Accident Analysis and Prevention 220 (2025) 108180

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Intersection crash analysis considering longitudinal and lateral risky driving behavior from connected vehicle data: A spatial machine learning approach

Lei Han *, Mohamed Abdel-Aty

Department of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL 32816, United States

ARTICLE INFO

Keywords:
Intersection safety
Connected vehicle data
Risky driving behavior
Spatial heterogeneity
Rear-end crashes
Turn crashes

ABSTRACT

Existing intersection safety analysis studies have primarily focused on macro-level static infrastructure and highly aggregated traffic features. The emergence of Connected Vehicle (CV) has enabled researchers to extract micro-level driving behavior attributes from CVs. Although longitudinal driving behaviors (e.g., hard braking) have been studied recently, critical lateral left and right turn behaviors, which are common and pose potential conflict risk at intersections, have been largely overlooked. Meanwhile, dealing with both spatial heterogeneity and nonlinear effects between crash frequency and multitudinous driving features is another critical challenge for intersection safety analysis. To address such gaps, this study extracted driving behavior features for both longitudinal movements and lateral left and right turns to comprehensively capture driving dynamics at intersections. A novel spatial ML framework was proposed to integrate nonlinear ML models (e.g., LightGBM) with geographically weighted regression: Besides a global ML model training on all samples to fit average estimations, distinct local ML models are trained for each spatial sample with its neighbors to capture localized spatial heterogeneity. Empirical experiments using CV data at a Florida county show that the inclusion of lateral turning behavior (e.g., hard left/right turns) leads to improved accuracy of intersection crash frequency prediction. Compared to traditional Rrandom Forest, XGBoost, LightGBM, and Multilayer Perceptron models, the spatial ML integrating LightGBM demonstrates significant improvements of 5.8%, 6.3%, and 5.6% in RMSE, MAE, and R^2, respectively. The results reveal the nonlinear impact of driving features and their spatial heterogeneity: In downtown, hard braking events primarily influence the risk of rear-end (RE) crashes. Drivers' acceleration also is more likely to lead to RE crashes in urban areas. While hard left turns show greater influence of sideswipe and left turn crashes at suburban intersections.

1. Introduction

Intersections have been recognized as crash-prone locations within the urban traffic network due to the complex vehicle movements, multimodal interactions, and conflicts from different approaches. In the United States, more than 25 % of traffic fatalities and 50 % of traffic injuries occur at or near intersections each year (FHWA, 2024). In 2022, 42,514 traffic fatalities were recorded, of which 12,036 involved intersections—accounting for 28.3 % of all fatalities, causing a significant toll on people's lives and property (FHWA, 2024). Consequently, enhancing intersection safety is a critical step toward saving human lives and realizing Vision Zero (USDOT). To achieve this goal, extensive efforts have been dedicated to intersection crash frequency modeling and safety evaluation (Gu et al., 2023; Kabir et al., 2021; Yuan and Abdel-Aty, 2018). Specifically, researchers have identified various contributing factors (e.g., intersection traffic, geometric design, etc.) and employed statistical and machine learning (ML) models to assess their impacts on intersection crash frequency (Kabir et al., 2021; Lee et al., 2023b), therefore providing valuable insights for traffic engineers to implement targeted countermeasures to reduce intersection crashes (Wang et al., 2024a; Wu et al., 2023; Wang et al., 2025).

However, previous studies mainly relied on macro-level infrastructure and traffic features, barely considering micro-level human driving behaviors. Crashes—particularly within intersection areas—are mainly caused by drivers' risky driving behavior and failure to interact appropriately with other vehicles (Han et al., 2024a,b; Shirazi and Morris,

* Corresponding author.
E-mail addresses: le966091@ucf.edu (L. Han), m.aty@ucf.edu (M. Abdel-Aty).

https://doi.org/10.1016/j.aap.2025.108180
Received 29 April 2025; Received in revised form 3 July 2025; Accepted 24 July 2025
Available online 30 July 2025

2017). Compared to macro features describing road traffic environment, micro driving behaviors inherently capture risky driving motions and interactions, which are more closely related to intersection crashes (Gu et al., 2023). Leveraging emerging high-resolution Connected Vehicle (CV) data, recent studies started to extract micro driving features and examine their impact on intersection crashes (Gu et al., 2023; Hunter et al., 2021; Kamrani et al., 2018). While these studies highlighted key impacts of risky driving behaviors on intersection crashes, they have primarily focused on longitudinal behaviors (e.g., hard acceleration and braking). In reality, vehicles at intersection engage in both longitudinal (e.g., driving straight) and lateral (e.g., left/right turn) interactions. Critical lateral driving behaviors—such as hard left and right turns—exhibit distinct patterns from longitudinal actions (e.g., car-following) to capture risky lateral interactions related to vehicle safety (Sander, 2017; Shirazi and Morris, 2017). For instance, a vehicle executing a hard left turn can often sideswipe an oncoming vehicle. If

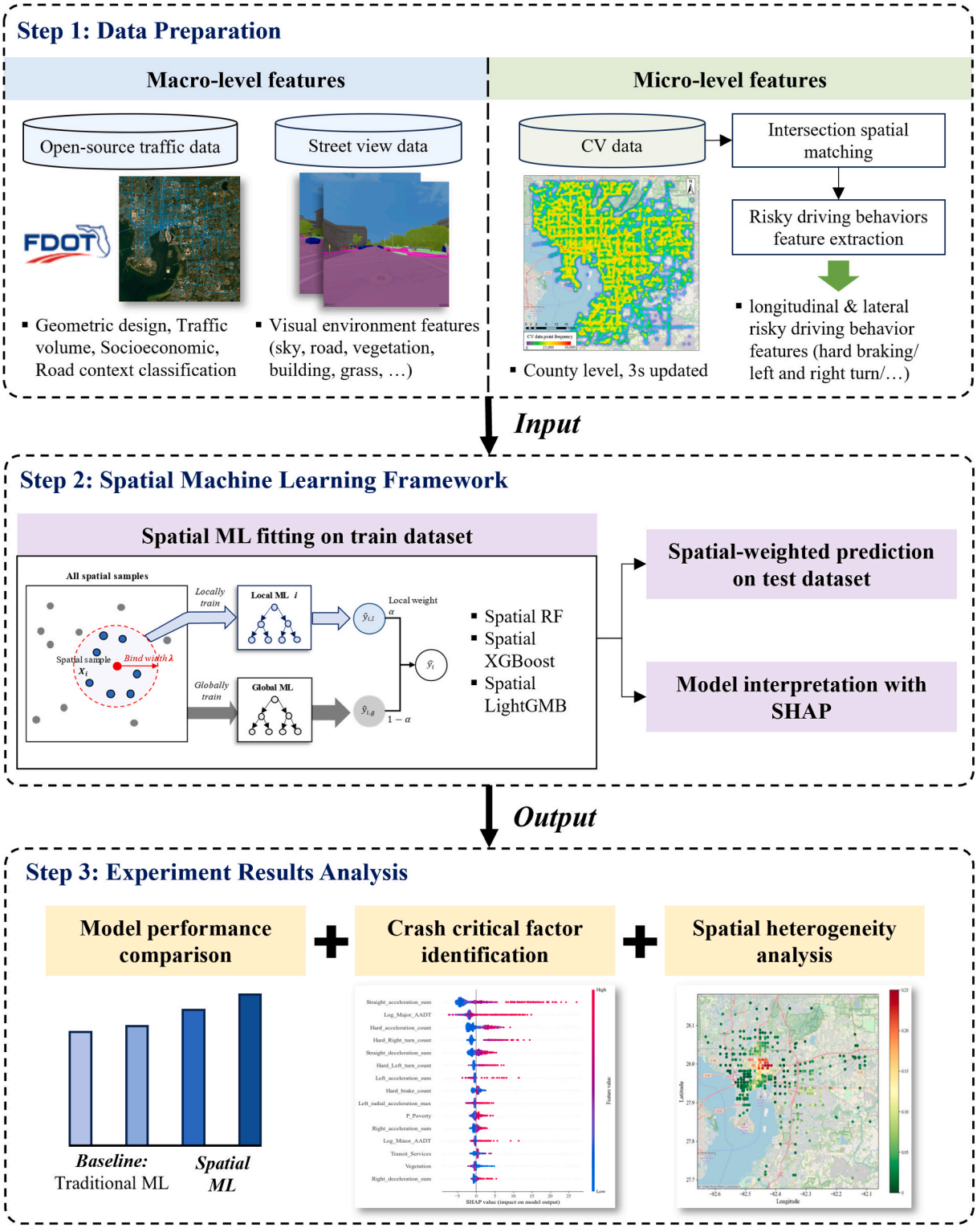


Fig. 1. Overall workflow of this study.

the oncoming vehicle brakes hard to avoid the collision, it may in turn be rear-ended by trailing traffic. Therefore, it is important to incorporate such risky turning maneuvers into intersection safety analysis. However, existing studies either overlooked these turning maneuvers or mixed them with longitudinal behaviors, making it difficult to assess their impact on intersection safety and introducing potential estimation biases. To the best of our knowledge, no existing study has yet analyzed risky driving behaviors specific to critical lateral turning movements and established their relationship with intersection crash frequencies.

For intersection crash frequency modeling, spatial approaches have emerged as a primary focus for addressing the spatial heterogeneity issue in crash analysis (Arvin et al., 2019; Tang et al., 2020; Wang et al., 2024). Various spatial statistical methods have been developed including random parameter models (Lee et al., 2023a,b; Wang et al., 2024b), spatial lag models (Hong et al., 2016), and geographically weighted regression (GWR) (Li et al., 2022). Although existing spatial statistical methods can capture complex spatial heterogeneity well, they still strongly rely on the linear relationship assumption and struggle with the high-dimensional traffic datasets (Wen et al., 2021; Zhou et al., 2023). To address such limitation, recent studies have shifted toward exploring integration spatial heterogeneity effects into ML frameworks as ML models are significantly effective in modeling nonlinear and high-dimensional datasets (Fan et al., 2023a; Wen et al., 2022; Zhou et al., 2023). For example, the Geographical Random Forest (GRF) model, which integrates Random Forest (RF) into geographically weighted model, has been employed to account for spatial heterogeneity and achieved notable prediction accuracy in recent traffic safety modeling (Wu et al., 2024; Wang et al., 2024). However, the potential of integrating other ML methods (e.g., XGBoost and LightGBM), known for robust nonlinear fitting capabilities, into spatial modeling has yet to be investigated (Sigrist, 2023). Therefore, developing novel spatial ML approaches that can capture both nonlinear effects and spatial heterogeneity in intersection crash modeling remains to be further investigated.

To address the aforementioned research gaps, this study aims to extract multiple risky driving behavior features at intersections and capture their nonlinear effects and spatial heterogeneity on intersection crashes. The overall workflow of this study is shown in Fig. 1. First, multiple macro-level features are extracted from open-source traffic and street view data, and micro-level driving behavior features are identified from county-scale CV data. Second, these features are fed into the proposed novel spatial ML framework for model testing and interpretation. Finally, the experimental analysis focuses on model performance, identifying crash-critical factors, and examining their spatial heterogeneity.

Overall, the main contributions of this paper include:

1) Identifying high risk driving behavior features for both longitudinal (e.g., hard braking and acceleration) and lateral turning (e.g., hard left and right turn) movements from CV trajectories to analyze their impact on intersection crashes.
2) Proposing a spatial ML framework that integrates multiple ML models (e.g., RF, XGBoost and LightGBM) with geographically weighted regression to account for spatial heterogeneity in intersection crash frequency modeling.

Following this, section 2 reviews related literature and section 3 details data preparation. Section 4 shows the proposed methodology and section 5 illustrates the experiment results. The discussion and conclusion of this study are presented in Section 6 and 7, respectively.

2. Literature review

2.1. CV-data-based intersection driving behavior features

Benefiting from the development of vehicle connectivity technologies, it is now possible to obtain micro-level driving behavior features from high-resolution CV data to reflect detailed driving interactions within intersections. Table 1 summarizes the existing intersection crash studies considering driving behavior features from CV datasets. For instance, a group of studies have utilized CV data from the Safety Pilot Model Deployment (SPMD) Program to extract several driving volatility measures (e.g., standard deviation, coefficient of variation of speed and longitudinal acceleration) and link them with intersection crash frequencies (Arvin et al., 2019; Hu et al., 2020; Kamrani et al., 2018; Wali et al., 2018). Similarly, Hunter et al. (2021) leveraged hard-braking events from Wejo CV data to predict rear-end crash frequency, indicating a strong correlation between these events and crashes.

Although these studies highlighted key impacts of risky driving behaviors on intersection crashes, existing features have primarily focused on longitudinal behaviors (e.g., hard acceleration and braking). While some studies have incorporated lateral accelerations, they simply aggregated these features at the overall intersection level without distinguishing specific vehicle movements (e.g., left/right turn) (Arvin et al., 2019; Gu et al., 2023). However, left- and right-turn behaviors at intersections exhibit distinct interaction patterns with surrounding traffic (Sander, 2017; Shirazi and Morris, 2017), yet are still overlooked. From a traffic management perspective, it is more important to identify safety issues at the level of specific vehicle movements rather than high-aggregated lateral behavior measures. For instance, frequent hard left-turn maneuvers may serve as a strong indicator of angle and left-turn crashes—insights that cannot be obtained from existing aggregated analyses. Therefore, there is still a research gap on capturing risky lateral left and right turning behaviors and examining their impacts on intersection crash frequencies.

2.2. Intersection crash frequency spatial modeling

Spatial heterogeneity has emerged as a critical concern in intersection crash analysis, reflecting how the effect of the same factor on crashes varies across different spatial contexts (e.g., intersection in city, urban, and rural areas). To address such issue, spatial statistical methods such as spatial-lag models (Cui et al., 2024; Hong et al., 2016), Bayesian spatial models (Wang et al., 2023), and geographically weighted regression (GWR) (Brunsdon et al., 1998; Tang et al., 2020) have been developed to capture complex spatial heterogeneity. However, as these models fundamentally rely on linear assumptions, they struggle to estimate the inherent nonlinear relationships between the outcome variable and contributing factors (Han et al., 2024a; Wu et al., 2024).

Table 1
Existing intersection crash studies with CV-based driving behavior features.

Studies	CV datasets	Driving behavior features	Crash types
Kamrani et al. 2017	SPMD	Longitudinal accelerations	Rear-end
Wali et al. 2018	SPMD	Longitudinal acceleration/ jerk, standard deviation of speed	All crashes
Arvin et al. 2019	SPMD	Longitudinal & lateral acceleration/jerk, standard deviation of speed	Rear-end, head-on
Hunter et al. 2021	Wejo	Hard braking events	Rear-end
Mohammadnazar et al. 2022	SPMD	Longitudinal & lateral acceleration, standard deviation of speed	All crashes
Gu et al. 2023	SPMD	Longitudinal & lateral acceleration, yaw rate, standard deviation of speed	Rear-end
Current study	Streetlight	Longitudinal: Driving-straight speed, acceleration, deceleration Lateral: hard left and right-turn, radial acceleration	All crashes

Moreover, these models discard highly correlated factors to prevent overfitting, which could lead to the loss of potential critical information within high-dimensional datasets (Zhai et al., 2025; Zhao et al., 2024).

To address the limitations of statistical models, notable efforts have been made to integrate spatial heterogeneity effects into ML in recent safety studies as summarized in Table 2. Among them, the Geographical Random Forest (GRF) model proposed by Georganos et al. (2021) is the most popular one, which establishes local RFs for each sample and its spatial neighbors to account for spatial heterogeneity. Several studies have utilized GRF in recent traffic safety modeling and achieved better predictive accuracy and interpretability (Wu et al. 2024; Wang et al. 2024). However, RF may underperform on high-dimensional dataset (Do et al., 2010; Nguyen et al., 2015; Qu et al., 2019), which limits the predictive performance of GRF. Although a few studies have attempted to integrate neural networks into spatial modeling, these approaches also face challenges such as high computational cost and unstable training on small-size data (Goel et al., 2023). In contrast, boosting-based models (e.g., XGBoost and LightGBM) excel at nonlinear fitting for high-dimensional data and demonstrate high computational efficiency. Therefore, they have been widely utilized and achieved strong performance in recent traffic research (Li et al., 2023; Yang et al., 2021). Nevertheless, the potential of combining such ML models with spatial modeling has not been explored (Sigrist, 2023). It is worthwhile investigating such novel spatial ML approaches to account for both spatial heterogeneity and nonlinear effects in intersection crash frequency modeling.

3. Data preparation

In this study, we initially identified 612 signalized intersections at Hillsborough County using the FDOT Geographic Information System.[1] After excluding 111 sites with missing data (e.g., lacking traffic AADT data or census-tract socioeconomic data), a total of 501 intersections were finally selected. For each intersection, five kinds of macro-level features were extracted: (a) traffic volume, (b) intersection geometric design, (c) socioeconomic characteristics, (d) road context classifications, and (e) visual environment features. In terms of micro-level factors, multiple driving behavior features were extracted from the high-resolution CV data. The collective time frame of macro-level features is fixed for 2024 to match the duration of the CV data (Jan 3–13 and Jan 30-Feb 8, 2024). To mitigate the randomness of short-time crash recording, three-year crash data (June 2021 to May 2024) are utilized to better reflect the safety level at intersections (Desai et al., 2021; Hunter et al., 2021; Wali et al., 2018). The details of each data processing are elaborated on in the following sections.

