside)	Lft shldr wdth (insd)	4.04	4.78	0	35
Median width	Median Width	24.91	24.47	0	99
Right shoulder width (outside)	Rght shldr width (otsd)	4.13	4.77	0	35
Right shoulder width (inside)	Rght shldr width (insd)	7.96	3.22	0	24
Travel way width (left/decreasing)	Surface Width	35.79	17.69	0	108
Environmental justice variables (Census Tract level information)					
Total Population	Total_pop	4439.32	2505.46	0	39,373
Percent of households with no car	%_no_car	24.95	247.35	0	4902.50
Average commute time to work	Avg_cmmt_time	28.72	7.40	1.2	67.43
Frequency of Transit Services per Sq Mi	Transit_frequency	34.05	109.68	0	1663.19
Jobs within a 45-min Drive	Jobs_45m_drive	119708.03	145988.20	6.90	828440.62
Estimated Average Drive Time to Points of Interest (min)	Drive Time to POI	63.80	1612.45	1.407	53926.23
Estimated Average Walk Time to Points of Interest (min)	Walk Time to POI	119.45	168.41	4.26	1255.53
Calculated average annual cost of Transportation as percent of household income	% cost of Trnsprttn	1.27	15.22	0.04	378.99
Traffic Fatalities per 100,000 people	Traffic Ftlts	10.62	30.11	0	370.98
Ozone level in the air	Ozone level	47.73	11.97	26.34	74.40
Particulate Matter 2.5 (PM2.5) level in the air	PM2.5 level	11.47	2.30	5.676	17.75
Diesel particulate matter level in air	Diesel particulate	0.32	0.19	0	1.42
Air toxics cancer risk	Air txcs cncr rsk	30.16	9.73	0	200
Percent of tract within 1 mile of known hazardous sites	% to hzrds sites	73.29	38.29	0	100
Percent of tract within 1 mile of known Toxics Release sites	% to txcs rls sites	45.58	42.55	0	100
Percent of tract within 1 mile of known Treatment and Disposal Facilities	% to trtmnt fclts	4.21	16.93	0	100
Percent of tract within 1 mile of known Risk Management Plan Sites	% to risk mngmnt sites	23.39	33.97	0	100
Percent of houses built before 1980	% of houses blt bfr 1980	60.55	25.01	0	100
Percent of tract within 1 mile of high volume roads	hghvlr	77.29	34.07	0	100
Percent of tract within 1 mile of railways	%_to_1mi_of_rlwys'	45.49	42.36	0	100
Percent of tract within 5 miles of airports	% wthn 5 mi of arprts	77.09	37.15	0	100
Percent of tract within 3 miles of ports	prts	4.09	19.09	0	100
Percent of tract that intersects with a Watershed containing impaired water(s)	% imprd water	73.54	39.76	0	100
Asthma prevalence	asthm	6.81	3.27	0	14.9
Cancer prevalence	Cncr prvlnc	3.58	2.04	0	13
High blood pressure prevalence	bldprs	19.15	9.51	0	49.60
Diabetes prevalence	dbts	7.12	3.96	0	23.39
Poor mental health prevalence	mntlhl	10.67	5.38	0	22.80
Percent of population with Income below 200 % of poverty level	% of population below poverty	32.97	18.25	0	100
Percent of people age 25 + with less than a high school diploma	% hgh schl dplm	18.24	14.23	0	73.8
Percent of people age 16 + unemployed	% unemployed	4.28	2.69	0	19.6
Percent of total housing units that are renter-occupied	% rntr occupd	48.33	23.39	0	100
Percent of occupied houses spend 30 % or more on housing with less than 75 k	% 30 % or more on housing	41.03	11.44	0	100
Percent of population uninsured	% uninsured	7.92	5.67	0	50.74
Percent of households with no internet subscription	% no internet	13.37	9.47	0	100
GINI Index	GINI Index	0.49	0.82	0.1302	22.24
Percent of population 65 years or older	% 65 or older	14.83	8.41	0	100
Percent of population 17 years or younger	% 17 or younger	22.17	7.39	0	55.02
Percent of population with a disability	% disability	11.96	5.93	0	100
Percent of population (age 5+) with limited English proficiency	% lmtd English	9.10	8.84	0	100
Percent of total housing units that are mobile homes	% mobile homes	6.23	11.89	0	100
Estimated annualized loss due to disasters	Annual dsstr loss	2946543.92	4795553.42	4014.13	55002605.85
Increase in number of days over 90 deg by mid-century	Incrs days over 90 deg	16.80	13.25	0.01	53.41
Number of days exceeding 99th percentile of precip by mid-century	extrmp	4.22	2.44	0.01	17.33
Percent change in number of days with less than 0.01 in. of precip	drghtd	3.38	2.23	−2.09	12.55
Percent of tract inundated by 0.5 sea level increase by 2100	pctnnd	0.11	0.58	0	8.44
Average Percent Land classified as Impervious Surface per Tract	% imprvs surface	44.13	27.36	0.08	97.30

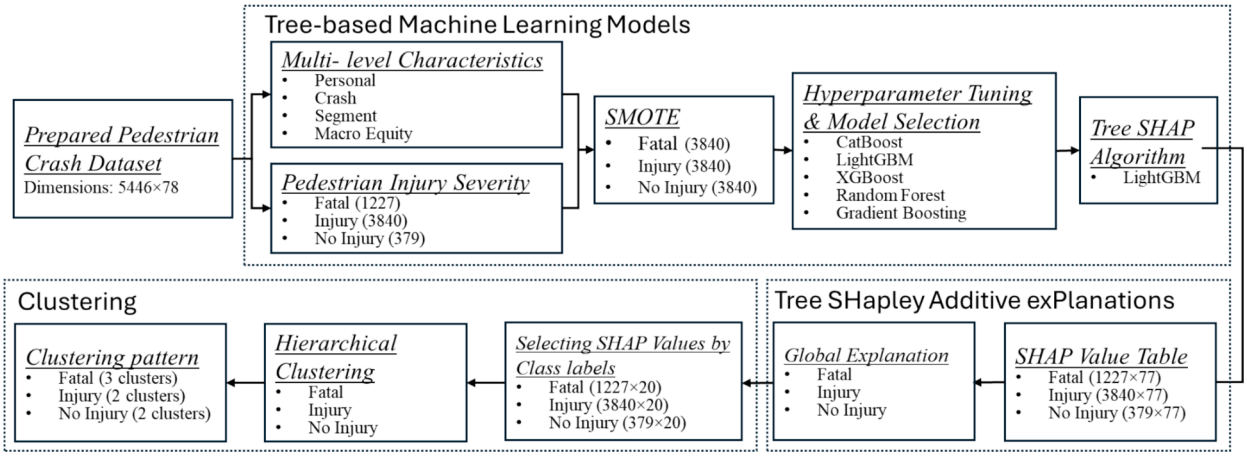


Fig. 2. Study Flowchart.

4.1. Explainable AI

Explainable AI methods analyze feature attributions to determine each feature's contribution to a machine learning model's predictions. SHAP allows users to explain the predictions of sophisticated machine learning models using input variables (Linardatos et al., 2020). In this experiment, SHAP values provide a measure of each feature's importance in the model. The goal of this method is to clarify pixel anomalies by evaluating each feature's contribution to severity. To explain a prediction for a single case x using tree-based model $f(x)$. $f(x)$ can be expressed as a sum of the importance of individual pixel features ϕ_i. The model's prediction can be described as follows:

$$f(x) = g(x') = \phi_0 + \phi_i x_i' \tag{1}$$

Here, ϕ_0 is what the prediction would be if this case had no features (a baseline), and x' is a vector indicating which features are included in the severity prediction. The function g represents an additive explanation model, where the overall prediction is broken down by the contribution of each feature.

SHAP is a method for such additive feature attribution proposed by Lundberg (Lundberg and Lee, 2017), inspired by the Shapley value from cooperative game theory (Shapley, 1997), which measures the contribution each player brings to the game. The Shapley value $\phi_i(v)$ is defined as:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{2}$$

This equation considers all subsets S of features that do not include feature i, where n is the total number of features. It computes the average marginal contribution of feature i, across all possible combinations of features. In this context, assumed of this like a game where each feature value is a player, and the payout is the prediction made by the model f. For an input instance x and a feature set S, the Shapley value $\phi_i(f, x)$ quantifies the contribution of feature i to the prediction $f(x)$ and is calculated as:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(n - |z'| - 1)!}{n!} (f(z') - f(z' \setminus \{i\})) \tag{3}$$

where z' is a binary vector representing a subset of features. This formula averages the impact of including feature i in the model across all subsets of features. By utilizing this method, this research seeks to explain severity data by computing the contribution of each severity label prediction. In this study, the SHAP value was calculated using Lundberg's (2019) explainer approach, which provides a speedy and accurate feature attribution method by employing an ensemble-based decision tree structure. The best machine learning model used in this approach is compatible with SHAP Explainer because it makes predictions using an ensemble-based decision tree. In this study, the *shap.TreeExplainer* Python package is utilized to perform SHAP analysis.

4.2. Machine learning models

Machine learning's prediction function has a variety of customized models, including random forest algorithms (Random Forest), extreme random tree algorithms (Extremely Randomized Trees), two boosting tree algorithms (Catboost and Light GBM), and ensemble learning approaches. The random forest and extremely randomized tree algorithms select the best variables using Entropy or Gini coefficients (Breiman, 2001). Entropy, a concept adopted from information theory, measures the uncertainty inherent in a collection of random variables. Similarly, the Gini coefficient is a measure of data impurity. Similar to information entropy, it is a

useful tool for feature selection, assisting in the identification of the most discriminative variables (Geurts et al., 2006; Kullback and Leibler, 1951).

The boosting tree technique is a machine learning method that uses decision trees and is optimized with the forward stepwise algorithm and linear tree combination (James et al., 2023). The gradient boosting decision tree strategy is often used to optimize the complexity of boosting tree models (Chen and Guestrin, 2016). Boosting is a mechanism for turning weak learners into strong learners. Tree based boosting methods include Catboost and Light GBM, which are based on gradient boosting decision trees and can handle massive amounts of data (Prokhorenkova et al., 2018). Furthermore, these algorithms are well-known for their resistance to overfitting and ability to handle categorical features directly, making them a popular choice for a variety of machine learning tasks (Erickson et al., 2020; Ke et al., 2017; Prokhorenkova et al., 2018).