Table 2
Existing spatial ML methods in traffic safety studies.

Model	Model innovations	Potential limitations
Geographical RF (Georganos et al., 2021; Gu et al., 2023; Wu et al., 2024)	Develop a global RF model for all samples and multiple local RF models for each sample with its spatial neighbors to capture spatial heterogeneity	RF models may underperform on high-dimensional dataset
XGBoost with geographic coordinates (Zhao et al., 2025)	Incorporate geographic coordinates (latitude and longitude) as inputs into the XGBoost model	Cannot capture spatial dependencies among input factors
Geographically Weighted Neural Network (GWNN) (Zhang et al., 2024)	Establish local Neural Network (NN) for each sample and its spatial neighbors	NN training needs longer computation time and may be unstable on small datasets
Geographically weighted convolutions neural network regression (GWCNNR) (Li et al., 2025)	Estimate spatial weights using CNN and then multiplied into the regression coefficients in the ordinary linear regression	Separately fit spatial and non-spatial features may ignore their interactions

[1] https://www.fdot.gov/statistics/gis/default.shtm.

3.1. Intersection crash identification

The crash data were obtained from the Florida Signal Four Analytics (S4A) system.[2] Each record includes precise crash time, location, collision type and severity, the number of vehicles involved, and other pertinent details. According to the S4A system, crashes that occur within 250 ft of the stop line are defined as "intersection-related crashes". Thus, within-intersection and intersection-related crashes were first identified, encompassing various vehicle-vehicle crash types (e.g., rear-end, sideswipe, left/right-turn, etc.). Referring to existing studies (Avelar et al., 2015; Kabir et al., 2021), an 80-ft spatial buffer around the centerline of intersection approaches was also created to ensure that the matched crashes did not occur in surrounding buildings or parking lots. Finally, a total of 24,707 intersection crashes were identified, and some examples are presented in Fig. 2.

3.2. Macro-level features matching

(1) Traffic volume and intersection geometric design.

From the FDOT Geographic Information System,[3] the average annual daily traffic (AADT) of all vehicles and large vehicles (e.g., truck, bus) on intersection major and minor roads were calculated. These features reflect intersection traffic volume as crash exposure as well as considering the impact of large vehicles. The geometric design of intersection approaches can also be obtained from the RCI system. For example, the posted speed limit is available for each roadway, allowing this information to be matched with the major and minor approaches at the intersections. Finally, a total of 26 geometric design features were extracted as shown in Table A1.

(2) Socioeconomic Data.

Socioeconomic data reflect the regional economic and demographic characteristics surrounding the intersection. To extract such features, census-tract-level socioeconomic features (e.g., average median income, population) were derived from the USDOT Equitable Transportation Community (ETC) Project.[4] Since one intersection may be near multiple census tracts, a 0.5-mile buffer (Avelar et al., 2015; Cai et al., 2018) was created around each intersection. Socioeconomic features from the census tracts that spatially overlapped with this buffer were aggregated to each intersections. A weighted average was utilized as suggested by existing studies (Huang et al., 2017; Pulugurtha and Sambhara, 2011). As an example, the average population variable E_i for the intersection buffer i can be calculated:

$$E_i = \sum_j \frac{A_{j,i}}{A_j} * E_j \quad (1)$$

where E_j is the population of census tract j, $A_{j,i}$ is the area of census tract j within buffer i, and A_j is the area of the census tract j.

(3) Road context classifications.

Considering that traffic patterns vary significantly between urban, suburban, and rural areas, the context classifications of roadways were obtained from the Florida Land Use & Infrastructure Plan.[5] Based on the

[2] https://signal4analytics.com.
[3] https://www.fdot.gov/statistics/gis/default.shtm.
[4] https://www.transportation.gov/priorities/equity/justice40/etc-explorer.
[5] https://hcfl.gov/government/county-projects/land-use-and-infrastructure-studies/land-use-and-infrastructure-other-publications.

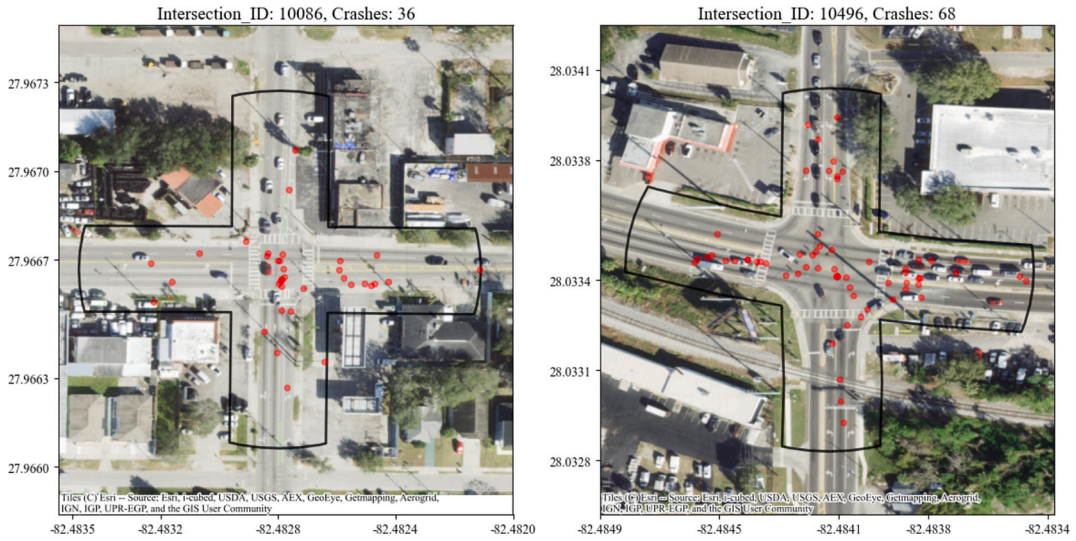


Fig. 2. Examples of crashes in the intersection buffers.

surrounding land use, roadways were divided into six categories (e.g., C3C for suburban commercial areas, C4 for urban areas, as shown in Table A1). Recent research indicates that the traffic patterns and safety are quite different among these roadway classifications (Al-Omari et al., 2021; Mahmoud et al., 2021). Therefore, it is essential to combine such features in intersection crash frequency modeling.

(4) Visual Environment Features.

Typically, the proportions of environmental objects in street-view images are utilized to reflect drivers' visual perception of their surroundings (e.g., grass indicating open space, vehicles representing traffic density, etc.). Using Google Maps API, the center point was identified as the viewpoint origin of street view images for each intersection. Following existing studies (Cai et al., 2022; Xue et al., 2024; Yue, 2024), 8 Google Street View (GSV) images were obtained at headings ranging from 0° (north) to 315°(northwest) to fully capture the entire intersection environment (Fig. 3**(a)**). It is noted that intersections undergoing construction were excluded to ensure a stable environment during the study period. Images at intersection center offer a more comprehensive view, whereas perspectives from individual travel directions often face obstructions from vehicles blocking other elements (e.g., buildings and trees). Finally, a total of 8*501 = 4008 GSV images were collected. To qualify for visual environment features in GSV images, Segmenter, a transformer-based segmentation model, was employed to pixel-level objects classification (over 90 % reported classification accuracy). A total of 7 types of objects in GSV images were labeled as shown in Fig. 3**(b)** (e.g., sidewalk, grass, vegetation, road, building, sky, and vehicle). Other types of elements (e.g., pedestrian, bicycle, animals, etc.) are excluded from the analysis, as they typically occupy a small pixel proportion and are often missed in most samples as suggested by Liu et al. (2025). Referring to existing studies (Abdel-Aty et al., 2024; Cai et al., 2022; Fan et al., 2023b), the pixel proportions of each object were calculated as visual environment features. Overall, detailed descriptive statistics of these macro-level features are depicted in Table A1.

3.3. Intersection driving behavior features extraction

In this study, the CV data are provided by StreetLight. It contains 3-second-interval vehicle trajectories collected from original equipment manufacturers (OEMs) using vehicle-to-cloud communication. Given that the fleet is composed of multiple types of non-commercial vehicles, it can better represent the vehicles on the roadways (Zhang and Abdel-Aty, 2022). The data includes journey ID, capture time, GPS location, heading, and speed as described in Table 3. The data spans two distinct periods: Jan 3–13 and Jan 30-Feb 8. On average, it includes over 4,692,975 CV trajectory points per day, derived from 154,997 journeys, providing full coverage of Hillsborough County with high market penetration of 4.17 %, as shown in Fig. 4.

3.3.1. Spatial matching of CV data with intersections

Fig. 5**(a)** presents the spatial matching workflow of CV trajectory data with intersections. To ensure data quality, erroneous CV data were filtered out following the criteria of speed $>$ 100 mph or updated time interval $>$ 5 s. Using the intersection and road GIS data, a joined spatial buffer was created consistent with those used in the crash matching: a 250-ft buffer was applied to identify CV points within the intersections and their approaches. An 80-ft buffer from the centerline of intersection roads was generated to exclude CV data not located on the roads (i.e., gray points outside the buffer are in a nearby parking lot in Fig. 5**(b)**). Through the spatial joining between the CV data and intersection spatial buffer, the final spatially matched CV data for each intersection were obtained.

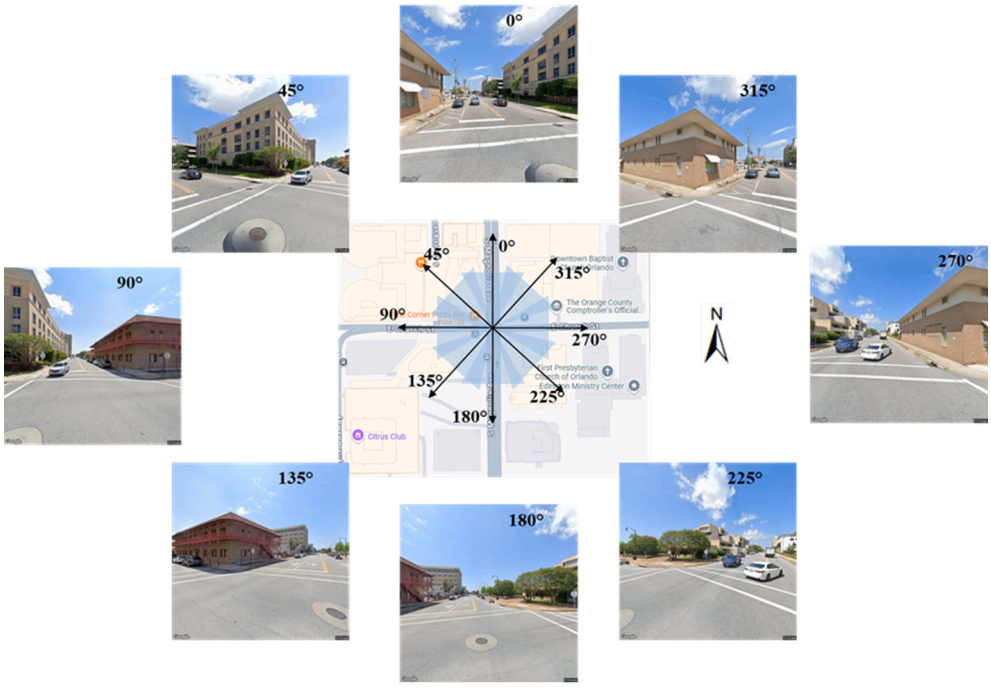
3.3.2. Identification of longitudinal and lateral risky driving behaviors

Risky driving behaviors refer to instantaneous driving actions involving abnormal vehicle operations (e.g., extreme acceleration or braking), which may lead to heavy surrounding traffic volatility and even crashes (Wali et al., 2018). Previous studies have identified risky longitudinal behaviors (e.g., hard acceleration and braking) based on the longitudinal/linear acceleration (Gu et al., 2023; Hunter et al., 2021). However, within the intersection area, drivers always encounter hard left turns, right turns and other lateral risky driving behaviors when navigating to another roadway direction. Therefore, based on the CV trajectories, we calculated both linear and radial accelerations, and thus identified risky longitudinal driving behaviors (hard acceleration and braking) as well as lateral driving behaviors (hard right turn and left turn).

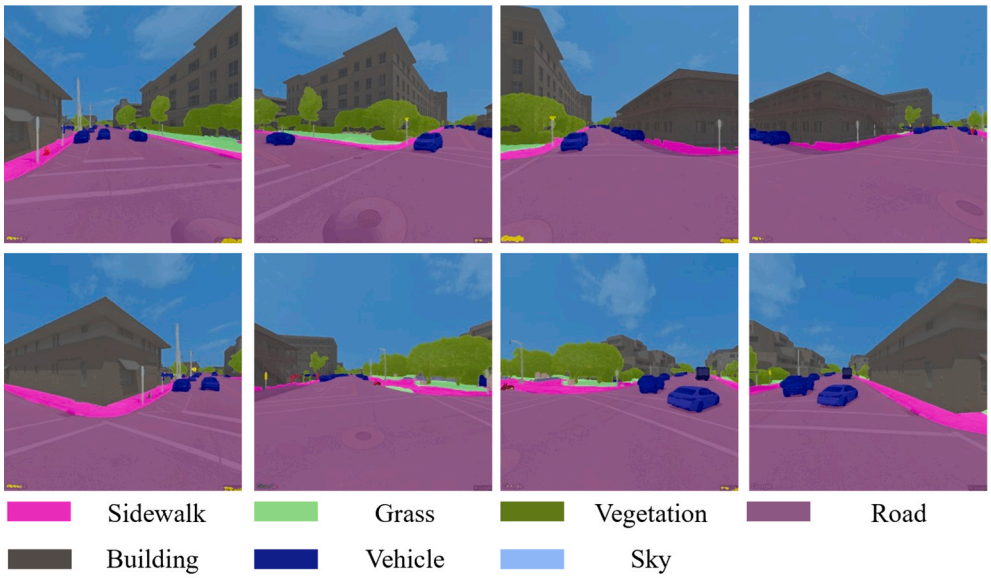
(1) Linear and radial acceleration calculations.

Fig. 6**(a)** shows the calculation of both linear and radial acceleration based on CV point. Mathematically, a CV trajectory point can be written as $P_t = (X_t, V_t, \theta_t)$, where $X_t = (lat_t, lon_t)$ is the GPS location, V_t is the speed, and θ_t is the travel heading at time t. Therefore, consider two CV points during turning maneuver: the earlier point $P_{t0} = (X_{t0}, V_{t0}, \theta_{t0})$ at time $t0$ and the subsequent point $P_{t1} = (X_{t1}, V_{t1}, \theta_{t1})$ at time $t1$. The linear acceleration can be easily calculated:

$$LinearACC_{t1} = \frac{\Delta V}{\Delta t} = \frac{V_{t1} - V_{t0}}{t1 - t0} \quad (2)$$



(a) Google street view images at one intersection



(b) Semantic segmentation results

Fig. 3. The illustration of visual environment features extraction from GSV images.

Table 3
Description of the raw CV data parameters.

Parameters	Description	Unit
Journey ID	Unique identifier for a trip (from ignition start to end).	–
Capture time	10-digit UTC timestamp	s
Latitude	North-South position of the vehicle	–
Longitude	East-West position of the vehicle	–
Heading	The heading of the vehicle travel. (e.g., 0: North; 90: East)	°
Speed	Speed of the vehicle at the instant the datapoint was captured	mph

As for the radial acceleration (a.k.a centripetal acceleration), it describes the acceleration of an object moving along a curved path directed toward the center of its trajectory. Thus, we need to first determine the turning radius r:

$$r = \frac{d/2}{\sin(\Delta\theta/2)} where d = Euclide and istance(X_{t1}, X_{t0}), \Delta\theta = |\theta_{t1} - \theta_{t0}| \quad (3)$$

Then, the radial acceleration can be calculated:

$$RadialACC_{t1} = \frac{\overline{V}}{r^2} = \frac{(V_{t1} + V_{t0})/2}{r^2} \quad (4)$$

Based on the above formulas, both linear and radial acceleration can be derived from the raw CV trajectory data. For example, Fig. 6**(b)-(d)**

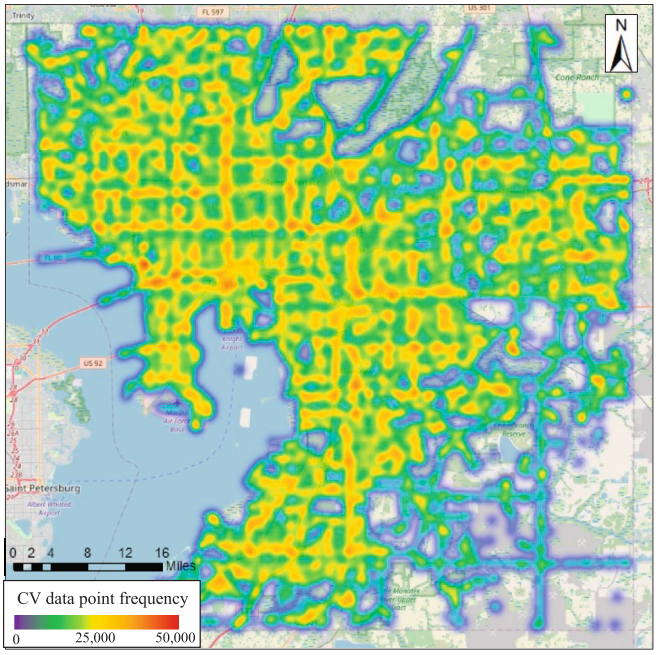


Fig. 4. CV data spatial distribution (January 30, 2024 as an example).

show the speed, linear and radial acceleration of an entire trip, respectively. Clearly, the linear acceleration can reflect the acceleration (blue points in Fig. 6**(c)**) and braking (red points in Fig. 6**(c)**) situations. While at intersections, the radial acceleration increases significantly to effectively capture the left and right turning behaviors.