4.3. Hierarchical clustering

To identify typical clustering patterns, this study conducted a hierarchical clustering analysis on the SHAP values for each severity level. Hierarchical clustering is based on a dendrogram, or cluster tree, which organizes data according to the similarity between individual data points. In hierarchical clustering, data points are grouped by merging similar clusters until all groups form one large cluster, a process known as the agglomerative strategy (Kaufman and Rousseeuw, 1990). Conversely, the divisive strategy starts with a large cluster that continuously splits into smaller clusters based on their heterogeneity. This study used the agglomerative strategy. To calculate the average inter-cluster distance using the average-linkage approach, the distance between each pair of observations is measured, summed, and then divided by the number of pairs in the cluster. The distance is estimated as follows:

$$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{([u]^*[v])}, \tag{4}$$

where, i and j are the data instances that correspond to each cluster, and u and v are the two clusters, respectively. Rousseeuw's Silhouette approach (1987) was used to determine the number of clusters from the resulting clustered tree. The silhouette coefficient was used in this method to assess the cohesiveness measure and the distance between the generated clusters. While cluster separation quantifies the degree to which each cluster is isolated from the others, cohesion quantifies how closely the data inside a cluster are related to one another. Equation (5) details the silhouette coefficient:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{5}$$

where, $s(i)$ is the silhouette coefficient of sample i, which is determined by comparing sample $a(i)$, which represents sample i as average distance to other samples in the same cluster, to sample $b(i)$, which represents sample i as average distance to all other clusters. The silhouette coefficient ranges from -1 to 1. A more positive value indicates a higher probability that a data point is correctly assigned to its cluster.

4.4. Evaluation metrics

To evaluate the performance of classification models, weighted metrics are selected considering the imbalanced nature of the injury severity. Here's an overview of accuracy, weighted precision, weighted recall, and weighted F1 score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Weighted\,Precision = \frac{\sum_i \left(\left(\frac{TP}{TP+FP} \right)_i \times Number\ of\ instances_i \right)}{Total\ Number\ of\ Instances}$$

$$Weighted\,Recall = \frac{\sum_i \left(\left(\frac{TP}{TP+FN} \right)_i \times Number\ of\ instances_i \right)}{Total\ Number\ of\ Instances}$$

$$Weighted\,F1\ Score = \frac{\sum_i \left(\left(\frac{Precision \times Recall}{Precision + Recall} \right)_i \times Number\ of\ instances_i \right)}{Total\ Number\ of\ Instances}$$

where TP is correctly predicted positive class, TN is correctly predicted negative class, FP is incorrectly predicted positive class, FN is incorrectly predicted negative class.

5. Results and discussion

5.1. Hyperparameter tuning and model selection

The first step of the pedestrian crash severity analysis involved fine-tuning the parameters and selecting the machine learning models. These models include CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. The dataset was split into two subsets, 70 % for training and the remaining 30 % for testing. Stratified sampling was applied to maintain balanced class distributions across both sets. During the training phase, the models were validated at every three iterations. The aim of the optimization was minimizing the loss function, also referred to as the cost, which signifies the predictive capability of the model. Lower costs denote better-performing models (Khan and Ahmed, 2020). Each model requires different parameters to be optimized. Grid search is a widely adopted approach for hyperparameter optimization (Chicco, 2017). To enhance computational efficiency, the hyperparameter tuning was confined to parameters that significantly affect model performance. Table 3 presents the hyperparameter tuning space for each model. Parameters such as learning rate, number of estimators, criterion, batch size, max depth, and others were fine-tuned. It is important to note that parameter tuning is not guaranteed for global optimal but local optimal.

The best parameter combinations were searched using 3-fold cross-validation in the tuning space. Fig. 3(a) illustrates the normalized accuracy of various machine learning models over a series of search iterations. The number of iterations for each model is determined by the defined tuning space, which is why there are varying iteration counts for each model. It can be observed that the normalized accuracy for all models tends to increase with the number of iterations, suggesting that parameter tuning is effectively enhancing model performance. The LightGBM and CatBoost models show more significant improvements as iterations increase compared to other models. In contrast, the Random Forest model shows less improvement in the first half of the iterations, maintaining a relatively lower accuracy throughout.

Fig. 3(b) presents the log loss for the same models over search iterations, a metric where lower scores indicate better model performance. There is a downward trend in log loss values across all models, suggesting a general improvement in model accuracy as the search progresses. Sharp increases in log loss at certain points can be interpreted as instances where the model parameters did not align well with the prediction task, potentially leading to poorer performance. Notably, CatBoost showed stability in log loss in the later iterations. Overall, CatBoost achieved the highest training accuracy and lowest training log loss. Fig. 3(c) and (d) present the training and validation accuracy in detail, alongside the validation confusion matrix for the CatBoost model.

After assessing the training performance, it is crucial to evaluate the predictive performance. As shown in Table 4, the CatBoost model produces the highest prediction accuracy of 89.84 %. This model also has the highest weighted average F1 score at 0.90, which is a harmonized metric accounting for both precision and recall. Furthermore, the CatBoost model demonstrates a commendable precision and recall of 0.90, which emphasizes its capability to correctly identify instances among different classes. These results establish CatBoost as the preferred model for the subsequent clustering analysis.

5.2. Tree SHAP value for CatBoost model

Utilizing tree SHAP algorithms, the study examined variable impacts on pedestrian crash severity predictions using the CatBoost model. Fig. 4 presents the top 20 features by absolute mean SHAP values. The feature importance results vary for each severity level. For fatal crashes, party sobriety impairment status emerges as the most impactful feature, indicating that impairment status can significantly affect the severity of pedestrian injuries. This influence likely arises from reduced awareness, diminished attention to traffic, and impaired judgment of safe crossing times associated with impairment. For instance, Hossain et al. (2023) found that alcohol-impaired older pedestrians are more likely to be involved in fatal crashes. Batouli et al. (2020) identified impairment as a significant factor associated with the severity of outcomes of motor vehicle crashes. Lighting conditions rank as the second most significant factor. Traffic fatalities per 100,000 people, also stand out, suggesting that fatal pedestrian crashes are closely associated with macro-level traffic fatality characteristics.

Interestingly, macro-level features such as the percent of people age 16+ unemployed, PM2.5 air quality levels, total population, AADT, and the average commute time to work are also among the top 20 variables associated with fatal crashes. Features like crash type and location type have considerable influence as well. The type of location is notable and potentially related to operating speeds. Several prior studies have suggested that higher speeds have more association with severe outcomes (Adanu et al., 2023; Islam, 2023). Traffic operations and the underlying cause of the crash also contribute to the model's predictions, reflecting the complexity of factors that can lead to a fatal crash. Moderately influential features pedestrian actions and drug or physical impairment, both having significant behavioral and condition impacts that may lead to severe outcomes. Injury crashes show a slightly different hierarchy of

Table 3
Hyperparameter tuning space.

Models	Parameters
CatBoost	'iterations': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1], 'depth': [5, 10,15], 'l2_leaf_reg': [1, 3, 5, 7]
LightGBM	'num_leaves': [31, 63, 127], 'learning_rate': [0.01, 0.05, 0.1], 'n_estimators': [100, 200, 300], 'min_split_gain': [0.0, 0.1, 0.2]
XGBoost	'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1], 'max_depth': [3, 5, 7] 'min_child_weight': [1, 3, 5]
Random Forest	'n_estimators': [100, 300, 500], 'criterion': ['gini', 'entropy', 'log_loss'],'max_depth': [2, 5, 7]
Gradient Boosting	'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1] 'loss': ['log_loss'], 'max_depth': [2, 3, 5]

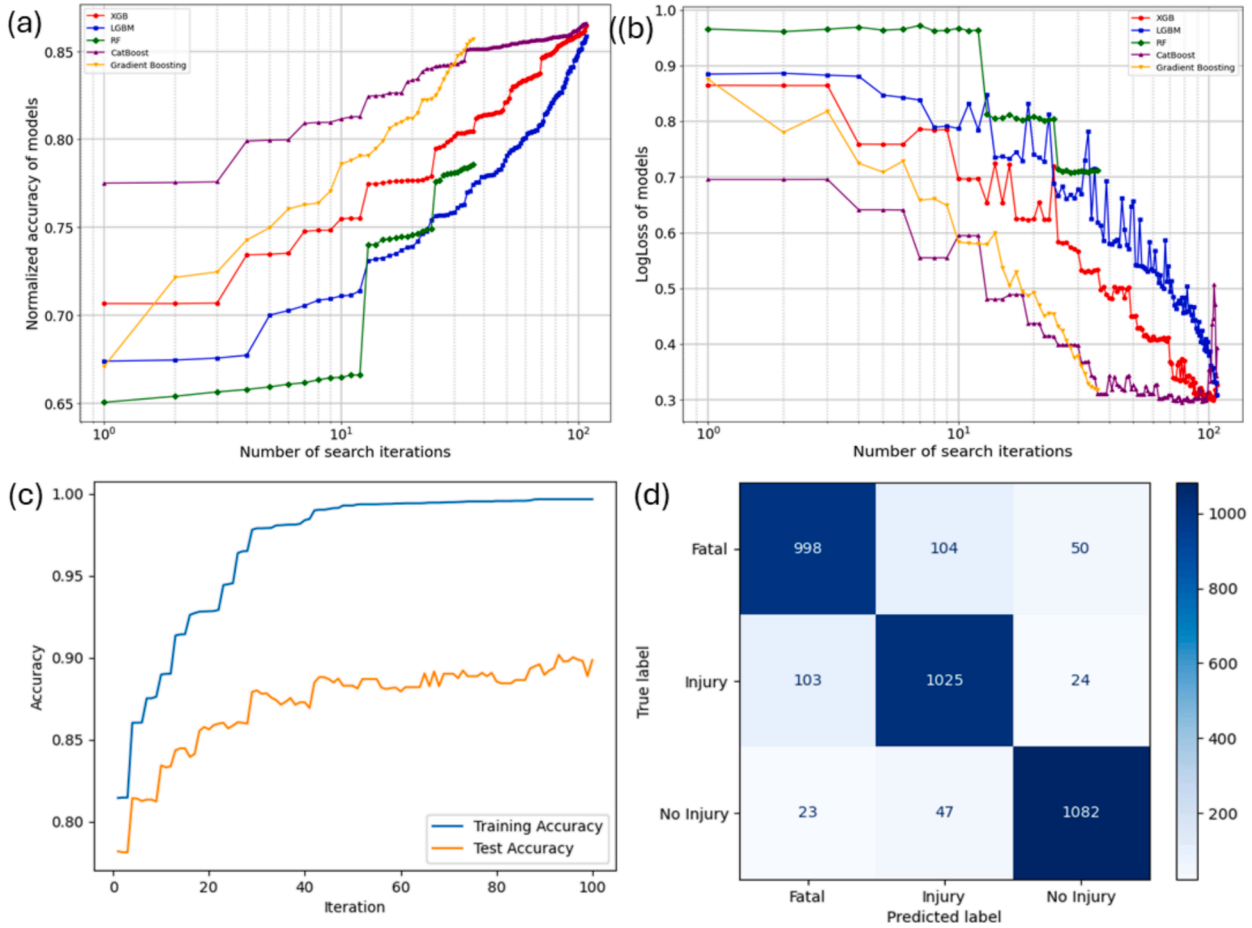


Fig. 3. Model performance across search iterations: (a) Normalized accuracy; (b) Log loss; (c) Training vs. validation accuracy of CatBoost; (d) Validation confusion matrix of CatBoost.

Table 4
Crash severity prediction performance.

Model	Optimized Parameters	Accuracy	Weighted Avg		
			Precision	Recall	F1
XGBoost	'n_estimators': 200, 'min_child_weight': 1, 'max_depth': 3, 'learning_rate': 0.1	89.50 %	0.89	0.89	0.89
Random Forest	'n_estimators': 300, 'max_depth': 7, 'criterion': 'entropy'	79.57 %	0.80	0.80	0.80
CatBoost	'learning_rate': 0.1, 'l2_leaf_reg': 1, 'iterations': 300, 'depth': 10	**89.84 %**	**0.90**	**0.90**	**0.90**
LightGBM	'num_leaves': 127, 'n_estimators': 300, 'min_split_gain': 0.1, 'learning_rate': 0.01	89.35 %	0.89	0.89	0.89
Gradient Boosting	'n_estimators': 100, 'max_depth': 2, 'loss': 'log_loss', 'learning_rate': 0.05	88.34 %	0.88	0.88	0.88

feature importance. Compared to fatal pedestrian crashes, the variables differ significantly, with more social and environmental-related variables – such as average commute time to work, percent of no-car households, percent of occupied houses spend 30 % or more on housing, GINI index, and percent of hazardous sites – presented as important variables in predicting injury crashes. The feature sobriety impairment status is also the most influential. Lighting conditions and location type also have considerable importance. Traffic fatalities and cause follow, highlighting the role of overall crash frequency and causative factors.

Compared to crashes resulting in injury, more social and environmental equity-related variables are identified as important factors in predicting the occurrence of no-injury pedestrian crashes. These factors include traffic fatalities, PM2.5 levels, total population, average commute time to work, AADT, percent of tract risk management sites, percent of population 17 or younger, percent of occupied houses spend 30 % or more on housing, percentage of households without a car, percent of tract within 1 mile of toxics release sites. This also suggests that fewer crash-related or road-related features are considered important in not-injured crashes. Crash type is the most influential feature in predicting no-injury crashes followed by lightning conditions. Sobriety impairment status still plays an important role, mostly because most of the cases are not under the influence in the no-injury subset.

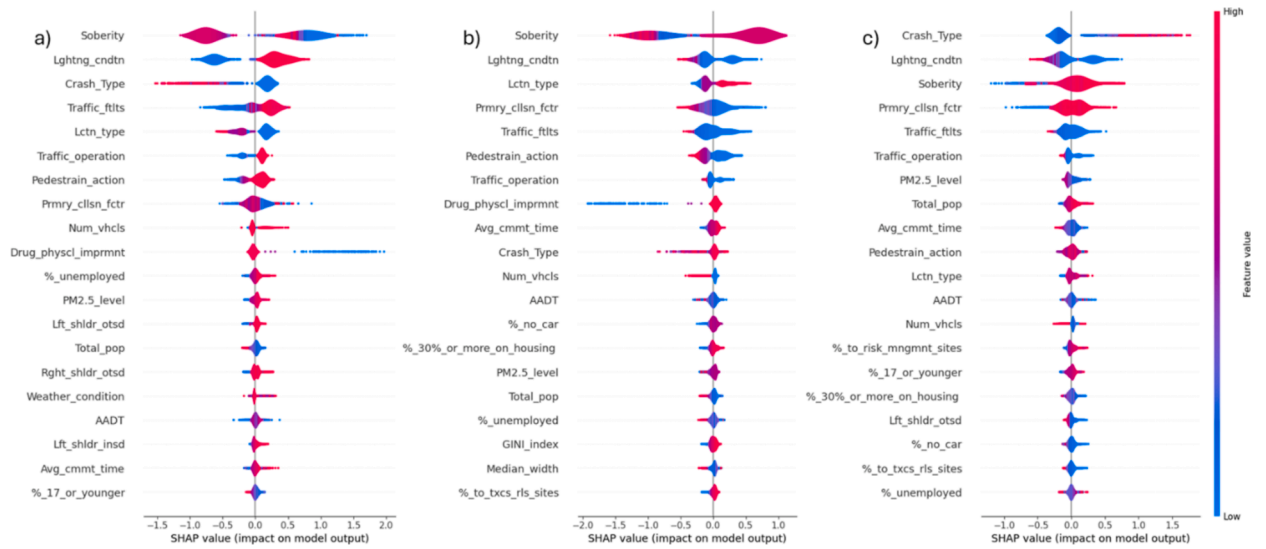


Fig. 4. Global importance of top 20 variables for each crash severity level using tree SHAP Explainer: a) Fatal crash; b) Injury crash; c) No injury crash.

5.3. Clustering patterns by severity

To further identify clustering patterns across different severity levels, hierarchical clustering was applied to the SHAP values. Fig. 5 presents the dendrogram for pedestrian-involved crashes at varying severities and includes corresponding silhouette scores for different numbers of clusters. The dendrograms for fatal (a), injury (b), and no injury (c) crashes illustrate the grouping of instances according to their SHAP values. On each dendrogram, the x-axis labels individual samples, while the y-axis quantifies the dissimilarity between clusters, with branch lengths signifying the point of cluster merging.

The silhouette score plot (d) presents cluster coherence over a range from 2 to 6 clusters, with scores between 0 and 1; higher scores signify well-differentiated clusters. Fatal crashes peak at three clusters, pointing to the presence of three distinct groups. Both injury and no injury crashes show the highest silhouette scores at two clusters. For fatal crashes (a) three distinct clusters are identifiable, with one notably more separated, indicating three groups with SHAP value profiles. The injury crash dendrogram (b) shows two clusters, suggesting a less complex SHAP value profile compared to fatal crashes. The dendrogram for no injury crashes (c) displays clusters that are closely positioned, revealing a higher similarity in SHAP values. Overall, dendrograms and silhouette scores illustrate a detailed picture of the clustering within the data. The subsequent study will investigate the patterns of these clusters in depth.

5.3.1. Fatal pedestrian crash clusters

Out of a total of 3,840 fatal pedestrian crashes, Cluster 1 includes 1,002 samples (82 %), Cluster 2 consists of 130 samples (11 %), and Cluster 3 has 95 samples (8 %). Fig. 6a) displays the mean SHAP values for the top 20 important features with the highest mean absolute SHAP values. Fig. 6b-i) zooms in on the first 8 features in Fig. 6a), providing a detailed SHAP value distribution by clusters. As shown in Fig. 6a), sobriety status stands out in Cluster 2 with a notably negative mean SHAP value of −0.60 (blue), suggesting an average negative association of this feature with fatal crash outcomes within this group. In contrast, the average association is positive in the other two clusters. Referring to Fig. 6b), the sobriety status shows varying impacts across clusters. For instance, the 'Sober' attribute, prominent in Cluster 2, presents negative SHAP values, whereas 'Influenced' and 'Drugged' have positive values and are major components of Cluster 1. Cluster 3, with a minor positive SHAP value of 0.60, mainly includes unstated impairment statuses and HBD sober. Moreover, drug or physical impairment status is positively associated with Cluster 3 but negatively associated with the other clusters, making this a significant factor contributing to fatal outcomes in Cluster 3 while not in the other two clusters. The following paragraphs provide a detailed analysis of each cluster and explore potential strategies with a focus on equity considerations.

Fatal Cluster 1: Nighttime Highway Crashes in Underserved Communities with Poor Visibility and Impairment Among Involved Parties

Cluster 1, which accounts for 82 % of all fatal crashes, is primarily characterized by cases where the party's sobriety impairment status is undetermined. However, SHAP analysis reveals positive values for the 'Unknown' sobriety category, suggesting that some of these unknown cases might involve impairment. In this cluster, 7 % of fatal pedestrian crashes were confirmed to be under the influence of alcohol, and 5 % were impaired by drugs, with both categories also exhibiting positive SHAP values. Most of these crashes occur on highways (69 %), where vehicles typically travel at higher speeds, reflected in common speed zones of 60–70 mph, contributing to the high risk. The light conditions are predominantly dark, with 40 % of crashes occurring in unlit dark environments and 33 % in lighted dark conditions, indicating that poor visibility plays a significant role. In this cluster, drug or physical impairment statuses of all crashes are not stated.