(2) Risky driving behavior identification with dynamic thresholds.

Risky driving behaviors are often identified as extreme outliers in terms of acceleration. For instance, if the linear acceleration exceeds a certain threshold, it is classified as "hard acceleration". Most existing studies applied a fixed threshold across all speed conditions (Guo et al., 2022, 2021; Han et al., 2024b; Hunter et al., 2021). However, the capability of a vehicle to accelerate/decelerate varies significantly at different speeds, i.e., at higher speeds the possible minimum deceleration values are considerably smaller than those observed at lower speeds. To overcome such limitation, a dynamic threshold approach proposed by Wali et al. (2018) was adopted in this study. Specifically, a series of 5mph speed bins (i.e., 0-5mph, 5-10mph, …70-75mph) were first defined. Within each speed bin, the 3 σ rule was used to identify these extreme acceleration outliers by calculating the upper bound following existing studies (Gu et al., 2023; Wali et al., 2018):

$$ACC\_threshold_k = \mu_k + 3^*\sigma_k \tag{5}$$

where μ_k and σ_k represent the mean and the standard deviation of acceleration within the respective speed bin k. Therefore, it ensures that thresholds can be dynamically adjusted according to speed considering the vehicle performance. However, as some speed bins contained insufficient samples, the resulting thresholds were inconsistent (e.g., the threshold at the 5–10 mph bin was lower than that for 10–15 mph). To address such cases, we fixed the thresholds for speeds below 15 mph at 2.50 m/s^2 for hard acceleration and –2.75 m/s^2 for hard deceleration. For speeds above 60mph, we set the thresholds to 0.71 m/s^2 for hard acceleration and –0.64 m/s^2 for hard deceleration. For the radial acceleration, we set the thresholds to 3.40 m/s^2 for speed bin of 0–10 mph and 1.52 m/s^2 for speed bin > 35mph. Using this approach, risky driving behaviors at intersections can be identified as shown in Fig. 7. Fig. 7**(a)** shows the linear acceleration distribution across different speed bins, with two threshold curves: the positive curve identifies hard acceleration events, while the negative curve captures hard braking events. Since radial acceleration is inherently positive, a single threshold curve was established to identify hard turning, as shown in Fig. 7**(b)**. These turning events are further categorized into hard right turns and hard left turns in subsequent analyses.

3.3.3. Calculation of intersection driving behavior features

Considering different vehicle movements at intersections, CV trajectories were further categorized into three types: drive-straight, left-turn, and right-turn as shown in Fig. 8. Compared to the drive-straight maneuvers, lateral turning maneuvers exhibit distinct spatial patterns and movement interactions. For the right-turn maneuver (Fig. 8 (a)), they are predominantly distributed around the outer parts of intersections. In contrast, left-turn maneuvers occupy the inner intersection area, often overlapping with movements in other directions (Fig. 8 (b)). However, existing studies typically either solely rely on longitudinal driving features or utilize the aggregated speed or acceleration measures of the whole maneuvers, largely overlooking the distinctive driving behaviors and interactions with other vehicles in lateral turning maneuvers. To address this limitation, multiple driving behavior features were calculated based on the three movement categories—drive-straight, left-turn, and right-turn—to comprehensively capture risky driving behaviors.

(1) Speed volatility.

Speed volatility is widely used to measure the variations in instantaneous driving, which is highly associated with aggressive drivers and unsafe outcomes (e.g., crashes and conflicts) (Kamrani et al., 2018; Wali et al., 2018; Yu et al., 2021). Referring to existing studies (Wali et al., 2018), the standard deviation of speed was calculated to reflect the driving volatility at intersections. First, at the maneuver level, the speed standard deviation (STD_Speed_i) for the i-th maneuver can be calculated as follows:

$$STD\_Speed_i = \sqrt{\frac{\sum_t^T \left(V_{i,t} - \overline{V_{i,t}}\right)^2}{T}} \tag{6}$$

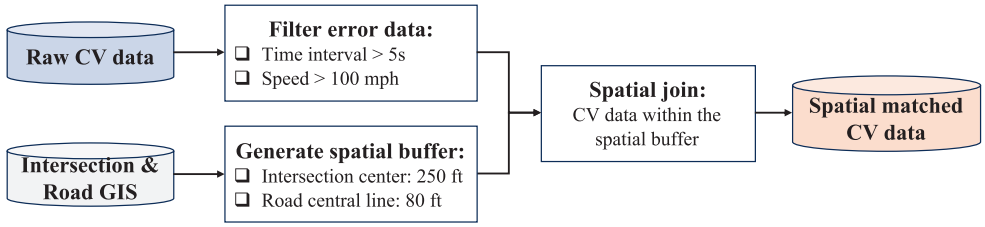
where $V_{i,t}$ and $\overline{V_{i,t}}$ are the speed at timepoint t and the mean speed over the i-th maneuver. T is the total number of timepoints recorded during the i-th maneuver at the intersection buffer. Here, stop points (speed = 0mph) were removed to eliminate the impact of signal-controlled stops on speed volatility (Wali et al., 2018). After obtaining the maneuver-level speed volatility, their mean and maximum values were further calculated to represent the average and extreme cases of speed volatility at the intersection level. Thus, for each kind of maneuver (i.e., drive-straight, left-turn, and right-turn), these speed volatility measures were separately aggregated to capture their driving volatility.

(2) Hard events.

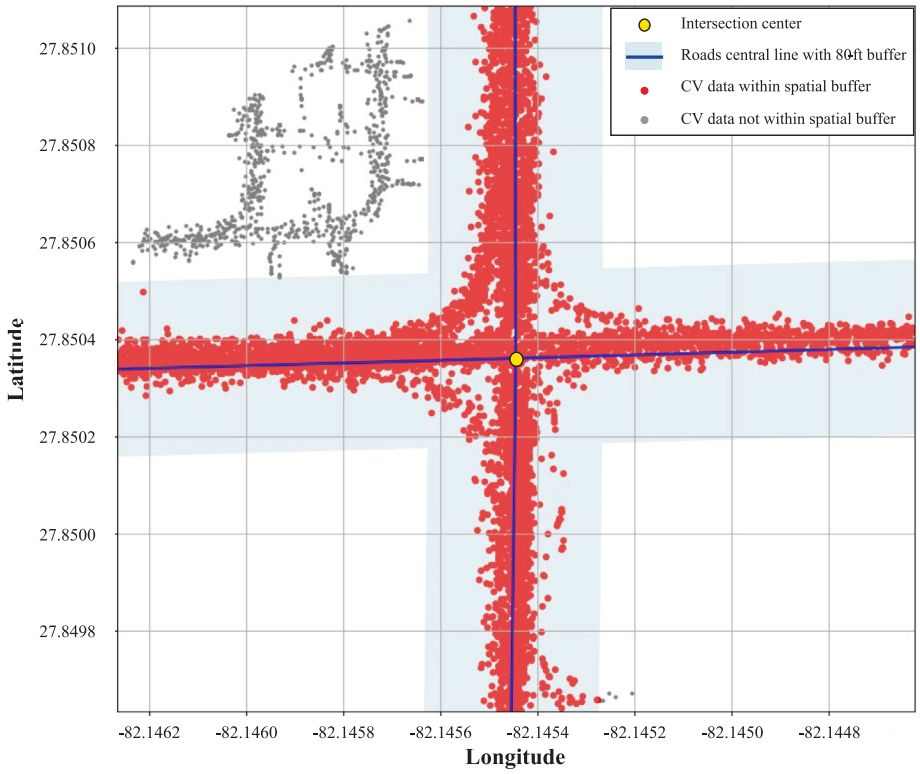
The count of hard events directly reflects the frequency of risky driving behaviors at intersections and has been widely used to reflect the potential risk level in recent traffic studies (Gu et al., 2023; Guo et al., 2021; Han et al., 2024b; Zhang and Abdel-Aty, 2022). Based on the identified risky driving behavior from the previous step, the counts of hard braking and hard acceleration events were considered for the drive-straight maneuvers. While for left-turn and right-turn maneuvers, the lateral hard turning was counted separately to represent hard left turns and hard right turns.

(3) Maneuver risky level: extreme acceleration

Although the count of hard events reflects the frequency of risky driving behaviors, it heavily relies on the specific cut-off threshold. More importantly, it loses the detailed acceleration information, making it fail to measure the risk or severity level of each driving behavior (Han et al., 2024b, 2024a; Kamrani et al., 2017). For instance, a behavior involving high acceleration poses a greater likelihood of causing crashes compared to one with low acceleration. To address this issue, the extreme acceleration observed in each maneuver was selected as an indicator of the severity of potential crash risk. Specifically, the highest acceleration and the lowest deceleration (negative acceleration) were extracted for all



(a) Data processing workflow



(b) Illustration of spatially matched CV data

Fig. 5. Spatial matching of CV data within intersections and road buffer.

drive-straight, left-turn, and right-turn maneuvers. While the extreme radial acceleration was only detected for left-turn and right-turn maneuver. These metrics were then aggregated to the intersection level using their mean, maximum, and sum values. It is noted that the sum of such extreme acceleration can be regarded as a risk accumulation index at an intersection, which offers two potential benefits. First, under stable penetration rates (3–5 %), it is proportional to the intersection traffic volume, serving as an indicator of exposure to crashes. Second, it can sensitively capture the risk level under similar traffic conditions, as it would significantly increase at intersections with frequent risky driving behaviors compared to those with little risky driving activities.

Finally, a total of 34 intersection driving behavior features were calculated from the CV trajectories as summarized in Table 4. Given that the CV data spans $D = 21$ days, these features were first calculated for each day, and their mean $\overline{F_j}$ was then computed to ensure a stable representation of the daily average intersection safety level:

$$\overline{F_j} = \frac{\sum_d^D F_{j,d}}{D} \tag{7}$$

where $F_{j,d}$ is the driving behavior features F (e.g., hard braking) on day d at intersection j. Detailed descriptive statistics of these micro-level features are depicted in Table A2.

4. Methodology

4.1. Spatial machine learning framework

In the traditional GWR framework, model coefficient estimates are allowed to vary across locations in order to address spatial heterogeneity. A simplified representation is:

$$Y_i = \boldsymbol{\beta}_{(u_i,v_i)}(\boldsymbol{X}_i) + \varepsilon_i, i = 1:n \tag{8-1}$$

$$\boldsymbol{\beta}_{(u_i,v_i)}(\boldsymbol{X}_i) = \beta_0(u_i,v_i) + \sum_{k=1}^{p}\beta_k(u_i,v_i)x_{ik} \tag{8-2}$$

where Y_i is the dependent variable for the i th observation; (u_i, v_i) represents the spatial coordinate (e.g., longitude and latitude) of sample i; $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ is its vector of independent variables. $\boldsymbol{\beta}_{(u_i,v_i)} = [\beta_0(u_i,v_i), \beta_1(u_i,v_i), \cdots \beta_p(u_i,v_i)]^T$ is the vector of location-specific linear parameters, including a local intercept $\beta_0(u_i,v_i)$ and coefficients $\beta_k(u_i,v_i)$ that vary with (u_i,v_i) to capture local spatial effects; ε_i is the random error term. Clearly, GWR relies on the linear function given by Equation (8)–(2) and therefore cannot estimate nonlinear relationships and struggle to deal with high-dimensional inputs (Deng et al., 2020; Georganos et al., 2021).

To address this issue, spatial ML framework is proposed to replace

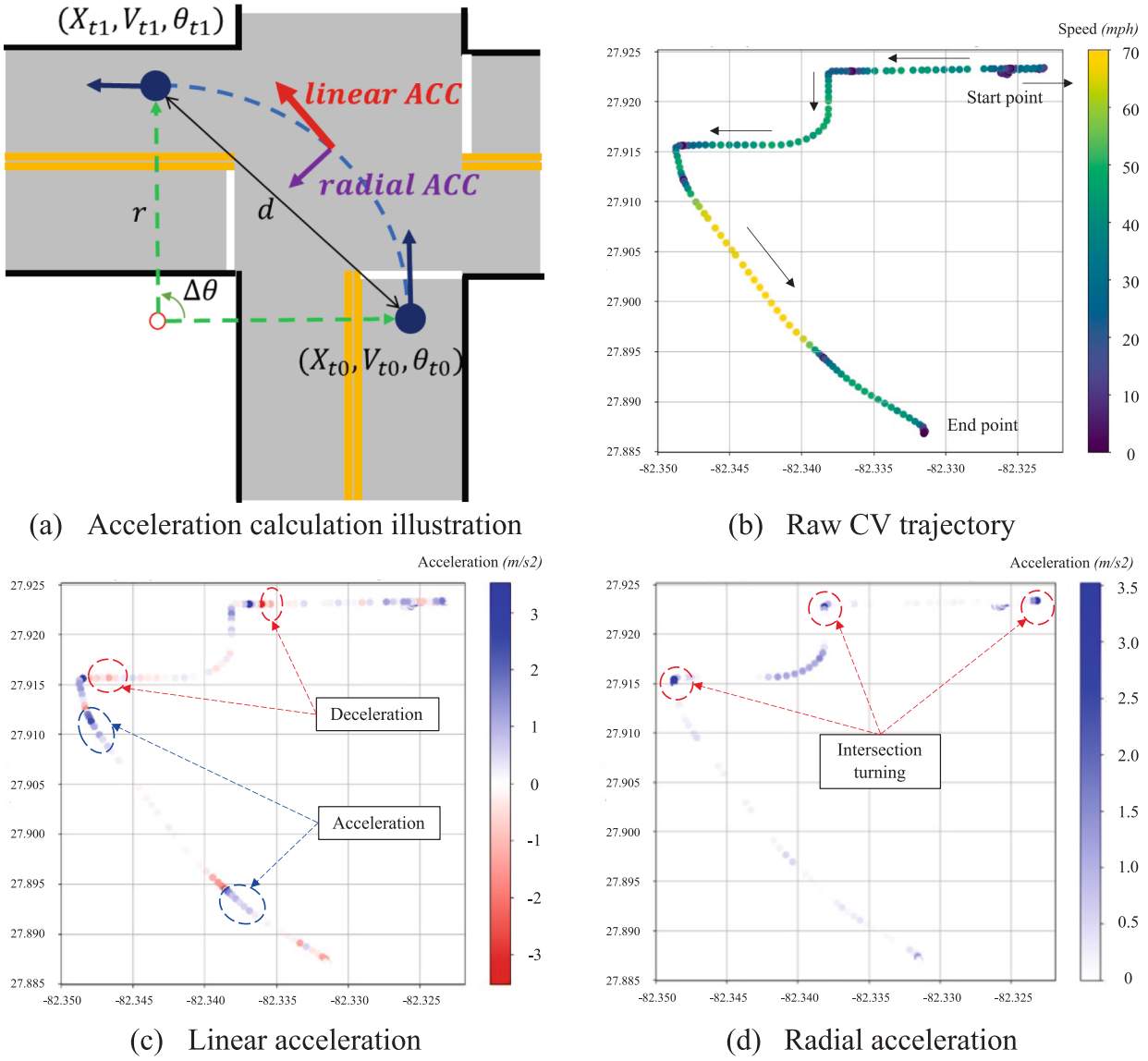


Fig. 6. Linear and radial acceleration calculation.

the GWR linear function to a nonlinear ML kernel:

$$Y_i = ML_{(u_i,v_i)}(\boldsymbol{X}_i) + \varepsilon_i, i = 1:n \tag{9}$$

Where $ML_{(u_i,v_i)}$ denotes a nonlinear ML model with varying parameters at each spatial coordinate (u_i, v_i) to capture the local spatial effects. In theory, $ML_{(u_i,v_i)}$ can be instantiated by any ML algorithm[6] without requiring an explicit formula. For example, if the ML kernel is a random forest, the framework becomes the GRF model (Georganos et al., 2021). Similarly, it can easily incorporate other ML models, such as the boosting-based XGBoost and LightGBM, which remain unexplored for their spatial modeling performance. To fill this gap, the proposed spatial ML framework integrates these different ML methods (e.g., RF, XGBoost, and LightGBM) as basic models within a geographically weighted framework. This flexibility enables researchers to select the most effective learner for their data.

To implement this spatial ML framework for modeling datasets, the processes of model fitting, testing, and interpretation are thoroughly described in the following sections.

4.1.1. *Spatial ML fitting on train dataset*

Fig. 9 illustrates the model fitting framework for integrating the ML models within the GWR framework. In the spatial ML framework, a global ML model is first trained using all samples to capture an unbiased estimate of the average relationship for the entire region. However, such relationships may vary across different spatial locations, forming the spatial heterogeneity (Brunsdon et al., 1998; Deng et al., 2020). To handle such issue, we assume that nearby samples are more likely to share similar patterns according to the Tobler's First Law of Geography (Fotheringham et al., 2017). Consequently, for each spatial sample $\boldsymbol{X}_i$, a local ML is further trained which only includes its nearby observations within a specified distance (defined via bandwidth λ) to capture the localized variations in the relationships. Finally, the prediction for sample $\boldsymbol{X}_i$ is obtained as a weighted combination of the predictions from both global and local models:

$$\hat{y}_i = \alpha^* \hat{y}_{li} + (1-\alpha)^* \hat{y}_{gi} \tag{10}$$

where $\hat{y}_i$ is the final prediction, $\hat{y}_{li}$ and $\hat{y}_{gi}$ are the predictions of the local and global ML models, respectively. α is the local weight hyper-

[6] The code is available at GitHub: https://github.com/UCFLeiHan/Spatial-ML.