Auto-pedestrian crashes are the most frequent type within this cluster, comprising 74 % of the crashes. Pedestrian actions leading to

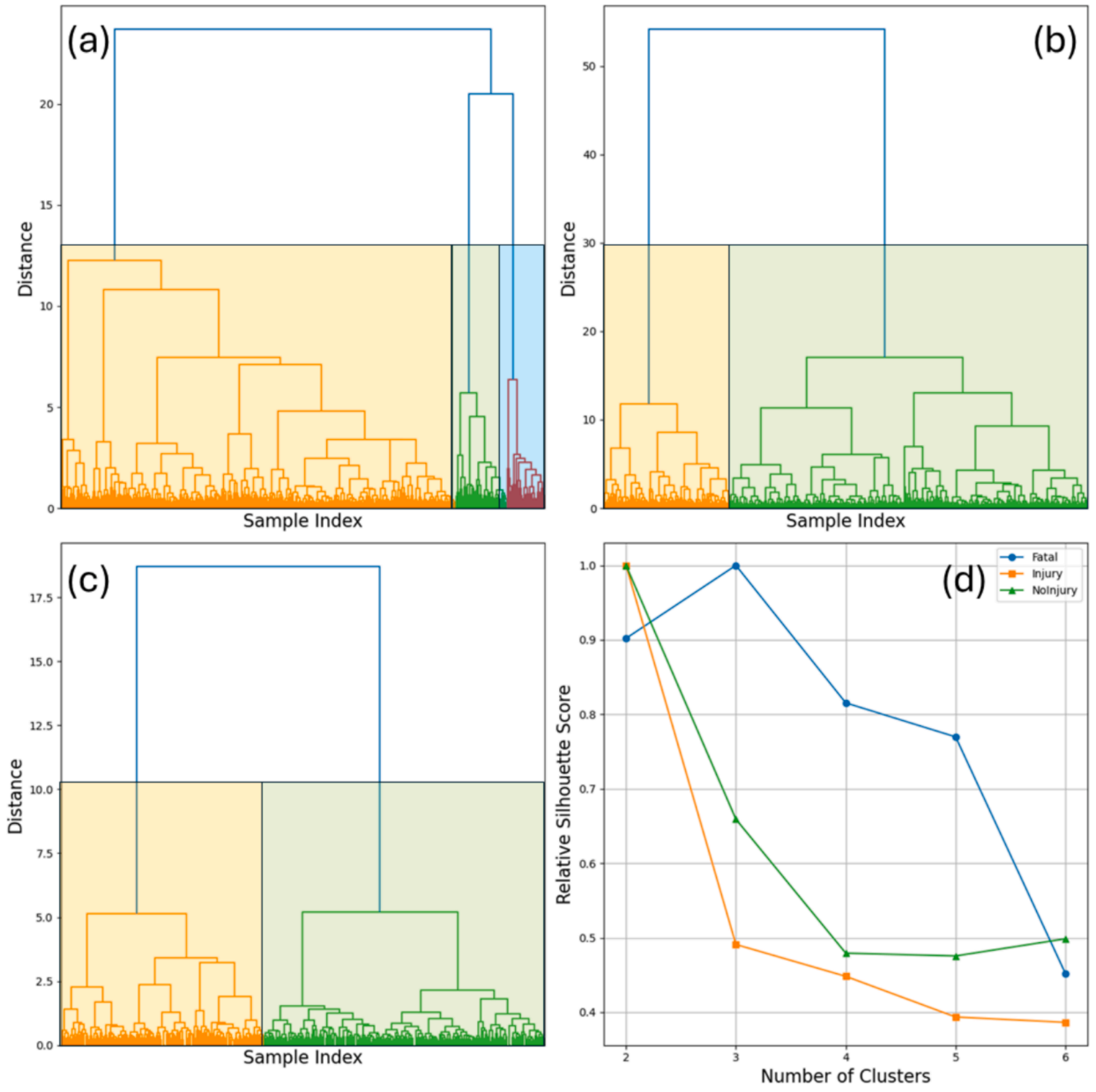


Fig. 5. Dendrogram for SHAP values and silhouette scores: a) Fatal crashes; b) Injury crashes; c) No injury crashes; d) Silhouette scores for different numbers of clusters.

crashes often involve the road shoulder (54 %) and non-crosswalk areas (21 %). The leading cause of crashes is failure to yield (46 %), followed by other violations (20 %) and speeding (7 %). This distribution suggests that driver education, law enforcement, and infrastructural changes might be necessary to address these issues. Single-vehicle crashes are the most common (63 %), with crashes involving 2–4 vehicles at 18 %, indicating that fatal pedestrian crashes often involve one vehicle and occasionally multiple vehicles. A lack of traffic control is evident, with 71 % of crashes happening in areas without traffic controls and only 10 % occurring in areas with controls. The right and left outside shoulder widths (9; 5) are larger than in the other clusters, indicating a better influence of these features.

In Cluster 1, several key factors contribute to fatal crashes. This cluster has a mean traffic fatality rate of 15 per 100,000 people, exceeding the study area's average of 11 per 100,000 people. The mean percent of unemployed is 4.62, AADT is 120,073, and the population is 4,519, both exceeding the area averages of 4.28, 102,307, and 4,439, respectively. PM2.5 levels stand at 11.64, slightly above the study area's 11.47. Commuting time in this cluster averages 28.73 min, longer than in other clusters, highlighting limited accessibility to work. To reduce pedestrian crashes in underserved communities in this cluster, it is essential to improve street lighting and visibility and enhance pedestrian infrastructure. Traffic calming measures, such as speed reductions and increased traffic control

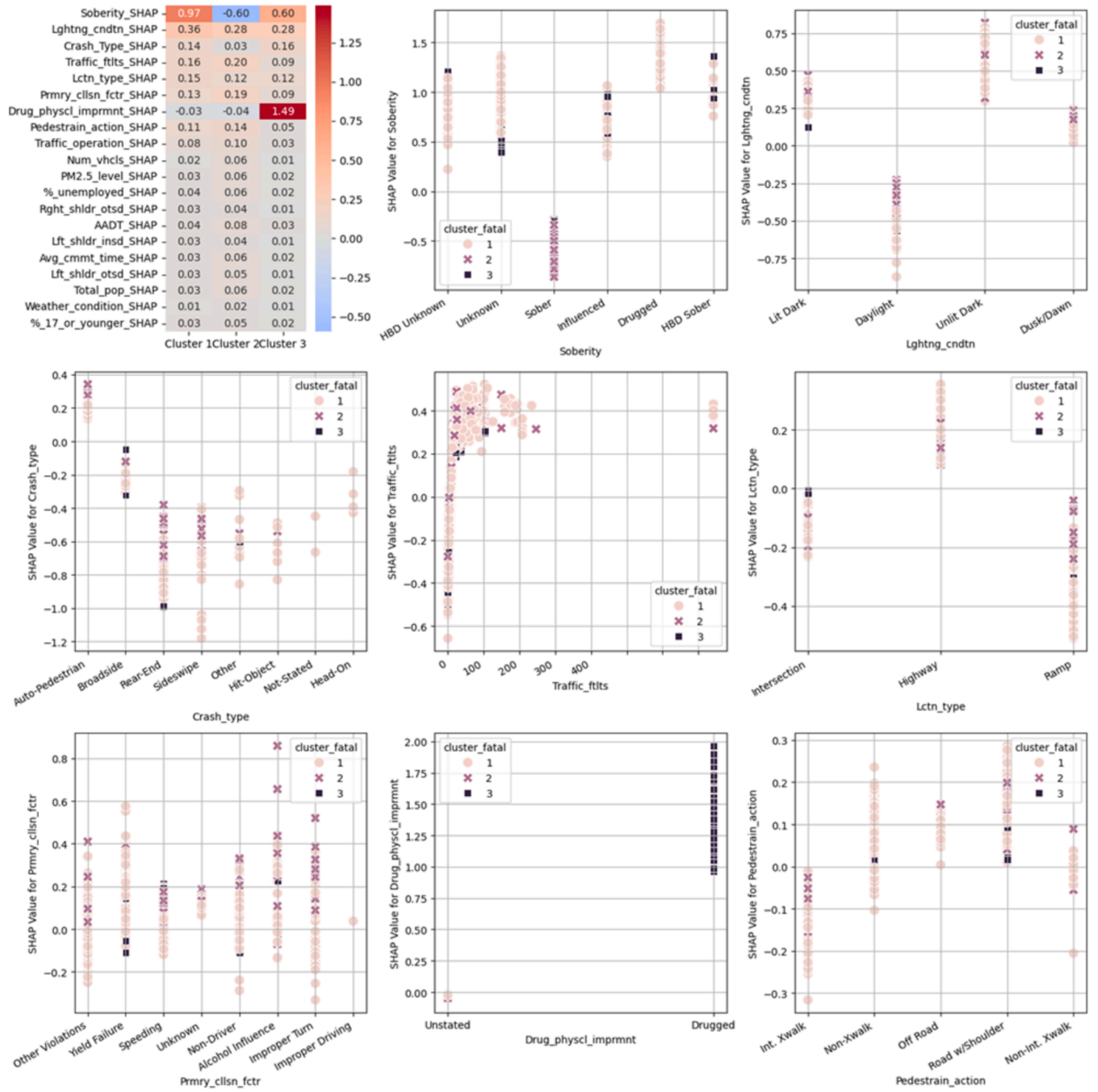


Fig. 6. Cluster analysis for fatal pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features.

with signals and crossings, should also be implemented. Public awareness campaigns and educational programs are vital for promoting road safety. Furthermore, offering public transportation subsidies and developing affordable housing near transit options may also help reduce commute times and transportation costs for residents.

Fatal Cluster 2: Failing to Yield, Sober Pedestrians on Highways with Shoulders in High Traffic Fatality Areas

Cluster 2 accounts for 11 % of fatal crashes. In this cluster, pedestrians are predominantly sober, with 11 % showing no signs of impairment. As illustrated in Fig. 6b), the 'sober' category displays negative SHAP values, indicating that there are other factors that are likely contributing to the crashes in this fatal cluster. The status of pedestrian drug or physical impairment is all unstated. Most crashes occur in dark conditions, either unlighted (4 %) or lighted (3 %), with a slightly higher proportion of crashes in daylight (2 %) than in Cluster 1, suggesting that visibility issues do not exclusively account for the crashes in this cluster.