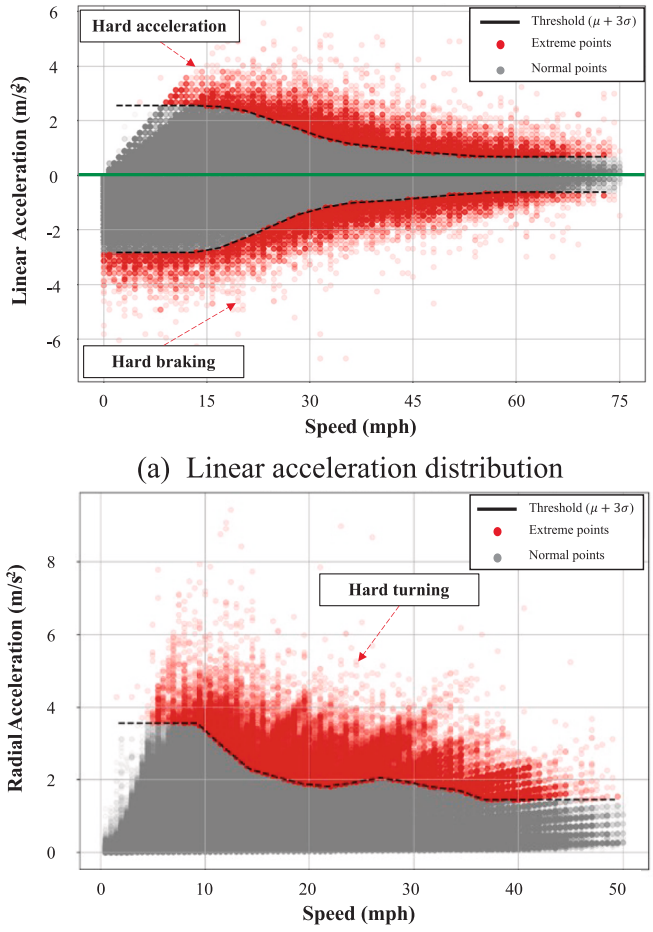


Fig. 7. Risky driving behavior identification with dynamic acceleration thresholds.

parameter whose value in the range $[0,1]$. Our preliminary experiments reveal that relying solely on local models often results in overfitting specific training samples, thereby limiting their ability to generalize to unseen data. The global model is necessary to capture overall unbiased coefficients to overcome this issue (more detail can be found in Sun et al., 2024). Overall, this approach enhances the ML models to be calibrated locally rather than globally, thus effectively capturing nonlinearity and simultaneously revealing spatial variations. As a result, it can improve both the model predictive power and its ability to provide localized explanations (Wu et al., 2024; Zhang et al., 2024).

Two hyperparameters, bandwidth λ and local weight α, largely impact model performance. For the λ, there are two types of bandwidths: "fixed kernel" and "adaptive kernel" as shown in Fig. 10. In the fixed kernel model, a bandwidth of N miles means selecting all neighbors located within N miles of the target intersection. In the adaptive kernel model, a bandwidth of N indicates that the N closest intersections are used in fitting the local model. The limitation of the fixed kernel is that it cannot capture enough samples when intersections are sparsely distributed (e.g., suburban and rural areas), while adaptive kernel can ensure a certain number of observations for fitting the local model. Therefore, existing spatial models commonly adopted the adaptive kernel to train the local models (e.g., Geographically weighted regression and Geographically weighted random forests in Gu et al., 2023; Wu et al., 2024). For this reason, we adopted the adaptive kernel during spatial modeling. Based on the finding of Sun et al. (2024), the optimal bandwidth can be determined though an incremental spatial autocorrelation test: the distance at which the z-score is the highest is used as bandwidth λ, and the global Moran's I index at that distance is local weight α:

$$\alpha = \begin{cases} Moran's\ I, & if\ Moran's\ I > 0,\ and\ p < 0.05 \\ 0, & otherwise \end{cases} \tag{11}$$

Specifically, **Algorithm 1** provides the model fitting procedure of the spatial ML model. Given the training set Ω_{train} and a ML algorithm (e.g., RF/XGBoost/LightGBM/…), a global ML is first fitted on all samples in Ω_{train}. Based on the spatial autocorrelation test on the targets $(y_1, y_2, \cdots, y_n)$, the optimal bandwidth λ and local weight α can be obtained. Then for each sample $(\boldsymbol{X}_i, y_i)$, the neighboring samples within bandwidth λ would be chosen to train a local model (ML_i). It is noted that during this local training process, each neighbor sample is assigned a spatial weight based on its distance to $\boldsymbol{X}_i$ (d_{ij}), so that closer samples can show stronger spatial influences on the local model. Here, the loss function $\mathscr{L}()$ of ML is chosen as to mean squared error loss. Once the local ML is fitted, the prediction of training sample $\widehat{y}_i$ can be calculated by combining the predictions from its local ML ($\widehat{y}_{li}$) and global ML ($\widehat{y}_{gi}$). Finally, the spatial ML framework consists of one global ML model (ML_{global}) and n local ML models ($ML_{local} = \{ML_i\}_i^n$). The predicted targets $\widehat{y}_{train}$ can be used for model hyperparameter tuning to get the best spatial ML.

Algorithm 1: Pseudo algorithm of spatial ML fitting

Input:

Ω_{train}: a set of n training samples including $\{(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \cdots, (\boldsymbol{X}_n, y_n)\}$; ML: a pre-initialized ML model (e.g., RF, XGBoost, LightGBM)

1: $ML_{global} \leftarrow$ train global ML given Ω_{train}
2: Initialize $ML_{local} \leftarrow []_n$, $\widehat{y}_{train} \leftarrow []_n$
3: Calculate λ and $\alpha \leftarrow$ spatial autocorrelation test on $(y_1, y_2, \cdots, y_n)$
4: **for** each $(\boldsymbol{X}_i, y_i)$ in Ω_{train} **do**
5: Calculate distance $d_{ij} \leftarrow (\boldsymbol{X}_i, \boldsymbol{X}_j)$
6: $\Omega_i \leftarrow \{j | d_{ij} < \lambda\}$ // select neighbors of $\boldsymbol{X}_i$ within bandwidth λ
7: $w_{ij} = \left(1 - \left(\frac{d_{ij}}{\lambda}\right)^2\right)^2$ // calculate training weight of each neighbor $j \in \Omega_i$
8: $ML_i = \underset{\theta}{argmin} \sum_{j \in \Omega_i} w_{ij} \mathscr{L}\left(\boldsymbol{X}_j, \theta, y_j\right)$ // train local ML using only neighbors $j \in \Omega_i$
9: $\widehat{y}_{li} \leftarrow ML_i(\boldsymbol{X}_i)$, $\widehat{y}_{gi} \leftarrow ML_{global}(\boldsymbol{X}_i)$
10: $\widehat{y}_i = \alpha * \widehat{y}_{li} + (1-\alpha) * \widehat{y}_{gi}$
11: ML_{local}. append(ML_i), $\widehat{y}_{train}$. append($\widehat{y}_i$)
12: **end for**

Output:

ML_{global}: the global ML model; ML_{local}: a set of n local ML models; $\widehat{y}_{train}$: a set of n predicted targets

4.1.2. Spatial-weighted prediction on test dataset

The well-trained spatial ML model can be used to make predictions for unseen test datasets. In existing spatial studies (Georganos et al., 2021; Gu et al., 2023; Wu et al., 2024), the local prediction for unseen test data is typically made using the closest local ML model. However, possible outliers within the closest local model can degrade its performance. Meanwhile, other nearby ML models also offer valuable insights for local prediction, yet are still overlooked in these spatial models. To address these issues, a spatial-weighted ensemble prediction was utilized as shown in Fig. 11**(a)**. Instead of relying solely on the single closest local ML model, this approach incorporates all local ML models within the specified bandwidth, combining their predictions in a spatially weighted manner:

$$\widehat{y}_{lj} = \frac{\sum_{k \in \Omega_j} w_{kj} * ML_k\left(\boldsymbol{X}_j\right)}{\sum_{k \in \Omega_j} w_{kj}} \tag{12}$$

where $\widehat{y}_{lj}$ is the ensemble local prediction for the j th test sample $\boldsymbol{X}_j$, ML_k are the k th nearby local ML models within the bandwidth λ; w_{kj} is the 'bisquare' kernel spatial weight determined by the distance between locations k and j as shown in Fig. 11**(b)** (Deng et al., 2020; Fotheringham et al., 2017):

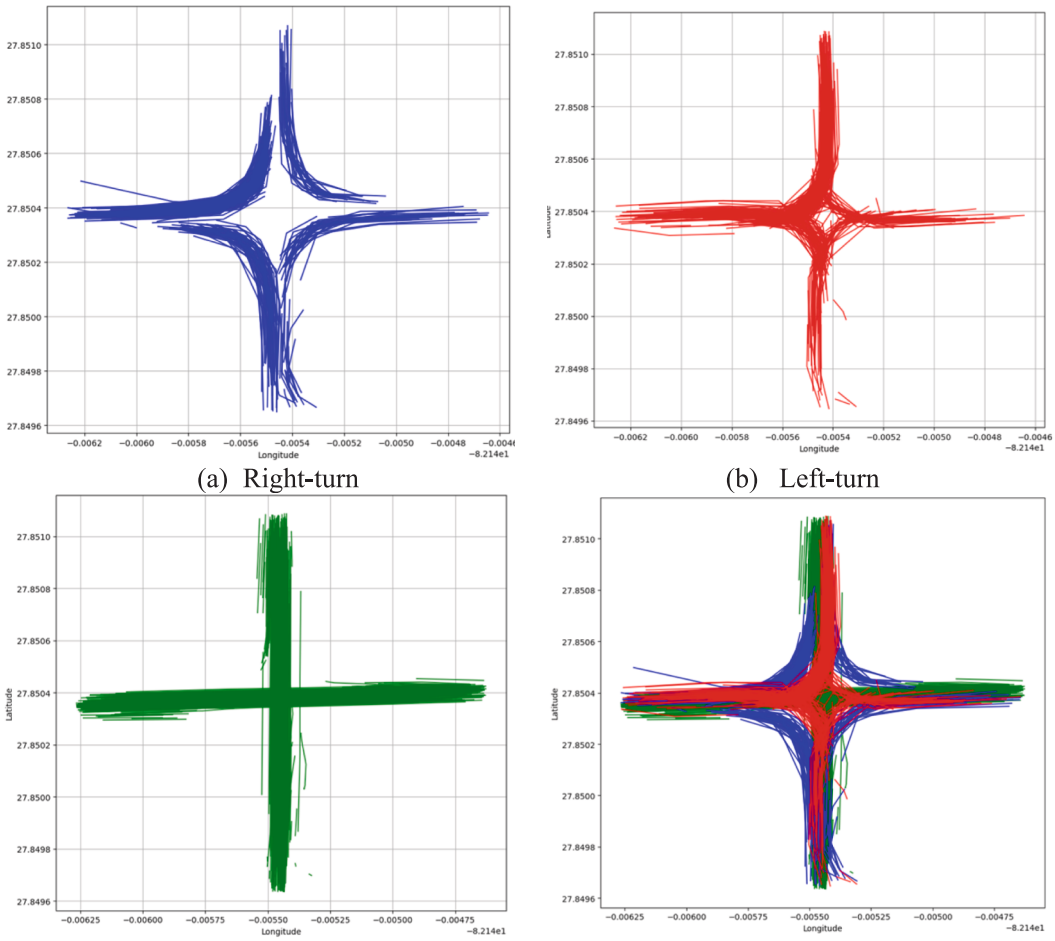


Fig. 8. Intersection CV trajectories by three movement types.

Table 4
34 intersection driving behavior features.

Driving behavior features		Maneuver categories		
		Drive-straight	Left turn	Right-turn
Speed volatility		Straight/left/right_speed_std_mean Straight/left/right_speed_std_max		
Hard events		Hard braking/ acceleration	Hard left turn	Hard right turn
Maneuver risky level: extreme acceleration	**Acceleration**	Straight/left/right_acc_mean Straight/left/right_acc_max Straight/left/right_acc_sum		
	Deceleration	Straight/left/right_dec_mean Straight/left/right_dec_max Straight/left/right_dec_sum		
	Radial acceleration	–	Left/ right_radial_acc_mean Left/ right_radial_acc_max Left/ right_radial_acc_sum	

$$w_{kj} = \left(1 - \left(\frac{d_{kj}}{\lambda}\right)^2\right)^2, d_{kj} \le \lambda \tag{13}$$

Therefore, the predictions from closer local ML models are assigned higher weights than those from farther away local ML models. Since this spatially weighted local prediction combines the predictions from all local ML models within the bandwidth, it is more robust and less susceptible to data outliers affecting a single local ML model.

4.1.3. Model interpretation with SHapley Additive exPlanations (SHAP)

Although Spatial ML can provide feature importances at both global and local levels, it cannot determine whether the impact is positive or negative. Therefore, we introduce the widely used SHAP framework to both global and local MLs to quantify how each feature contributes to the targets. The SHAP, proposed by Lundberg and Lee (2017), aims to describe the performance of a machine learning model based on game theory (Štrumbelj and Kononenko, 2014) and local explanations (Ribeiro et al., 2016). It offers an easy and effective measure to estimate the feature contributions and has been widely utilized in machine learning interpretation (Han et al., 2024a; Yu et al., 2024). Assume a ML model where a group F (with n features) is used to predict an output. In SHAP, the contribution of each feature to the model output $f(F)$ is allocated based on its marginal contribution (Lundberg and Lee, 2017). The SHAP value $\varnothing_i$ of the i th feature is calculated through:

$$\varnothing_i = \sum S \subseteq F\{i\} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \tag{14}$$

where S represents all feature subsets from F after removing the i th feature. $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ represents the probability weight of S calculated after feature permutation and combination. $f_{S\cup\{i\}}$ and f_S represent the model predictions with and without the i th feature, respectively, and x_S represents the values of the input features in the set S.

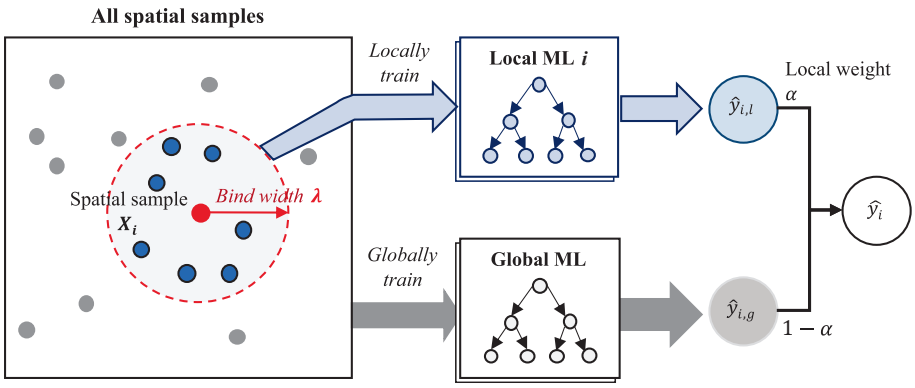


Fig. 9. The framework of the spatial ML fitting.

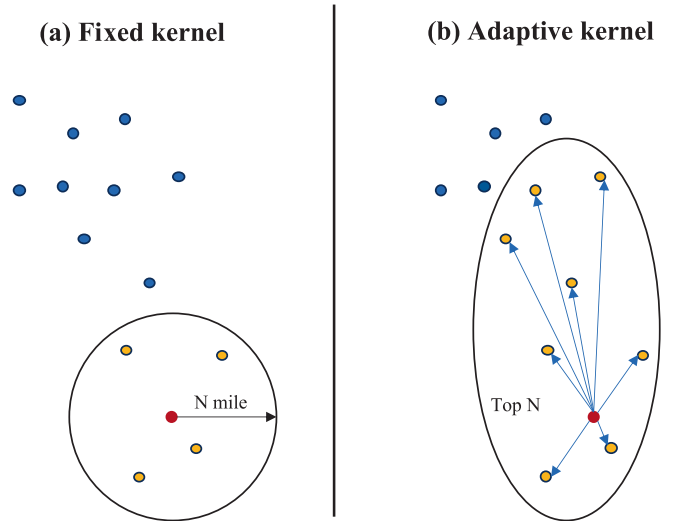


Fig. 10. Two types of adaptive kernals for bandwidth.