Auto-pedestrian crashes still make up a significant portion, though less than in Cluster 1, indicating diversity in crash types, evidenced by rear-end and sideswipe collisions. Highways remain the most common location for these crashes, but there are notable occurrences at intersections and ramps. Pedestrian actions leading to crashes often involve the road shoulder, similar to Cluster 1, but

crashes at marked crosswalks also occur, indicating that pedestrian safety measures in these areas could be failing. Yield failure remains a leading cause of crashes, with other violations and speeding also contributing significantly. Crashes typically involve a single vehicle or up to four vehicles, reflecting a broader distribution of crash contexts. The lack of traffic control is significant, with many crashes occurring in areas without controls. Speed limits remain high in crash zones, indicating that high vehicle speeds pose a significant risk despite the sober state of pedestrians.

Cluster 2 shows economic and environmental changes, a slightly lower unemployment rate at 4.25 %, similar to Cluster 1, a modest reduction in PM2.5 levels, a reduction in AADT and commute time. However, the traffic fatality rate is significantly higher at 17.70, compared to the average of 10.62, influenced by factors such as vehicle speed, pedestrian traffic volume, and the effectiveness of traffic control measures. While Cluster 2 shares some similarities with Cluster 1, it differs in having a higher proportion of sober pedestrians, a wider range of crash types, and more incidents occurring at controlled intersections. Additionally, economic, societal, and environmental conditions reveal subtle differences between the two clusters.

Fatal Cluster 3: Non-Crosswalk, Narrow Shoulder Crashes with Alcohol or Drug-Impaired Pedestrians in Underserved Areas

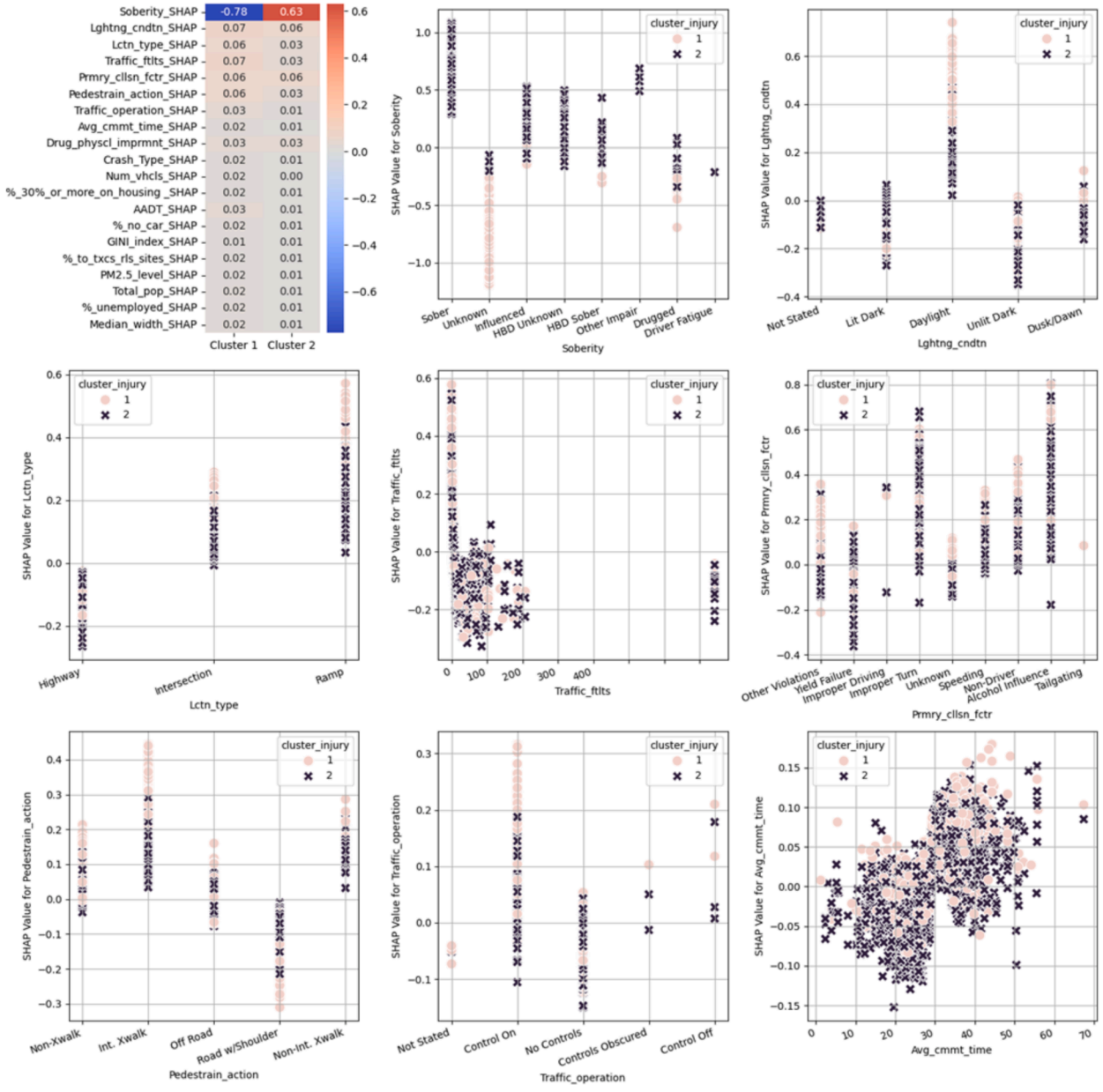


Fig. 7. Cluster analysis for injured pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features.

Cluster 3, which accounts for 8 % of fatal crashes, is notably characterized by a significant proportion of 7.74 % of pedestrian crashes involving drug-impaired individuals, which means almost all crashes in this cluster involved drug-impaired individuals. In this cluster, 4 % of crashes involve parties under the influence of alcohol. Lighting conditions reveal many crashes occurring in lighted dark environments (5 %) and unlighted dark environments (2 %). Daylight and dusk/dawn crashes are less frequent but still present, indicating pedestrian risks across various lighting conditions. Non-crosswalk pedestrian actions (3 %) and crashes on the road shoulder (3 %) highlight the need for improvements in pedestrian infrastructure and behavior modification measures. Other violations are the leading cause of crashes, indicating the need to address various walking behaviors beyond yielding and speed control.

Most crashes in Cluster 3 involve a single vehicle, similar to Clusters 1 and 2, with a lower incidence of multi-vehicle collisions, emphasizing single-vehicle crashes as a critical area for intervention. A notable portion of crashes (6 %) occurred in areas without traffic controls, indicating different traffic management challenges compared to Cluster 1. Infrastructure in this cluster is characterized by narrower shoulder widths, potentially reducing the safety margin for pedestrians. Socioeconomic factors show a slightly lower AADT of 67,580, and the traffic fatality rate stands at 8.50, lower than in the other clusters, suggesting a comparatively safer community than those represented in Clusters 1 and 2. To improve safety, it is essential to enhance pedestrian infrastructure and address behavioral issues related to alcohol and drug impairment. Developing and maintaining pedestrian pathways, especially in areas without crosswalks and on narrow shoulders, is crucial for reducing fatalities.

5.3.2. Injured pedestrian crash clusters

The clustering patterns and the representative factors contributing to injury outcomes within each cluster are visualized in Fig. 7. Compared to fatal pedestrian crashes, the variables differ significantly, with more social and environmental equity-related variables – such as average commute time to work, houses spending 30 % or more on housing with less than 75 k, percent of no-car households, GINI index, and percent tract within a of toxics sites – presented as important variables in predicting injured crashes. Interestingly, the party sobriety impairment status has a positive value, indicating a positive association with fatal outcomes, whereas it tends to have a negative association with fatal outcomes in Cluster 2. Lightning conditions also show a mean positive association in both clusters.

Injured Cluster 1: Poor Lighted, Lacking Pedestrian Infrastructure, Exacerbated by Social Disparities

Cluster 1, representing 26 % of the injury subset, includes many crashes where the impairment status is not reported and sober. According to Fig. 7b), unknown impairments are associated with negative values, probably indicating no obvious impairment. The lighting conditions under which these crashes occur vary, with 100 % happening in dark-lighted environments, 9 % during daylight, and 55 % during dark-unlighted conditions, indicating a higher risk at night. Highways are noted in 14 % of cases and intersections in 7 %. The predominant causes of crashes – 'yield failure' at 10 % and 'other violations' at 9 % – point to broader issues of non-adherence to traffic laws contributing to pedestrian injuries. Traffic operation conditions vary, with 16 % of injuries occurring where there are 'no controls' and 10 % where 'control is on'. Injuries happen on roads with shoulders (10 %), at intersection crosswalks (8 %), and in non-crosswalk areas (6 %).

Demographic and socio-economic factors in areas where these crashes occur reveal an average total population of 4,4335. Additionally, 27 % of households do not own a car, which is higher than the average value of 25 % in the study area. Traffic fatalities further underscore safety concerns, with a rate of 9, significantly lower than the fatal crash subset. The average PM2.5 level is 11, but as it reaches 12, the association with injuries becomes inverse. 46 percent of the area lies within one mile of known toxic release sites. The economic burden is evident, with 41 % of occupied houses spending 30 % or more on housing costs despite earning less than $75,000. The average GINI index is 0.49. Many of these values exceed the average, which may indicate an equity gap between these areas. To address the gap, it is essential to enhance lighting and visibility through the installation of streetlights and reflective road markings, particularly in dark-lighted and dark-unlighted environments.

Injured Cluster 2: Sober Pedestrian Injuries Influenced by Driver Violations and Inadequate Infrastructure

Cluster 2 encompasses most of the dataset, with 74 % of pedestrian injuries. In this cluster, 64 % of cases involved individuals recorded as 'sober' at the time of the crash, with 'influenced' at 4 %. Interestingly, 'sober' has a positive value opposite to the value in the fatal crashes, indicating a positive association with injuries, whereas it tends to have a negative association with fatal severity. This suggests that sobriety tends to reduce the likelihood of fatal crashes but not injuries. The lighting conditions during these crashes were predominantly 'daylight' at 33 % and 'lit dark' at 24 %, indicating that good visibility and other factors contribute to these crashes.