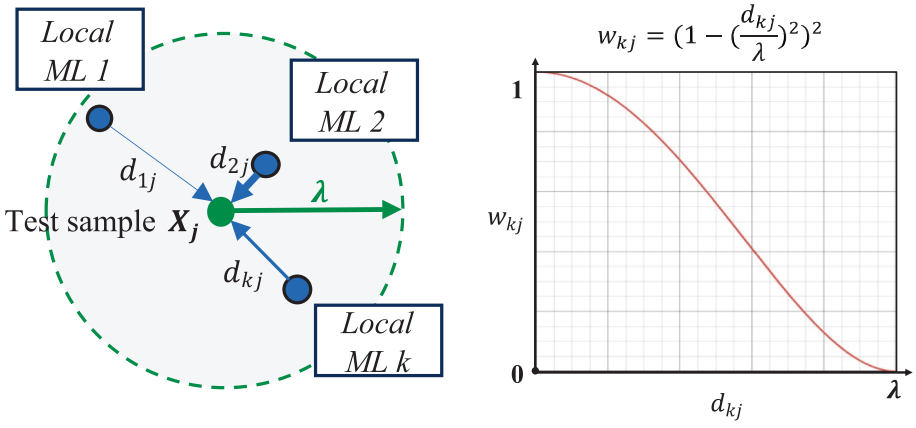
4.2. Baselines and model evaluation metrics

To evaluate the model performance of our proposed spatial ML, we selected five widely used models in existing studies as baselines: Geographically Weighted Regression (GWR), RF, XGBoost, LightGBM, and Multilayer Perceptron network (MLP). For these ML models, their hyperparameters are tuned using random grid search to get the optimal values as shown in Table 5.

To evaluate the prediction performance of candidate models, three measures include the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2. For these metrics, lower RMAE and MSE and higher R^2 indicate better model prediction performance.

Table 5
The optimal model hyperparameters setting.

Models	Hyperparameters	Tuning range	Selected value
RF	Number of estimators	20, 50, 100, 200	100
	Maximum tree depth	10, 20, 30	20
	Maximum features	'sqrt', 'log2','1/3'	'1/3'
XGboost, LightGBM	Objective function	'squarederror', 'squaredlogerror'	'squarederror'
	Learning rate	0.2, 0.1, 0.01, 0.001	0.1
	Number of estimators	20, 50, 100, 200	50
	Maximum tree depth	10, 20, 30	20
MLP	Layer number	1, 2, 4, 6	2
	Nodes of 1st, 2nd layer	16, 32, 64, 128	32 & 16
	Loss function	MSE, MAE, HuberLoss	MSE
	Learning rate	5e-3, 1e-3, 5e-4, 1e-4	1e-4
	Batch size	32, 64, 128, 256	64



(a) Spatial-weighted ensemble prediction (b) Function curve of bi-square function

Fig. 11. Spatial-weighted prediction on test dataset.

5. Results

5.1. Intersection crash spatial autocorrelation test

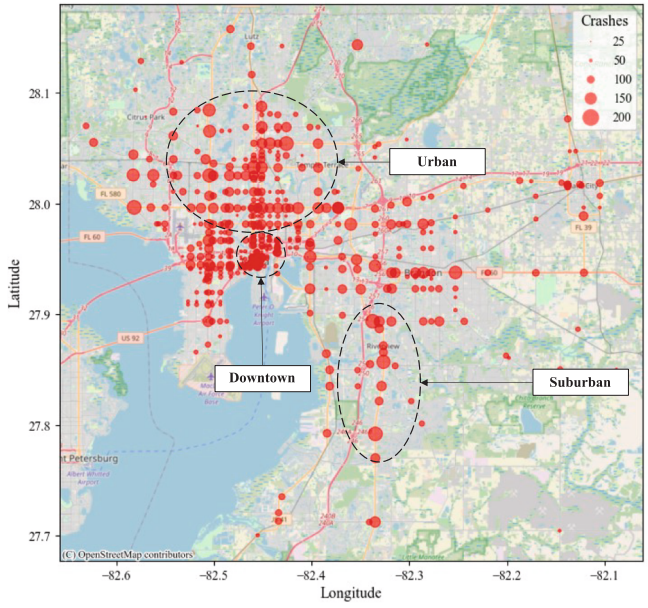
The prerequisite for spatial modeling is to confirm the presence of spatial autocorrelation. Fig. 12**(a)** shows the spatial distribution of intersection crashes, revealing clear spatial aggregation patterns in downtown, urban, and suburban areas. Also, similar distribution trends are observed along certain high-volume arterial roads. However, high-crash intersections are primarily distributed in downtown and urban areas with heavy traffic and dense human activity. In contrast, suburban intersections exhibit lower crash frequencies and are more sparsely distributed, demonstrating the complex spatial heterogeneity. To further quantify their spatial autocorrelation and heterogeneity, the commonly used indicator Moran's I (Anselin, 1995) was applied:

$$Moran's\ I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i}^{N}\sum_{j}^{N} w_{ij}(y_i-\bar{y})\left(y_j-\bar{y}\right)}{\sum_{i}^{N}(y_i-\bar{y})^2} \tag{15}$$

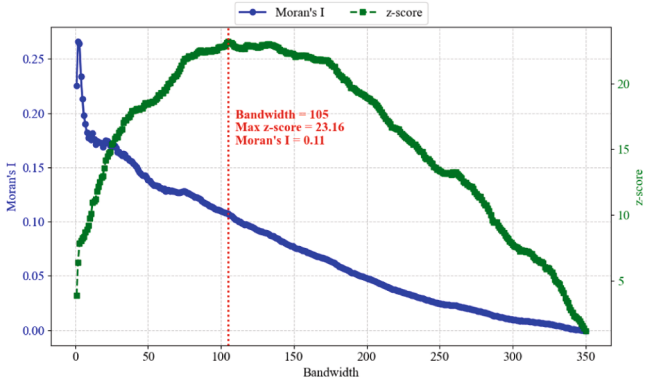
where N is the total number of intersections, y_i denotes the observed crash frequency in intersection i, and $\bar{y}$ is the mean crash frequency across all intersections. The spatial weight w_{ij} is defined as the reciprocal of the distance between intersections i and j. Accordingly, the z-score can be calculated to represent the statistical significance of Moran's I compared to completely random distribution:

$$z-score = \frac{Moran's\ I - E(I)}{\sqrt{Var(I)}} = \frac{I - \left[-\frac{1}{N-1}\right]}{\sqrt{Var(I)}} \tag{16}$$

where $E(I)$ and $Var(I)$ represent the expectation and variance of Moran's I under the assumption that the intersection crash data are completely randomly distributed. Fig. 12**(b)** illustrates the Moran's I and their corresponding z-score at different bandwidth. Moran's I indexes are consistently greater than 0 (p-value < 0.05), indicating the significantly positive spatial clustering of intersection crashes. The z-score reaches highest at bandwidth $= 105$, indicating that under this bandwidth the spatial patterns of intersection crashes are most pronounced. In other words, it can be seen as the optimal spatial scale that best explains the spatial patterns of intersection crashes. The Z-score peak method has been widely used to determine the best bandwidth in spatial analyses (e.g., arcGIS) and modeling studies (Gedamu et al., 2024; Zheng et al., 2024; Sun et al., 2024). Therefore, this band width of 105 was used in our subsequent spatial ML models.



(a) Spatial distribution of intersection crashes



(b) Moran's I and z-score *v.s.* Bandwidth

Fig. 12. Spatial autocorrelation analysis results.

5.2. Model performance comparison

To compare model performance, 5-fold cross-validation was utilized to mitigate the impact of random data partitioning and ensure reliable performance estimation (Han et al., 2024b). To be specific, the entire modeling dataset was partitioned into five equally sized subsets. In each iteration, one-fold (20 %) is designated as the test set while the remaining four folds (80 %) are used for training the model. This process ensures that every data point is used for both training and testing exactly once. The performance metrics from each iteration are then averaged to provide a comprehensive assessment of model performance. Meanwhile, two recent novel spatial MLs in traffic safety studies were also implemented including the GWNN in Zhang et al. (2024) and GWCNNR in Li et al., (2025). Table 6 presents the comparison of baselines between three spatial ML models incorporating XGBoost, RF, and LightGBM. Among these baselines, the GWR model performs the weakest, with the highest RMSE (28.38 ± 2.54), MAE (20.32 ± 2.51), and the lowest R^2 (0.576 ± 0.12). Although GWR can account for potential spatial heterogeneity, it cannot fit the nonlinear dependencies between intersection crashes and impact factors. On the contrary, other ML models can effectively capture such nonlinear correlations, thus demonstrating better prediction performance. Among the baseline ML models, RF provides the best prediction performance with an RMSE of 25.32 ± 3.55, MAE of 17.73 ± 2.48, and R^2 of 0.666 ± 0.10. As expected, the proposed spatial ML models achieve significantly improved prediction performance compared to conventional ML models. Overall, the three spatial ML models demonstrate significantly lower RMSE (<24.81) and MAE (<16.93), with the R^2 improving to over 0.68 compared to baseline

Table 6
Model performance comparison between baselines and three spatial ML models.

Models		RMSE*	MAE*	R^{2}*
Non-spatial models				
Baselines	GWR	28.38 (2.54)	20.32 (2.51)	0.576 (0.12)
	MLP	25.78 (2.69)	17.67 (2.38)	0.657 (0.08)
	XGBoost	26.60 (3.57)	17.75 (2.28)	0.631 (0.11)
	RF	25.32 (3.55)	17.73 (2.48)	0.666 (0.10)
	LightGBM	25.96 (2.28)	17.88 (1.87)	0.646 (0.09)
Spatial models				
GWNN (Zhang et al., 2024)		24.38 (2.80)	16.83 (1.62)	0.696 (0.05)
GWCNNR (Li et al., 2025)		25.33 (3.71)	17.08 (2.15)	0.701 (0.06)
Ours	Spatial XGBoost	24.81 (2.62)	16.84 (1.89)	0.686 (0.05)
	Spatial RF	24.17 (3.37)	16.93 (2.03)	0.701 (0.04)
	Spatial LightGBM	**23.85 (2.21)**	**16.62 (1.19)**	**0.703(0.04)**

*: Values in parentheses represent the standard deviation of the metric across the 5-fold cross-validation.

models. Notably, the spatial LightGBM model outperforms all other models, achieving the lowest RMSE (23.85 ± 2.21), MAE (16.62 ± 1.19) and highest R^2 (0.703 ± 0.04). Compared to the best baseline model (RF), it demonstrates improvements of 5.8 %, 6.3 %, and 5.6 % in RMSE, MAE, and R^2, respectively. Furthermore, compared to standard XGBoost, LightGBM, and RF models, their spatial-version models achieved average improvements of 6.5 %, 5.6 % and 7.6 % in RMSE, MAE, and R^2, respectively. Both GWNN and GWCNNR achieved similar results with our spatial MLs, indicating that integrating diverse MLs within spatial modeling is a promising way to improve prediction performance. Overall, the spatial LightGBM still achieved relatively better results with smaller fluctuations (i.e., standard deviation).

Recent studies (Abdel-Aty et al., 2024; Cai et al., 2022; Wali et al., 2018) have identified visual environment features and driving behavior features as two key impact factors to intersection crashes, with the former representing drivers' visual perception and the latter reflecting micro-level driving dynamics at intersections. To investigate the benefits of introducing such emerging features, a set of ablation experiments were conducted using the best spatial LightGBM, comprising four different models:

1) **Macro (base):** Includes only the four macro-level features (i.e., traffic volume; geometric design, socioeconomic data, and road context classifications).
2) **Macro + V:** Utilizes four macro-level features with visual environment features.
3) **Macro + Lon D:** Combines the four macro-level features with only longitudinal driving behavior features (e.g., hard acceleration and braking during drive-straight).
4) **Macro + V + Full D:** Incorporates the four types of macro-level features with both visual environment and longitudinal and lateral driving behavior features.

Ablation results in Fig. 13 show that the *Macro* model using only macro-level features exhibits the highest RMSE (29.83), MAE (21.11), and lowest R^2 (0.55). While introducing visual environment features, the *Macro + V* model achieves lower RMSE (28.41), lower MAE (21.11), and higher R^2 (0.59), indicating that drivers' visual perception at intersections also impacts their driving operations and safety (Xue et al., 2024; Yue, 2024). Overall, the *Macro + V + Full D* performs best (RMSE = 23.85, MAE = 16.62, and R^2 = 0.70). Compared to the *Macro + Lon D,* it achieves a 9.2 % reduction in RMSE, a 6.5 % reduction in MAE, and a 7.7 % increase in R^2. It highlights that beyond existing studies focusing solely on longitudinal driving behaviors, the inclusion of lateral turning driving features (e.g., hard left/right turns) and visual environment features enables capturing more driver's risky driving interactions and visual perception at intersections, thereby significantly enhancing the crash prediction accuracy (Guo et al., 2021; Hu and Cicchino, 2020).

5.3. Model results interpretation

This section delves into the interpretation of the spatial LightGBM model results. We first identify key influencing factors and their nonlinear effects on the frequency of intersection crashes at the global level (Section 5.3.1). Then, we examine the spatial heterogeneity of these features at the local level, offering a comprehensive understanding of their varying impact across different spatial locations (Section 5.3.2). To account for randomness in model training, we fixed the random seed at 42 (*a common default*), ensuring that feature interpretation results are both reproducible and consistent.

5.3.1. Global level: Nonlinear effects of factors

In the spatial LightGBM model, the global LightGBM model provides globally averaged estimates of the impact of features on intersection crashes across all intersections. Fig. 14 visualizes the top 15 features ranked by importance and their SHAP values. Among these, all 9 micro-level driving behavior features show positive impact, with high values (red points in Fig. 14) corresponding to positive SHAP values, highlighting the strong correlations between risky driving behaviors and intersection crashes. Additionally, six macro-level features exhibit significant impact, including two traffic volume indicators, two socioeconomic features (the poverty ratio (P_poverty) and transit service frequency (Transit_Services) within surrounding census tracts), and two visual environment features (the vegetation and grass proportion in GSV images).

(1) Driving behavior features

Fig. 15 compares the frequencies of intersection crashes and the SHAP values for three key driving behavior features associated with drive-straight, left-turn, and right-turn maneuvers. Based on the results, the following conclusions can be drawn:

1) ***Drive-straight:*** The top contributing factor is the sum of extreme accelerations during drive-straight maneuvers (Straight_acceleration_sum), representing the daily cumulative risk of acceleration behaviors within intersection areas. Fig. 15**(a)** shows a clear positive correlation between this feature and intersection

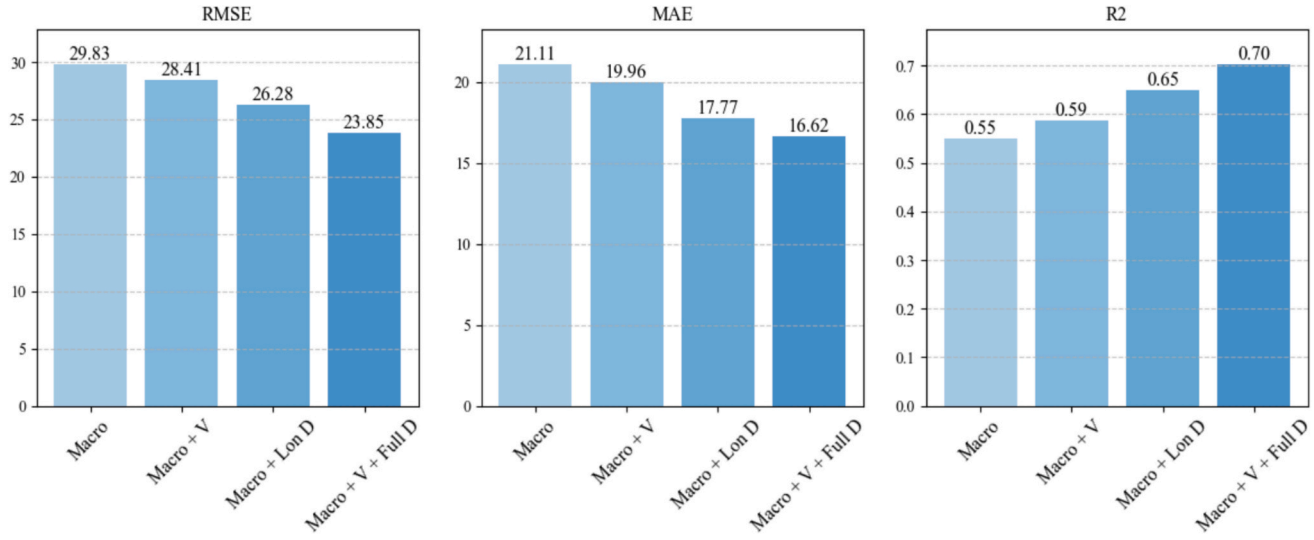


Fig. 13. Ablation experiment results.