Pedestrian actions involving 'road with shoulder' crashes (31 %) and 'intersection crosswalk' crashes (25 %) show higher figures than in Cluster 1. Highways remain a significant risk, with 39 % of injuries occurring there, similar to Cluster 1. The primary collision factors –'yield failure' at 31 %, 'other violations' at 17 %, and 'speeding' at 11 % – highlight that driver behavior is critical in the occurrence of these injuries. Auto-pedestrian crashes dominate at 56 %, but other types such as 'rear-end', 'sideswipe', 'broadside', and 'hit-object' also contribute to the injury crashes, indicating a more diverse range of crashes. Regarding traffic operations, there is a higher incidence of pedestrian injuries in areas with 'no controls' (44 %) and 'control on' (30 %). To address crashes in this cluster, it is essential to focus on improving infrastructure and addressing driver behavior.

Demographic and socio-economic characteristics in areas where these crashes occur reveal an average total population of 4,435. 97,931 AADT lower than Cluster 1. 23 % of households do not own a car, a figure higher than the study area's average of 25 %. Traffic fatalities highlight safety concerns, with a rate of 8.9 lower than Cluster 1. Additionally, the average PM2.5 level is 11, and 49 % of the area lies within one mile of toxic release sites. Furthermore, 41 % of occupied households spend 30 % or more of their income on housing costs, despite earning less than $75,000. The average GINI index is 0.48, which is positively correlated with injury crashes. These findings are similar to those in Cluster 1.

5.3.3. Not injured pedestrian crash clusters

Fig. 8 visualizes the clustering patterns and the representative factors contributing to not-injured crashes within each cluster. Compared to crashes that result in injury, more social and environmental equity-related variables are presented as the top 20 important variables in predicting the occurrence of no-injury pedestrian crashes. These include traffic fatalities, total population, percent of no-car households, average commute time to work, AADT, percent of tract within 1 mile of risk management plan sites, GINI index, percent of houses spend 30 % or more on housing with less than 75 k, percent of unemployed and percent of tract within 1 mile of toxics release sites. It also means fewer crashes or road related features are presented as important features in results of no injury. It can be observed from the average values of these indicates that these communities do not present as disadvantaged as those in the fatal and injury pedestrian crash categories.

Not Injured Cluster 1: Pedestrian Non-Injury Crashes Without Direct Auto-Pedestrian Contact in Resilient Communities

Cluster 1 represents 42 % of crashes with no reported injuries and encompasses all eight crash types. Notably, rear-end collisions occur frequently at 13 %, and hit-object crashes account for 11 %. These incidents primarily take place in daylight (18 %) and unlit dark conditions (13 %), reflecting a mix of visibility and attentiveness among pedestrians and drivers. Alcohol involvement shows 23 %

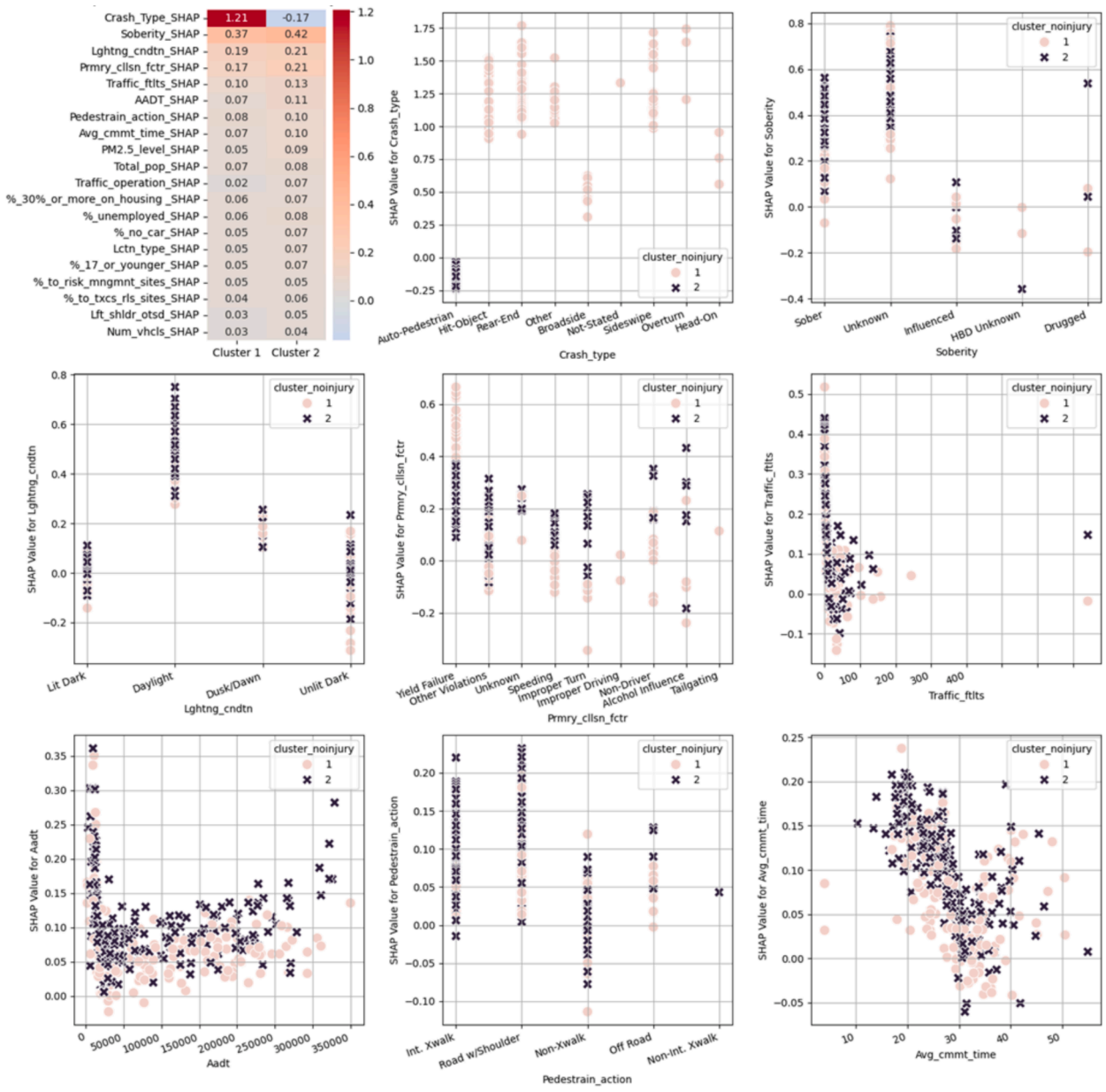


Fig. 8. Cluster analysis for not injured pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features.

unknown status and 16 % sober cases, which, as illustrated in Fig. 8(c), positively correlate with the likelihood of no injury. The primary causes include 'yield failure' at 15 % and 'other violations' at 10 %, underscoring the need for stronger adherence to right-of-way laws. Most incidents involve 2–4 vehicles (22 %), and they commonly occur on highways (31 %) or roads with shoulders (27 %), suggesting that pedestrian caution at crosswalks could mitigate injury risks.

Demographically, these findings reflect less vulnerable community characteristics. The average traffic fatalities stand at 9.62 per 100,000, with an AADT of 92,161. Approximately 22.26 % of sites implement risk management measures, and 20 % or more of households allocate a significant portion of income to housing costs. Additionally, 38.61 % of the area falls within toxic release zones. The population averages around 4,310, with only 7 % of households lacking a vehicle, indicating limited dependence on walking. With a moderate average commute time of 27.5 min, these moderately urbanized communities appear resilient. Overall, it is essential to maintain and enhance the factors contributing to the resilience of these moderately urbanized communities.

Not Injured Cluster 2: Not Injured Crashes in Prevalence of Controlled Settings in High-Density Urbanized Communities

Cluster 2 accounts for 58 % of no-injury crashes, all of which are auto-pedestrian crashes that are negatively associated with non-injury outcomes. Most pedestrians involved in these cases were unimpaired, a factor positively contributing to the likelihood of no injury, as shown in Fig. 8(c). 23 % of these crashes occurred at locations with active traffic control. Crashes in this cluster frequently occur under 'lit dark' (19 %), 'unlit dark' (13 %), and 'daylight' conditions (24 %), underscoring the need to address safety across diverse lighting environments. Yield failure is the leading cause at 33 %, followed by other violations (14 %) and speeding (3 %). A substantial portion (26 %) of crashes occur on roads with shoulders, highlighting a need for enhanced pedestrian pathways, with 19 % of crashes taking place at intersections.

Demographically, this cluster shows a high percentage of car-less households (54 %), significantly above Cluster 1. The traffic fatality rate is 11.22, surpassing Cluster 1. With a slightly larger average population (4,537) and a higher-than-average AADT of 108,869, this cluster likely represents a denser urban environment. The longer average commute time (30 min) suggests a need for improved transportation options, as 54 % of households lack a vehicle-a rate much higher than the study area average of 25 %. With a 30 % longer commute time compared to the area's average, these residents likely depend more heavily on public transit. Policymakers might consider infrastructure improvements such as enhanced street lighting, reflective signage, and pedestrian isolation facilities to reduce the risk of pedestrian-related crashes. Community-specific interventions, like upgrading pedestrian pathways and ensuring safe crossing points, can effectively improve safety in this high-density urban environment.

6. Conclusion

This study utilizes comprehensive crash data from California, spanning from 2018 to 2021, which was gathered from HSIS. It also incorporates environmental justice features sourced from the Equitable Transportation Community, providing a comprehensive dataset that includes extensive socioeconomic, demographic, and build-environment data at the tract level. The dataset encompasses four tiers of information: unit level, crash level, roadway details, and zonal data, from which 77 variables were chosen to develop models predicting pedestrian crash severity across three categories: fatal, injury, and no injury. The study used the Tree SHAP algorithm to explain the influence of each variable on the predictions. Hierarchical clustering is applied to these SHAP values to reveal distinct patterns within clusters.

The analysis using the CatBoost model and tree SHAP values highlights distinct factors influencing the severity of pedestrian crashes. For fatal crashes, key factors include pedestrian impairment, lighting conditions, and overall traffic fatality rates, highlighting the importance of immediate situational factors. These variables are also significant in injury crashes. For non-injury crashes, beyond the key immediate factors, more social equity-related variables emerged as important, indicating a closer association of these factors with property-damage-only pedestrian crashes. The hierarchical clustering of SHAP values provides insights into crash patterns that reveal the importance of integrating crash data with environmental justice concerns. By identifying vulnerable communities where infrastructure inadequacies and socio-economic disparities contribute to higher injury rates, policymakers can prioritize investments in these areas, promoting both safety and equity. For instance, addressing the needs of vulnerable populations by enhancing pedestrian infrastructure and mitigating high housing costs in urban areas could yield more equitable safety outcomes.

Policy implications arising from these findings are complex. First, the strong influence of sobriety status and lighting conditions on fatal crashes suggests that targeted interventions are necessary, such as stricter sobriety enforcement, enhanced street lighting, and improved pedestrian crossings, especially in areas with poor visibility. For injury crashes, the role of social and environmental variables, such as income inequality and commuting challenges, emphasizes the need for policymakers to address systemic inequities. Policies that focus on improving walkability and pedestrian infrastructure in underserved communities could mitigate injury outcomes. Additionally, broader economic interventions that tackle underlying societal issues, like affordable housing and access to public transportation, may have far-reaching impacts on pedestrian safety. The study also highlights the importance of designing safety interventions that are tailored to specific community contexts. In denser urban areas, improving street lighting, expanding pedestrian pathways, and creating safer multimodal transportation networks can significantly reduce the risk of pedestrian crashes. Meanwhile, in suburban or moderately urbanized areas, strengthening adherence to traffic control measures and enhancing roadway design are essential for reducing crash severity. These targeted, context-specific interventions align with the principles of the Safe System approach, which advocates for systemic safety measures that accommodate human error while minimizing fatal and severe crash outcomes.

While this study presents a comprehensive analysis of pedestrian crash severity, it is important to acknowledge its limitations. One key limitation is the absence of pedestrian demographic details, such as age and gender, which are critical factors influencing crash outcomes. Another limitation lies in the clustering methods used. The current study employs a hierarchical clustering approach, but

alternative techniques, such as density-based or partitioning methods, could be explored to better capture the complex, non-linear relationships between crash factors. Incorporating more granular severity levels, beyond the traditional categories of fatal, injury, and no injury, could also reveal subtler patterns. For instance, distinguishing between minor and major injuries crashes could provide deeper insights into the progression of crash severity and potential intervention points. Automated machine learning (AutoML) can automate model selection, hyperparameter tuning, and feature engineering, making the analysis more efficient and less prone to human bias in model selection. There, the application of AutoML represents an innovative future research direction.

CRediT authorship contribution statement

Jinli Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gian Antariksa:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Shriyank Somvanshi:** Validation, Investigation. **Subasish Das:** Writing – review & editing, Supervision, Conceptualization.

Acknowledgement

The authors would like to thank the reviewers for their valuable feedback, with special appreciation to one reviewer who went above and beyond, offering exceptional insights that greatly enhanced the quality of this paper.

Data availability

The authors do not have permission to share data.

References

Adanu, E.K., Dzinyela, R., Agyemang, W., 2023. A comprehensive study of child pedestrian crash outcomes in Ghana. Accid. Anal. Prev. 189, 107146. https://doi.org/10.1016/j.aap.2023.107146.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid. Anal. Prev. 41, 359–364. https://doi.org/10.1016/j.aap.2008.12.014.

Atakishiyev, S., Salameh, M., Yao, H., Goebel, R., 2021. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions [WWW Document]. arXiv.org. URL https://arxiv.org/abs/2112.11561v4 (accessed 4.22.24).

Baireddy, R., Zhou, H., Jalayer, M., 2018. Multiple Correspondence Analysis of Pedestrian Crashes in Rural Illinois. Transp. Res. Rec. 2672, 116–127. https://doi.org/10.1177/0361198118777088.

Batouli, G., Guo, M., Janson, B., Marshall, W., 2020. Analysis of pedestrian-vehicle crash injury severity factors in Colorado 2006–2016. Accid. Anal. Prev. 148, 105782. https://doi.org/10.1016/j.aap.2020.105782.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Anal. Methods Accid. Res 3–4, 56–91. https://doi.org/10.1016/j.amar.2014.10.001.

Bíl, M., Andrášik, R., Janoška, Z., 2013. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accid. Anal. Prev. 55, 265–273. https://doi.org/10.1016/j.aap.2013.03.003.

Braun, L.M., Le, H.T.K., Voulgaris, C.T., Nethery, R.C., 2023. Who benefits from shifting metal-to-pedal? Equity in the health tradeoffs of cycling. Transp. Res. Part D: Transp. Environ. 115, 103540. https://doi.org/10.1016/j.trd.2022.103540.

Breiman, L., 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). ss 16, 199–231. doi: 10.1214/ss/1009213726.

CDC, 2023. Pedestrian Safety | Transportation Safety [WWW Document]. URL https://www.cdc.gov/transportationsafety/pedestrian_safety/index.html (accessed 4.24.24).

Chang, I., Park, H., Hong, E., Lee, J., Kwon, N., 2022. Predicting effects of built environment on fatal pedestrian accidents at location-specific level: application of XGBoost and SHAP. Accid. Anal. Prev. 166, 106545.

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. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. doi: 10.1145/2939672.2939785.

Chen, P., Zhou, J., 2016. Effects of the built environment on automobile-involved pedestrian crash frequency and risk. J. Transp. Health Built Environ Transp. Health 3, 448–456. https://doi.org/10.1016/j.jth.2016.06.008.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. Biodata Min. 10, 1–17. https://doi.org/10.1186/s13040-017-0155-3.

Chimba, D., Musinguzi, A., Kidando, E., 2018. Associating pedestrian crashes with demographic and socioeconomic factors. Case Stud. Transp. Policy 6, 11–16. https://doi.org/10.1016/j.cstp.2018.01.006.

Cilínio, M., Pereira, M., D., D., L., M., P, V., 2023. Explainable fault analysis in mobile networks: A SHAP-based supervised clustering approach. In: Presented at the 16th International Conference on Signal Processing and Communication System (ICSPCS), IEEE, pp. 1–9.

Das, S., 2021. Association rules mining applied to autonomous vehicle safety ratings in BikePGH surveys. In: Proceedings of the 15th International Conference on Pervasive Computing Technologies for Healthcare, pp. 1–5.

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., Jalayer, M., 2019. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int. J. Urban Sci. 23, 30–48. https://doi.org/10.1080/12265934.2018.1431146.

Das, S., Sun, X., 2015. Factor association with multiple correspondence analysis in vehicle-pedestrian crashes. Transp. Res. Rec. 2519, 95–103. https://doi.org/10.3141/2519-11.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40, 1257–1266. https://doi.org/10.1016/j.aap.2008.01.007.

Ding, C., Chen, P., Jiao, J., 2018. Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach. Accid. Anal. Prev. 112, 116–126. https://doi.org/10.1016/j.aap.2017.12.026.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A., 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data.

Escriva, E., Aligon, J., Excoffier, J.-B., Monsarrat, P., Soulé-Dupuy, C., 2023. How to make the most of local explanations: effective clustering based on influences. In: Abelló, A., Vassiliadis, P., Romero, O., Wrembel, R. (Eds.), Advances in Databases and Information Systems. Springer Nature, Switzerland, Cham, pp. 146–160. https://doi.org/10.1007/978-3-031-42914-9_11.