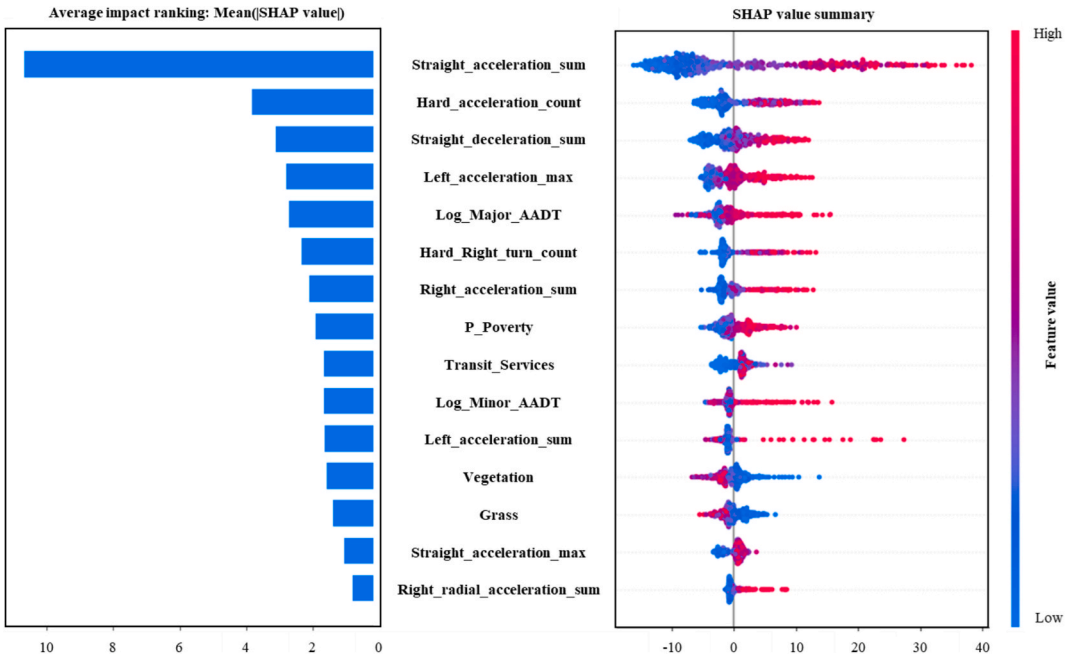


Fig. 14. Global-level feature importance ranking and their SHAP values (Top-15).

crashes, indicating that risky longitudinal accelerations may reduce reaction time and cause dangerous gap distances with front vehicles, strongly increasing the likelihood of crashes (Azadani and Boukerche, 2022; Wali et al., 2018). Notably, these impacts exhibit a nonlinear trend as shown in Fig. 15**(b)**: when the value is below approximately 700–800 m/s^2, its SHAP values remain negative, indicating limited impact on intersection crashes. Conversely, if it exceeds this threshold (>800 m/s^2), corresponding SHAP value shifts positive and increases sharply, meaning that a high cumulation of risky acceleration behaviors would cause substantial impact to contribute to more crashes.

2) ***Left-turn:*** The extreme acceleration of left-turn maneuvers (Left_acceleration_max) ranks as the fourth key factor. While less influential than drive-straight features, it shows a significant positive correlation with the intersection crash frequency (Fig. 15**(c)**). It indicates that left-turn behaviors with high acceleration may increase the potential of conflicts involving the left-turn and opposing drive-straight vehicles (Appiah et al., 2020). Based on its SHAP distribution in Fig. 15**(d)**, its impact present nonlinear trend: when the extreme acceleration is below 2.6–3.0 m/s, the SHAP value remains negative, indicating minimal impact on intersection crashes. While if it exceeds this threshold, the SHAP value becomes positive and increases sharply to lead to more crashes.
3) ***Right-turn:*** The daily count of hard right-turn events (Hard_Right_turn_count) is the sixth most important factor to exhibit a positive impact on intersection crashes. Fig. 15**(e)** reveals the positive correlation between this feature and the intersection crash frequency. As noted by Potts et al. (2013), right-turn traffic may lead to potential conflicts with vehicles in the turning direction or opposing lanes. Hard right-turn behaviors may seriously exacerbate these traffic conflicts, especially when drivers fail to yield, thereby causing crashes. Similarly, its impact follows a nonlinear trend as shown in Fig. 15**(f)**: when the value is below 50, the SHAP value remains negative. However, once this threshold is exceeded, the SHAP value becomes positive at high levels (SHAP > 2.5), which would significantly increase intersection crash frequencies.

(2) Macro-level features

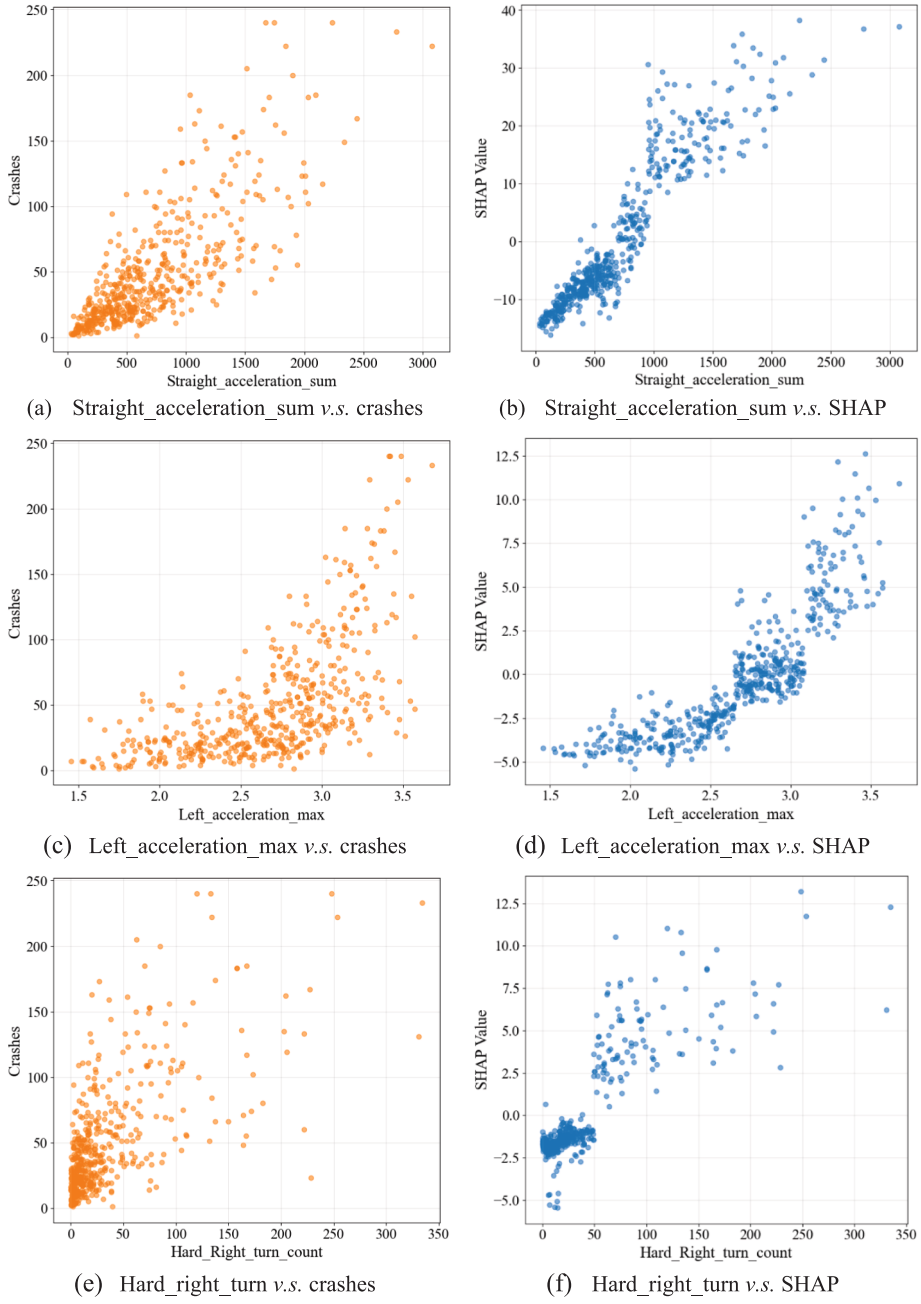
Fig. 16 visualizes the SHAP values of four critical macro-level features. For traffic volume, both the log-transformed AADT of major and minor roads (Log_Major_AADT and Log_Minor_AADT) exhibit positive and nonlinear correlations with crashes. Their SHAP values increase sharply when these features exceed a threshold of 10 (i.e., 22,026 pcu/day). It indicates that compared to low-traffic intersections, high traffic on major or minor roads leads to more exposure to risky vehicle interactions from different directions, thereby directly causing more crashes. Within the socioeconomic features, the surrounding poverty ratio (P_poverty) shows a positive correlation with intersection crashes, indicating that these high-poverty areas may experience higher rates of crashes than other areas, which need more attention for safety improvement (Li et al., 2022; Patwary et al., 2024). For visual environment features, the proportion of vegetation in intersection GSVs shows a negative relationship with crashes. Intersections with high vegetation coverage (5–10 %) have negative SHAP values, reflecting potential benefits of vegetation to reduce crashes.

5.3.2. Local level: Spatial heterogeneity of factors

Unlike global models, which estimate the overall impact of factors without considering their spatial variations, local models capture distinct and location-specific relationships between these factors and crashes to effectively reveal their spatial heterogeneity at each intersection. To further investigate such spatial heterogeneity, feature importance across different intersections and their varying impact at the intersection level are analyzed:

(1) Spatial distribution of feature importances

Taking the four types of hard driving events as an example, Fig. 17 presents the spatial distribution of their importances across different intersections. Variable importance score is derived from each local LightGBM model, reflecting the contribution of each variable to intersection crash frequency prediction. For the same variable, its importance scores can vary between 0 (no contribution) and 1(full contribution) across different intersection sites. A redder color indicates relatively higher importance, while a greener color reflects lower importance. Notably, the uneven distribution of point color highlights the evident spatial dependency of crash frequency. From the results, we can



(a) Straight_acceleration_sum *v.s.* crashes (b) Straight_acceleration_sum *v.s.* SHAP

(c) Left_acceleration_max *v.s.* crashes (d) Left_acceleration_max *v.s.* SHAP

(e) Hard_right_turn *v.s.* crashes (f) Hard_right_turn *v.s.* SHAP

Fig. 15. Scatter plots of intersection crash frequency and SHAP with driving behavior features.

conclude that:

1) ***Hard acceleration:*** Fig. 17**(a)** shows that most of the redder points are primarily located in urban areas, suggesting that hard acceleration events have a greater impact on crashes at urban intersections compared to other regions (e.g., downtown or suburban areas). To quantify the spatial differences of this feature, the model's local importance scores show that intersections in urban areas exhibit higher scores between 0.05 and 0.24 (mean = 0.13), whereas other areas range from 0.01 to 0.15 (mean = 0.03). This indicates that in such urban areas, the importance of hard acceleration frequency is nearly four times greater than other locations. In other words, hard accelerations play a pivotal role in intersection crash frequency prediction in the urban areas. Specifically, most of these urban intersections are often situated on high-volume arterials, characterized by larger sizes and higher speed limits. In over 75 % of these intersections, the speed limits of major roads are higher than that of the minor roads— sometimes by as much as 10mph (e.g., 35mph on the minor and 45mph on the major). This result reveals that in such expansive intersections with high speed limit difference, drivers may be more likely to riskily accelerate to reach higher speeds, thereby significantly increasing the risk of crashes (Arvin et al., 2019; Wali et al., 2018).
2) ***Hard braking:*** As shown in Fig. 17**(b)**, the redder points are highly concentrated in downtown residential areas, while green points are distributed across the urban and suburban areas. It reveals that hard braking is a significant contributor to crashes at downtown intersections. The distribution of local importance scores indicates that intersections in downtown areas exhibit values between 0.08 and 0.20 (mean = 0.15), while those in other areas range from 0.02 to 0.04 (mean = 0.03). These quantitative results mean that hard-braking frequency is roughly five times more influential in

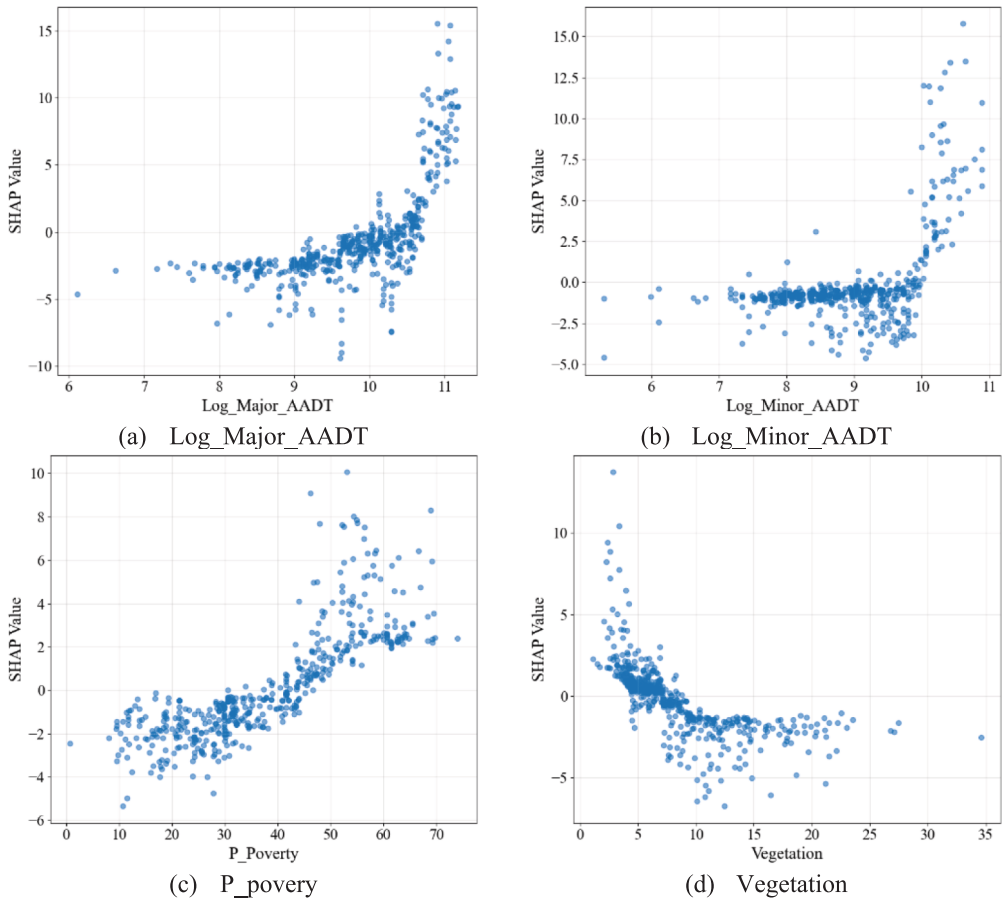


Fig. 16. SHAP plots of four macro-level features.

predicting intersection crashes in downtown areas than elsewhere. This might be because these are highly dense street networks and heavier roadside activities in downtown areas. For instance, intersection spacing in these downtown zones is typically 150–250 m, which is significantly less than that of urban (300–350 m) and suburban area (800–100 m). Moreover, higher traffic and human activities are presented in this area. As a results, such traffic environment makes drivers need to brake frequently when approaching intersections (Dumbaugh et al., 2023; Xie et al., 2014), making hard braking events more likely to occur to become a primary risky issue for crashes (Gu et al., 2023).

3) ***Hard left-turn:*** Unlike the spatial patterns of hard acceleration and braking events, Fig. 17**(c)** illustrates the redder points of hard left-turn events are mainly distributed at specific urban and suburban areas. The local important scores in these urban and suburban intersections range from 0.06 to 0.27, with an average of 0.14—nearly seven times higher than the average score in other areas (0.02). This highlights the substantial spatial variation in the impact of hard left-turn behavior on intersection crash frequency. Specifically, these points are mainly located at intersections along with specific main arterials or freeways. This indicates that these intersections may have more turning traffic exiting from high-volume roadways, thus causing more left-turn conflicts with other direction vehicles. CV data confirm that the ratio of left-turn traffic accounts for 33–39 % of the total volume at these intersections, higher than the average level (25.2 %). As a result, risky left-turn behaviors could lead to crashes with a higher probability in these intersections.
4) ***Hard right-turn:*** The redder points of hard right-turn events are distributed in the southern and northern suburban areas without a specific spatial pattern as shown in Fig. 17**(d)**. For most intersections, the local importance scores of hard right turns are less than 0.01, indicating a relatively minor effect on intersection crashes. However, in these southern and northern suburban intersections, these scores rise significantly to 0.10–0.20, suggesting a greater impact of risky hard right turns on crash occurrences in these regions. This indicates that risky right-turn behavior is a more significant issue to lead to crash on these suburban highways and arterials. Therefore, controlling such risky turning behaviors can be an effective way to reduce crashes for suburban intersection management (Hu and Cicchino, 2020).

(2) Case studies: feature impact at high-crash intersections

Fig. 18 depicts an urban intersection with a speed limit of 45 mph on the major road. Based on crash records, the pie chart shows the ratio of major crash types at this intersection, with rear-end crashes accounting for the majority (63 %). The model SHAP values in red shapes represent each feature's contribution to the estimated crash numbers (f(x)), indicating how many crashes deviate from the average level (E[f(x)]). The results show that the crashes at this intersection are mainly driven by the longitudinal acceleration driving features (e.g., the count of hard acceleration and the sum of extreme accelerations of drive-straight maneuvers (Straight_acceleration_sum)). A lot of hard acceleration events are observed at the four exit approaches of the intersection, meaning that drivers would accelerate rapidly upon exiting the intersection, largely leading to more rear-end crashes. Meanwhile, risky right-turn behaviors (e.g., Right_acceleration_max) and risky left-turn behaviors (e.g., Left_acceleration_max) also show critical contributions and are estimated to increase by 28.1 and 20.6 crashes, respectively. These lateral turning events are also recorded at these exit approaches, which may result in frequent conflicts between such turning and drive-straight vehicles, thus leading to sideswipe and other types of crashes.