Fan, Y., Zhu, X., She, B., Guo, W., Guo, T., 2018. Network-constrained spatio-temporal clustering analysis of traffic collisions in Jianghan District of Wuhan, China. PLOS ONE 13, e0195093.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.
Gu, Z., Peng, B., 2021. Investigation into the built environment impacts on pedestrian crash frequencies during morning, noon/afternoon, night, and during peak hours: a case study in Miami County, Florida. J. Transp. Safety Secur. 13, 915–935. https://doi.org/10.1080/19439962.2019.1701164.
Habibovic, A., Tivesten, E., Uchida, N., Bärgman, J., Ljung Aust, M., 2013. Driver behavior in car-to-pedestrian incidents: an application of the Driving Reliability and Error Analysis Method (DREAM). Accid. Anal. Prev. 50, 554–565. https://doi.org/10.1016/j.aap.2012.05.034.
Haddad, A.J., Mondal, A., Bhat, C.R., Zhang, A., Liao, M.C., Macias, L.J., Lee, M.K., Watkins, S.C., 2023. Pedestrian crash frequency: unpacking the effects of contributing factors and racial disparities. Accid. Anal. Prev. 182, 106954. https://doi.org/10.1016/j.aap.2023.106954.
Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid. Anal. Prev. 81, 14–23. https://doi.org/10.1016/j.aap.2015.04.025.
Henary, B.Y., Ivarsson, J., Crandall, J.R., 2006. The influence of age on the morbidity and mortality of pedestrian victims. Traffic Inj. Prev. 7, 182–190. https://doi.org/10.1080/15389580500516414.
Hezaveh, A.M., Cherry, C.R., 2018. Walking under the influence of the alcohol: a case study of pedestrian crashes in Tennessee. Accid. Anal. Prev. 121, 64–70. https://doi.org/10.1016/j.aap.2018.09.002.
Hossain, A., Sun, X., Thapa, R., Codjoe, J., 2022. Applying association rules mining to investigate pedestrian fatal and injury crash patterns under different lighting conditions. Transp. Res. Rec. 2676, 659–672. https://doi.org/10.1177/03611981221076120.
Hossain, A., Sun, X., Islam, S., Alam, S., Hossain, M.M., 2023. Identifying roadway departure crash patterns on rural two-lane highways under different lighting conditions: association knowledge using data mining approach. J. Saf. Res. 85, 52–65. https://doi.org/10.1016/j.jsr.2023.01.006.
Huang, B., Hooi, B., Shu, K., 2023. TAP: a comprehensive data repository for traffic accident prediction in road networks. In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23. Association for Computing Machinery, New York, NY, USA, pp. 1–4. 10.1145/3589132.3625655.
Hussain, M.S., Goswami, A.K., Gupta, A., 2023. Predicting pedestrian crash locations in urban India: an integrated GIS-based spatiotemporal HSID technique. J. Transp. Safety Secur. 15, 103–136. https://doi.org/10.1080/19439962.2022.2048759.
Islam, M., 2023. An exploratory analysis of the effects of speed limits on pedestrian injury severities in vehicle-pedestrian crashes. J. Transp. Health 28, 101561. https://doi.org/10.1016/j.jth.2022.101561.
James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J., 2023. Statistical learning. In: An introduction to statistical learning, Springer Texts in Statistics. Springer International Publishing, Cham, pp. 15–67. doi: 10.1007/978-3-031-38747-0_2.
Kaplan, S., Prato, C.G., 2013. Cyclist-motorist crash patterns in Denmark: a latent class clustering approach. Traffic Inj. Prev. 14, 725–733. https://doi.org/10.1080/15389588.2012.759654.
Karim, M.M., Li, Y., Qin, R., 2022. Toward explainable artificial intelligence for early anticipation of traffic accidents. Transp. Res. Rec. 2676, 743–755. https://doi.org/10.1177/03611981221076121.
Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
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.
Kemnitzer, C.R., Pope, C.N., Nwosu, A., Zhao, S., Wei, L., Zhu, M., 2019. An investigation of driver, pedestrian, and environmental characteristics and resulting pedestrian injury. Traffic Inj. Prev. 20, 510–514. https://doi.org/10.1080/15389588.2019.1612886.
Khan, M.N., Ahmed, M.M., 2020. Trajectory-level fog detection based on in-vehicle video camera with TensorFlow deep learning utilizing SHRP2 naturalistic driving data. Accid. Anal. Prev. 142. https://doi.org/10.1016/j.aap.2020.105521.
Khan, M.N., Das, S., Liu, J., 2024. Predicting pedestrian-involved crash severity using inception-v3 deep learning model. Accid. Anal. Prev. 197, 107457.
Kim, K., Yamashita, E.Y., 2007. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transp. 41, 69–89. https://doi.org/10.1002/atr.5670410106.
Kong, X., Das, S., Zhang, Y., Wei, Z., Yuan, C., 2023. In-depth understanding of pedestrian-vehicle near-crash events at signalized intersections: an interpretable machine learning approach. Transp. Res. Rec. 2677, 747–759. https://doi.org/10.1177/03611981221136138.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.
Kumar, S., Toshniwal, D., 2016. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). J Big Data 3, 13. https://doi.org/10.1186/s40537-016-0046-3.
LaScala, E.A., Johnson, F.W., Gruenewald, P.J., 2001. Neighborhood characteristics of alcohol-related pedestrian injury collisions: a geostatistical analysis. Prev. Sci. 2, 123–134. https://doi.org/10.1023/A:1011547831475.
Lasota, D., Goniewicz, M., Kosson, D., Ochal, A., Krajewski, P., Tarka, S., Goniewicz, K., Mirowska-Guzel, D., 2019. The effect of ethyl alcohol on the severity of injuries in fatal pedestrian victims of traffic crashes. PLoS One 14, e0221749.
Li, D., Ranjitkar, P., Zhao, Y., Yi, H., Rashidi, S., 2017. Analyzing pedestrian crash injury severity under different weather conditions. Traffic Inj. Prev. 18, 427–430. https://doi.org/10.1080/15389588.2016.1207762.
Li, X., Yu, S., Huang, X., Dadashova, B., Cui, W., Zhang, Z., 2022. Do underserved and socially vulnerable communities observe more crashes? A spatial examination of social vulnerability and crash risks in Texas. Accid. Anal. Prev. 173, 106721. https://doi.org/10.1016/j.aap.2022.106721.
Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable ai: a review of machine learning interpretability methods. Entropy 23, 18.
Lu, S., Chen, R., Wei, W., Belovsky, M., Lu, X., 2022. Understanding heart failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions. AMIA Annu. Symp. Proc. 2021, 813–822.
Lundberg, S.M., Erion, G.G., Lee, S.-I., 2019. Consistent Individualized Feature Attribution for Tree Ensembles.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Mafi, S., AbdelRazig, Y., Doczy, R., 2018. Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. Transp. Res. Rec. 2672, 171–183. https://doi.org/10.1177/0361198118794292.
Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M.A., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S.E., Davis, K.C., Megahed, F.M., 2020. A Review of Data Analytic Applications in Road Traffic Safety. Part 1: Descriptive and Predictive Modeling. Sensors 20, 1107. https://doi.org/10.3390/s20041107.
Niebuhr, T., Junge, M., Rosén, E., 2016. Pedestrian injury risk and the effect of age. Accid. Anal. Prev. 86, 121–128. https://doi.org/10.1016/j.aap.2015.10.026.
Pour-Rouholamin, M., Zhou, H., 2016. Investigating the risk factors associated with pedestrian injury severity in Illinois. J. Saf. Res. 57, 9–17. https://doi.org/10.1016/j.jsr.2016.03.004.
Prange, M., Heller, M., Watson, H., Iyer, M., Ivarsson, B.J., Fisher, J., 2010. Age Effects on Injury Patterns in Pedestrian Crashes. SAE Int. J. Passeng. Cars – Mech. Syst. 3, 789–820. https://doi.org/10.4271/2010-01-1164.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.
Rahim, M.A., Hassan, H.M., 2021. A deep learning based traffic crash severity prediction framework. Accid. Anal. Prev. 154, 106090. https://doi.org/10.1016/j.aap.2021.106090.
Rahman, M.S., Abdel-Aty, M., Hasan, S., Cai, Q., 2019. Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. J. Saf. Res. 70, 275–288. https://doi.org/10.1016/j.jsr.2019.04.008.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Sasidharan, L., Wu, K.-F., Menendez, M., 2015. Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. Accid. Anal. Prev. 85, 219–228. https://doi.org/10.1016/j.aap.2015.09.020.
Shapley, L., 1997. 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317., In: Kuhn, H.W. (Ed.), Classics in Game Theory. Princeton University Press, pp. 69–79. doi: 10.1515/9781400829156-012.
Sheykhfard, A., Haghighi, F., Papadimitriou, E., Van Gelder, P., 2021. Review and assessment of different perspectives of vehicle-pedestrian conflicts and crashes: passive and active analysis approaches. J. Traff. Transp. Eng. (Engl. Ed.) 8, 681–702. https://doi.org/10.1016/j.jtte.2021.08.001.

Sivasankaran, S.K., Balasubramanian, V., 2021. Severity of pedestrians in pedestrian - bus crashes: an investigation of pedestrian, driver and environmental characteristics using random forest approach. In: Black, N.L., Neumann, W.P., Noy, I. (Eds.), Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021). Springer International Publishing, Cham, pp. 825–833. doi: 10.1007/978-3-030-74608-7_101.

Song, L., Li, Y., Fan, W. (David), Wu, P., 2020. Modeling pedestrian-injury severities in pedestrian-vehicle crashes considering spatiotemporal patterns: insights from different hierarchical Bayesian random-effects models. Anal. Methods Accid. Res 28. https://doi.org/10.1016/j.amar.2020.100137.

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

Taamneh, M., Taamneh, S., Alkheder, S., 2017. Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. Int. J. Inj. Contr. Saf. Promot. 24, 388–395. https://doi.org/10.1080/17457300.2016.1224902.

Toran Pour, A., Moridpour, S., Tay, R., Rajabifard, A., 2018. Influence of pedestrian age and gender on spatial and temporal distribution of pedestrian crashes. Traffic Inj. Prev. 19, 81–87. https://doi.org/10.1080/15389588.2017.1341630.

Truong, L.T., Somenahalli, S.V.C., 2011. Using GIS to identify pedestrian-vehicle crash hot spots and unsafe bus stops. J. Public Transp. 14, 99–114. https://doi.org/10.5038/2375-0901.14.1.6.

United States Department of Transportation, 2024. Highway Safety Information System (HSIS) | FHWA [WWW Document]. URL https://highways.dot.gov/research/safety/hsis (accessed 4.22.24).

US Department of Transportation, 2024. Equitable Transportation Communitys [WWW Document]. URL https://www.transportation.gov/priorities/equity/justice40/download-data (accessed 4.22.24).

Wang, X., Yang, J., Lee, C., Ji, Z., You, S., 2016. Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accid. Anal. Prev. 96, 12–21. https://doi.org/10.1016/j.aap.2016.07.028.

Weiss, H.B., Kaplan, S., Prato, C.G., 2016. Fatal and serious road crashes involving young New Zealand drivers: a latent class clustering approach. Int. J. Inj. Contr. Saf. Promot. 23, 427–443. https://doi.org/10.1080/17457300.2015.1056807.

World Health Organization, 2024. Road traffic injuries [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed 4.24.24).

Xie, Z., Yan, J., 2013. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. J. Transp. Geogr. 31, 64–71. https://doi.org/10.1016/j.jtrangeo.2013.05.009.

Younes, H., Noland, R.B., Von Hagen, L.A., Meehan, S., 2023. Pedestrian-and bicyclist-involved crashes: associations with spatial factors, pedestrian infrastructure, and equity impacts. J. Saf. Res.

Yue, H., 2024. Investigating the influence of streetscape environmental characteristics on pedestrian crashes at intersections using street view images and explainable machine learning. Accid. Anal. Prev. 205, 107693. https://doi.org/10.1016/j.aap.2024.107693.

Zhang, Y., Han, L.D., Kim, H., 2018. Dijkstra's-DBSCAN: fast, accurate, and routable density based clustering of traffic incidents on large road network. Transp. Res. Rec. 2672, 265–273. https://doi.org/10.1177/0361198118796071.