Fig. 19 is an example of a large-size suburban intersection with speed

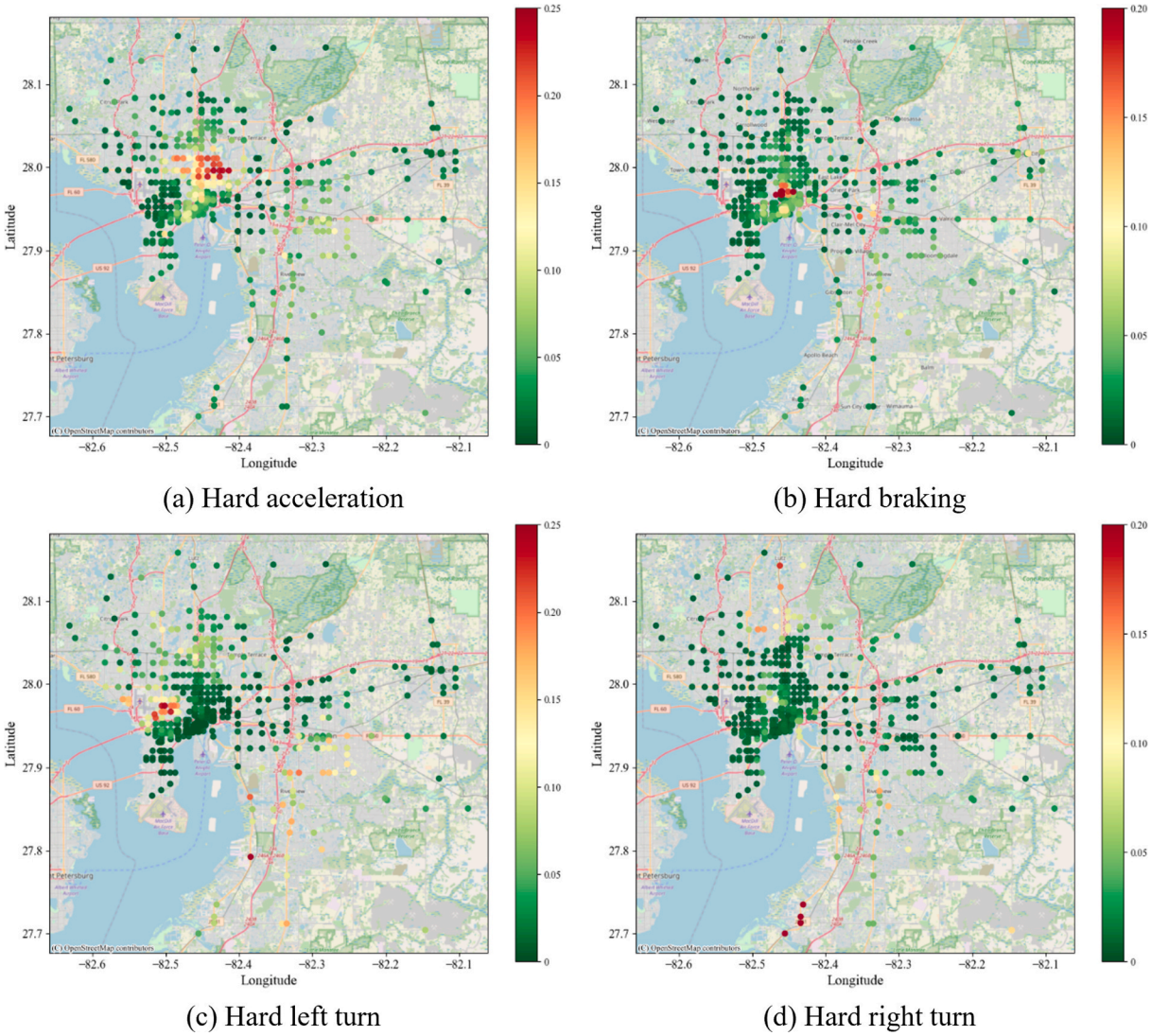


Fig. 17. Spatial distribution of importance scores of four hard event features.

limit of 50 mph on the major road. In this intersection, sideswipe and left-turn crashes are the main concerns, accounting for 44 % of the total crashes. Results show that risky left-turn behaviors are the primary contributing factor of crashes. Specifically, more than 388 hard left-turn events are identified daily, and 70.1 crashes are estimated to be attributed to left-turn features (e.g., the sum of extreme accelerations during left-turn maneuvers, Left_acceleration_sum), highlighting a significant safety concern regarding left-turn movements. Another safety issue is that risky acceleration during drive-straight maneuvers (Straight_acceleration_sum) contributes to an increase of 43 crashes compared to the average level. Due to the high speed limit, drivers tend to accelerate to reach the speed limit or beyond. Thus, a lot of hard acceleration events occurred along the road, which may lead to rear-end crashes. Meanwhile, a significant number of hard right-turn events are also recorded, contributing to about 17 crashes.

6. Discussion

6.1. Crash contributing factors comparison

In general, the overall relationship between macro-level contributing factors and intersection crash frequency is consistent with previous studies (Cai et al., 2018; Park et al., 2020; Pulugurtha and Sambhara, 2011; Yue, 2024). For example, as the most important traffic exposures for crashes, both the AADT of major and minor roads are found to have a strong positive correlation with crash frequency, which is consistent with existing studies (Cai et al., 2018; Xue et al., 2024; Yue, 2024). The surrounding poverty ratio (P_poverty) is found to show a positive correlation with intersection crashes. Li et al. (2022) and Patwary et al. (2024) also found that high poverty areas have a statistically significant higher crash frequency than other areas. Interestingly, the visual environment feature—the proportion of vegetation at intersections—shows a negative relationship with crashes, also echoing with recent studies by Abdel-Aty et al., 2024; Cai et al., 2022; and Yue, 2024. They found that roadside vegetation may make drivers feel a narrow road and exercise more caution, thereby reducing speeding, distracted driving, and crashes.

As for the micro-level crash contributing factors, this study reveals several distinctive patterns linking risky driving behaviors with intersection crash frequency:

1) Nonlinear impacts of risky driving behaviors on intersection crashes commonly exist. For example, the daily cumulative risk of acceleration behaviors is identified as the leading contributor to intersection crashes. While its impact remains low under a certain threshold (<1000 m/s^2), exceeding this threshold causes the corresponding SHAP value to turn positive and increase sharply, leading to a

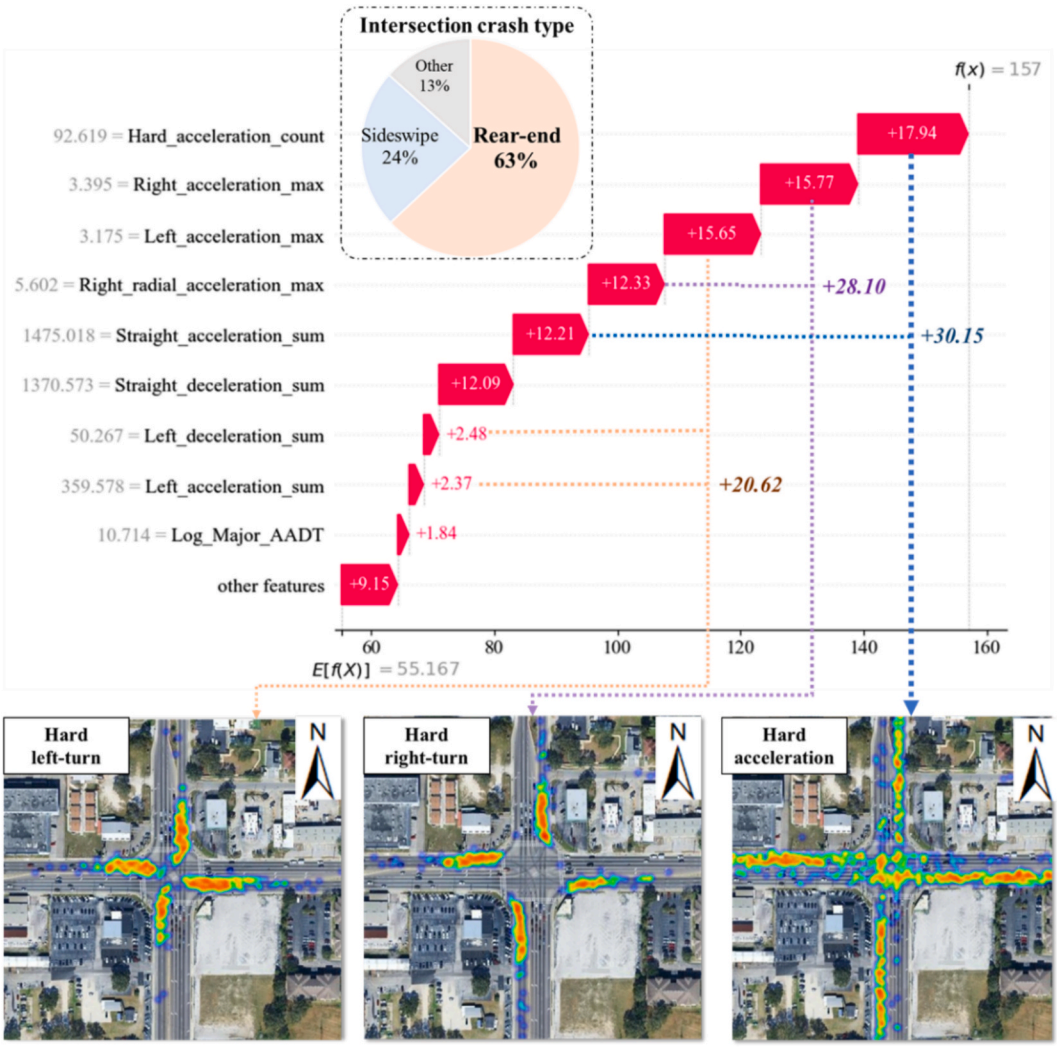


Fig. 18. Case study 1: Urban high-crash intersection.

substantial rise in crashes. Similar nonlinear effects are also found in Gu et al., (2023).

2) Compared to models that only use longitudinal driving behaviors, models further considering lateral left and right turning behaviors achieve superior predictions. The model interpretation also shows that several lateral driving behavior features (e.g., the extreme acceleration of left-turn maneuvers, daily count of hard right-turn events) are significantly positively associated with crash frequency. All these results highlight the critical role of lateral turning maneuvers for intersection crash modeling.
3) Local models in spatial ML reveal pronounced spatial heterogeneous relationships between risky driving behavior features and crashes across different intersections. Findings indicate that in downtown areas with dense street networks, hard braking events are more likely to occur and become a primary risk factor for crashes. While for urban areas, drivers may be more likely to risky accelerate to reach high speed limit, thereby significantly increasing the risk of crashes. At some suburban intersections, left-turn traffic may experience significant conflicts with vehicles traveling in other directions, highlighting the need for further risky-turn managements to enhance safety.

6.2. Potential safety countermeasures and policies

Based on the main findings as discussed at **Section 6.1**, several important implications and safety countermeasures can be derived for intersection safety practices and policies.

(1) ***Pay more attention to improving intersection safety at high-volume, underdeveloped communities:*** The results show that the intersection traffic and poverty ratio of communities surrounding intersections is positively correlated with intersection crashes. It indicates that such areas always have heavy traffic and more low-income population suffer more crashes and safety issue, which need increased attention and government support. For example, future transportation plans and policies should prioritize developing or maintaining road infrastructure to enhance safety in these underdeveloped communities.
(2) ***Coordinated signal timing in urban areas to reduce hard braking and acceleration at intersections:*** It is found that most intersections in urban and urban-core areas experience frequent hard braking and acceleration events, leading to frequent rear-end crashes. Therefore, implementing coordinated signal timing on major arterials may be an effective way to reduce stop-and-go traffic, thereby decreasing abrupt acceleration and braking maneuvers. Additionally, lowering the speed limits at locations with high rates of hard acceleration—for example, reducing from 45 mph to 40 mph or 35 mph—may discourage rapid acceleration and enhance overall safety.

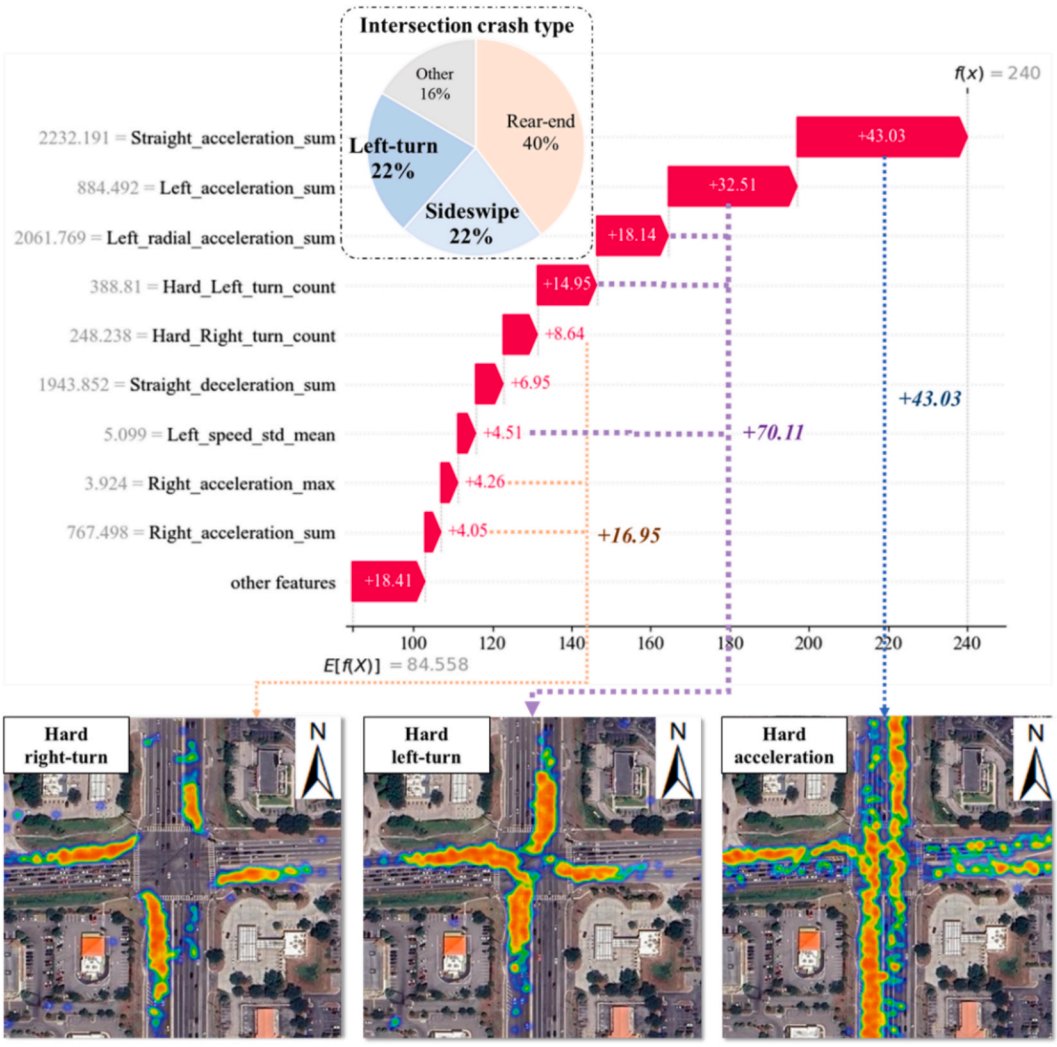


Fig. 19. Case study 2: Suburban high-crash intersection.

(3) ***Implement protected left/right-turn signals at suburban intersections with frequent hard turns:*** Several suburban intersections exhibit high proportions of left-turn and sideswipe crashes, driven primarily by frequent hard left/right turning maneuvers. To improve safety at these locations, consider adding or extending protected left- and right-turn signal phases to avoid potential conflicts. Additionally, implementing the "No Right-turn on Red" restriction can also reduce conflicts between right-turning and through traffic, thereby lowering the incidence of right-turn related crashes.

7. Conclusion

Existing studies have developed both statistical and machine learning models to identify factors contributing to intersection crashes (Kamrani et al., 2017; Wali et al., 2018). However, due to data limitation, previous studies have focused on macro-level static infrastructure and highly aggregated traffic features, ignoring the influence of micro-level human driving behaviors on intersection crashes (Gu et al., 2023; Wang et al., 2021). Leveraging emerging CV data, some studies have explored the impact of longitudinal driving behaviors (e.g., hard acceleration and braking) on intersection crashes, but critical lateral behaviors like left and right turns have largely been overlooked (Hunter et al., 2021). In terms of modeling approaches, existing studies have mainly integrated spatial effects into statistical methods to account for complex spatial heterogeneity (Li et al., 2022). These models, however, rely heavily on the linear assumptions, resulting in poor prediction performance and difficulties in handling high-dimensional traffic data (Zhou et al., 2023).

To address the first research gap, we divided intersection CV trajectories into drive-straight, left-turn, and right-turn movements. Driving behavior features at both longitudinal (e.g., hard braking and acceleration) and lateral turning (e.g., hard left and right turn) maneuvers were identified to capture the micro-level driving dynamics within intersections. To overcome modeling limitations, we proposed a novel spatial ML framework integrating nonlinear ML models (e.g., RF, XGBoost, and LightGBM) with geographically weighted regression. High-resolution CV data at Hillsborough County were utilized for experiments. Based on the empirical results, two key insights can be concluded:

1) The inclusion of risky turning driving features (e.g., hard left/right turns) enables capturing more lateral driving interactions to enhance intersection crash prediction. For instance, risky left-turn behaviors with high acceleration rank as the 4th most important feature with positive association with intersection crashes. It indicates that such risky left-turn behaviors may increase potential conflicts between left-turn vehicles and other traffic streams, thus undermining intersection safety.

2) Spatial heterogeneity of micro driving behaviors reveals that intersections at different locations experience different risky-driving behavior issues. In downtown, hard braking events primarily influence intersection crashes. Drivers' hard acceleration is more likely to lead to rear-end crashes in urban areas, aligning with findings in Gu et al. 2023. In contrast, at suburban intersections with high left-turn volume, hard left turns show greater influence on sideswipe and left turn crashes. By pinpointing intersections with risky left/right turn behaviors, it can provide new actionable insights for traffic management and targeted safety interventions.

The proposed method can be used for intersection safety evaluation and management. On the one hand, even short-term CV data enables researchers to proactively identify hotspot intersections where crashes are waiting to happen. These hotspots may be those intersections with low crash frequency but high risky driving events (Kamrani et al., 2017; Wang et al., 2024c). On the other hand, utilizing V2X communication, proactive warnings could be generated at these high-risk intersections to inform drivers about potential hazards, which could enhance drivers' situational awareness and prevent crashes (Arvin et al., 2019; Chen et al., 2024). Nonetheless, there are still a few limitations in the current study. First, the current study was based on a specific region. Future research should consider testing the model in diverse geographic contexts to evaluate its generalizability. Second, the penetration of CV data may vary across different intersections, which would potentially affect the feature values and further affect the local model performance. Third, within the CV dataset, over 53 % of vehicles are multipurpose passenger vehicles (e.g., SUV, Vans, and Minivans), 31.5 % are large-size pickup trucks, and 15.3 % are sedans. Regarding fuel type, 91.8 % run on gasoline, 7 % on diesel, and only 1.2 % are electric vehicles (EVs). Although these results indicate that the CV dataset comprises a variety of vehicle types, the proportion of EVs is relatively small in the Hillsborough area, which may introduce sampling bias if the method is applied directly to regions with high EV penetration (e.g., Shanghai or Beijing). Finally, short-time CV data were used in this study, yet intersection driving behaviors may have changed over the studied period. Due to data limitations, existing studies typically collected CV data over a limited duration (e.g., 1–3 months) (Arvin et al., 2019; Hunter et al., 2021; Joshi et al., 2024). It is necessary to collect more extensive CV data to examine the temporal stability of driving behaviors and their correlations with crashes.

CRediT authorship contribution statement

Lei Han: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Mohamed Abdel-Aty:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work was supported by the National Science Foundation (NSF) and Center for Smart Streetscapes (CS3) under NSF Cooperative Agreement No. EEC-2133516.

Appendix

Table A1
Descriptive Statistics of the target and macro-level variables.

Variable	Definition	Min	Max	Mean	STD
Crashes	Crashes per intersection from June 2021 to May 2024	0	240	49.32	44.72
Traffic volume					
Log_Major_AADT	The log value of AADT on major road in 2021–2023 (pcu)	6.11	11.18	9.79	0.86
Log_Minor_AADT	The log value of AADT on minor road in 2021–2023 (pcu)	5.30	10.89	8.90	0.90
Log_Major_TruckAADT	The log value of truck AADT on major road in 2021–2023 (pcu)	3.56	9.01	7.18	0.88
Log_Minor_TruckAADT	The log value of truck AADT on minor road in 2021–2023 (pcu)	2.77	8.56	6.36	0.91
Geometric design					
Legs_4	4-legged (yes = 1)	0	1	0.73	0.45
Major_lanes	Major road lanes > 4 (yes = 1)	0	1	0.20	0.40
Minor_lanes	Minor road lanes > 4 (yes = 1)	0	1	0.02	0.14
Major_speed_low	Speed limit on major road < 40 mph (yes = 1)	0	1	0.38	0.49
Major_speed_medium	Speed limit on major road in 40–50 mph (yes = 1)	0	1	0.59	0.49
Major_speed_high	Speed limit on major road > 50 mph (yes = 1)	0	1	0.03	0.18
Minor_speed_low	Speed limit on minor road < 40 mph (yes = 1)	0	1	0.66	0.47
Minor_speed_medium	Speed limit on minor road in 40–50 mph (yes = 1)	0	1	0.34	0.47
Minor_speed_high	Speed limit on minor road > 50 mph (yes = 1)	0	1	0.00	0.06
Major_surface_width	The surface width of major road (feet)	13	96	43.19	17.21
Minor_surface_width	The surface width of minor road (feet)	12	69	31.23	10.83
Major_minor_collector	Major road class is minor collector (yes = 1)	0	1	0.02	0.14
Major_major_collector	Major road class is major collector (yes = 1)	0	1	0.24	0.43
Major_minor_arterial	Major road class is minor arterial (yes = 1)	0	1	0.35	0.48
Major_major_arterial	Major road class is major arterial (yes = 1)	0	1	0.39	0.49
Minor_minor_local	Minor road class is local road (yes = 1)	0	1	0.02	0.13
Minor_minor_collector	Minor road class is minor collector (yes = 1)	0	1	0.20	0.40
Minor_major_collector	Minor road class is major collector (yes = 1)	0	1	0.56	0.50
Minor_minor_arterial	Minor road class is minor arterial (yes = 1)	0	1	0.19	0.40
Minor_major_arterial	Minor road class is major arterial (yes = 1)	0	1	0.04	0.19
Major_median_marking	The median type of major road is a traffic marking (yes = 1)	0	1	0.53	0.50
Major_median_separator	The median type of major road is a raised traffic separator (yes = 1)	0	1	0.28	0.45
Major_median_curb	The median type of major road is curb and vegetation (yes = 1)	0	1	0.13	0.34
Minor_median_marking	The median type of major road is a traffic marking (yes = 1)	0	1	0.71	0.45

(*continued on next page*)

Table A1 (*continued*)

Variable	Definition	Min	Max	Mean	STD
Minor_median_separator	The median type of minor road is a raised traffic separator (yes = 1)	0	1	0.18	0.39
Minor_median_curb	The median type of minor road is curb and vegetation (yes = 1)	0	1	0.09	0.29
Socioeconomic variables					
Population	Average population (1000 people)	0.59	10.66	4.06	1.41
Median_Income	Average median income (10000$)	1.28	15.07	5.79	2.52
P_Over65	Percent of population 65 years or older (%)	2.87	78.74	13.57	4.94
P_Under17	Percent of population 17 years or younger (%)	0.42	37.70	19.65	7.17
P_Unemployed	Percent of people age 16 + unemployed (%)	0	14.60	3.73	1.95
P_Poverty	Percent of population with income below average poverty level (%)	0.72	74.02	37.56	15.38
P_Uneducation	Percent of people age 25 + with less than a high school diploma (%)	0.51	41.90	14.51	7.84
P_Disability	Percent of population with a disability (%)	3.19	62.10	13.30	4.49
P_Mobile_Homes	Percent of total housing units that are mobile homes (%)	0	79.10	4.79	9.94
P_Nocar	Percent of households with no car (%)	0	35.17	10.46	7.46
Commute_Time	Average commute time to work (min)	7.77	39.08	25.09	4.48
Transit_Services	Frequency of Transit Services per Sq Mi	0	293.02	26.98	36.41
Road context classifications					
C3R	Intersection at suburban residential area (yes = 1)	0	1	0.20	0.40
C3C	Intersection at suburban commercial area (yes = 1)	0	1	0.14	0.35
C4	Intersection at urban general area (yes = 1)	0	1	0.42	0.49
C5 & C6	Intersection at urban center area (yes = 1)	0	1	0.08	0.27
Visual environment features					
Sidewalk	Proportion of sidewalk area within the intersection GSV images (%)	0	7.44	1.44	1.22
Grass	Proportion of grass or terrain within the intersection GSV images (%)	0	20.36	2.09	2.12
Vegetation	Proportion of vegetation (e.g., tree, grass) within the intersection GSV images (%)	1.11	34.60	8.43	4.90
Road	Proportion of road within the intersection GSV images (%)	16.67	45.80	41.05	3.43
Buildings	Proportion of buildings within the intersection GSV images (%)	0	36.31	4.34	6.33
Vehicle	Proportion of vehicles within the intersection GSV images (%)	0	29.94	2.08	2.33
Sky	Proportion of sky visible within the intersection GSV images (%)	9.06	47.00	37.80	7.25

Table A2
Descriptive Statistics of the micro-level driving behavior variables.

Variable	Definition	Min	Max	Mean	STD
Speed volatility					
Straight_speed_std_mean	Mean of speed standard deviation of drive-straight trajectories (m/s)	2.36	9.45	5.62	1.12
Left_speed_std_mean	Mean of speed standard deviation of left-turn trajectories (m/s)	2.97	8.91	6.12	1.26
Right_speed_std_mean	Mean of speed standard deviation of right-turn trajectories (m/s)	1.85	8.54	6.22	1.18
Straight_speed_std_max	Maximum of speed standard deviation of drive-straight trajectories (m/s)	11.43	29.03	19.62	3.45
Left_speed_std_max	Maximum of speed standard deviation of left-turn trajectories (m/s)	7.24	25.57	15.15	2.58
Right_speed_std_max	Maximum of speed standard deviation of right-turn trajectories (m/s)	4.36	24.12	15.54	2.96
Hard events					
Hard braking	Number of hard braking at intersection per day	0	120.19	25.14	25.63
Hard acceleration	Number of hard accelerations at intersection per day	0	239.57	37.98	40.35
Hard left turning	Number of hard left turn at intersection per day	0	388.81	43.28	57.50
Hard right turning	Number of hard right turn at intersection per day	0	334.24	33.42	48.16
Risky maneuver with extreme accelerations					
Straight_acc_mean	Mean of acceleration of drive-straight trajectories (m/s^2)	0.24	1.34	0.78	0.17
Left_acc_mean	Mean of acceleration of left-turn trajectories (m/s^2)	0.37	1.48	1.12	0.23
Right_acc_mean	Mean of acceleration of right-turn trajectories (m/s^2)	0.33	1.56	1.11	0.24
Straight_acc_max	Maximum of acceleration of drive-straight trajectories (m/s^2)	1.64	5.47	3.49	0.48
Left_acc_max	Maximum of acceleration of left-turn trajectories (m/s^2)	1.45	3.68	2.67	0.46
Right_acc_max	Maximum of acceleration of right-turn trajectories (m/s^2)	0.57	3.97	2.80	0.51
Straight_acc_sum	Sum of acceleration of drive-straight trajectories (m/s^2)	31.27	3075.24	725.57	503.28
Left_acc_sum	Sum of acceleration of left-turn trajectories (m/s^2)	8.49	884.49	177.70	159.04
Right_acc_sum	Sum of acceleration of right-turn trajectories (m/s^2)	1.12	954.57	179.29	157.04
Straight_dec_mean	Mean of deceleration of drive-straight trajectories (m/s^2)	0.26	1.61	0.85	0.21
Left_dec_mean	Mean of deceleration of left-turn trajectories (m/s^2)	0.11	1.04	0.42	0.16
Right_dec_mean	Mean of deceleration of right-turn trajectories (m/s^2)	0.17	1.46	0.81	0.21
Straight_dec_max	Maximum of deceleration of drive-straight trajectories (m/s^2)	2.41	5.88	3.94	0.57
Left_dec_max	Maximum of deceleration of left-turn trajectories (m/s^2)	1.20	3.72	2.47	0.39
Right_dec_max	Maximum of deceleration of right-turn trajectories (m/s^2)	0.52	4.07	2.84	0.51
Straight_dec_sum	Sum of deceleration of drive-straight trajectories (m/s^2)	38.16	2381.07	725.79	429.43
Left_dec_sum	Sum of deceleration of left-turn trajectories (m/s^2)	5.14	643.55	54.51	47.15
Right_dec_sum	Sum of deceleration of right-turn trajectories (m/s^2)	1.46	638.96	129.77	115.18
Left_radial_acc_mean	Mean of radial acceleration of left-turn trajectories (m/s^2)	1.01	3.05	2.23	0.37
Right_radial_acc_mean	Mean of radial acceleration of right-turn trajectories (m/s^2)	0.71	2.81	2.10	0.35
Left_radial_acc_max	Maximum of radial acceleration of left-turn trajectories (m/s^2)	2.87	8.83	4.94	0.88
Right_radial_acc_max	Maximum of radial acceleration of right-turn trajectories (m/s^2)	1.02	7.84	4.68	0.74
Left_radial_acc_sum	Sum of radial acceleration of left-turn trajectories (m/s^2)	15.92	2061.77	357.37	325.46
Right_radial_acc_sum	Sum of radial acceleration of right-turn trajectories (m/s^2)	2.11	1888.46	349.25	320.00

Data availability

The authors do not have permission to share data.

References

Abdel-Aty, M., Ugan, J., Islam, Z., 2024. Exploring the influence of drivers' visual surroundings on speeding behavior. Accid. Anal. Prev. 198, 107479. https://doi.org/10.1016/j.aap.2024.107479.

Al-Omari, M.M.A., Abdel-Aty, M., Cai, Q., 2021. Crash analysis and development of safety performance functions for Florida roads in the framework of the context classification system. J. Saf. Res. 79, 1–13. https://doi.org/10.1016/j.jsr.2021.08.004.

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. Geogr. Anal. 27, 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Appiah, J., King, F.A., Fontaine, M.D., Cottrell, B.H., 2020. Left turn crash risk analysis: Development of a microsimulation modeling approach. Accid. Anal. Prev. 144, 105591. https://doi.org/10.1016/j.aap.2020.105591.

Arvin, R., Kamrani, M., Khattak, A.J., 2019. How instantaneous driving behavior contributes to crashes at intersections: Extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133. https://doi.org/10.1016/j.aap.2019.01.014.

Avelar, R.E., Dixon, K.K., Escobar, P., 2015. Evaluation of Signalized-Intersection Crash Screening Methods based on Distance from Intersection. Transp. Res. Rec. 2514, 177–186. https://doi.org/10.3141/2514-19.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically Weighted Regression. J. r. Stat. Soc. Ser. Stat. 47, 431–443. https://doi.org/10.1111/1467-9884.00145.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Anal. Methods Accid. Res 19, 1–15. https://doi.org/10.1016/j.amar.2018.05.001.

Cai, Q., Abdel-Aty, M., Zheng, O., Wu, Y., 2022. Applying machine learning and google street view to explore effects of drivers' visual environment on traffic safety. Transp. Res. Part C Emerg. Technol. 135, 103541. https://doi.org/10.1016/j.trc.2021.103541.

Chen, K., Xu, C., Liu, P., Li, Z., Wang, Y., 2024. Evaluating the performance of traffic conflict measures in real-time crash risk prediction using pre-crash vehicle trajectories. Accid. Anal. Prev. 203, 107640. https://doi.org/10.1016/j.aap.2024.107640.

Cui, P., Yang, X., Abdel-Aty, M., Zhang, J., Yan, X., 2024. Advancing urban traffic accident forecasting through sparse spatio-temporal dynamic learning. Accid. Anal. Prev. 200, 107564. https://doi.org/10.1016/j.aap.2024.107564.

Deng, L., Adjouadi, M., Rishe, N., 2020. Inverse Distance Weighted Random Forests: Modeling Unevenly distributed Non-Stationary Geographic Data. In: In: 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS). Presented at the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 41–46. https://doi.org/10.1109/ICACSIS51025.2020.9263208.

Desai, J., Li, H., Mathew, J.K., Cheng, Y.-T., Habib, A., Bullock, D.M., 2021. Correlating Hard-Braking activity with Crash Occurrences on Interstate Construction Projects in Indiana. J. Big Data Anal. Transp. 3, 27–41. https://doi.org/10.1007/s42421-020-00024-x.

Do, T.-N., Lenca, P., Lallich, S., Pham, N.-K., 2010. Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees, in: Guillet, F., Ritschard, G., Zighed, D.A., Briand, H. (Eds.), Advances in Knowledge Discovery and Management. Springer, Berlin, Heidelberg, pp. 39–55. Doi: 10.1007/978-3-642-00580-0_3.

Dumbaugh, E., Haule, H., Stiles, J., Khattak, A., Collaborative Sciences Center for Road Safety, Florida Atlantic University, University of Tennessee, K., 2023. A Safe Systems Approach to Motorcycle Safety (No. CSCRS-R40).

Fan, C., Xu, J., Natarajan, B.Y., Mostafavi, A., 2023a. Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. Comput. Civ. Infrastruct. Eng. 38, 2013–2029.

Fan, Z., Zhang, F., Loo, B.P.Y., Ratti, C., 2023b. Urban visual intelligence: Uncovering hidden city profiles with street view images. Proc. Natl. Acad. Sci. 120, e2220417120. https://doi.org/10.1073/pnas.2220417120.

Federal Highway Administration (FHWA), 2024. Intersection Safety. U.S. Department of Transportation. https://highways.dot.gov/research/research-programs/safety/intersection-safety.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale Geographically Weighted Regression (MGWR). Ann. Am. Assoc. Geogr. 107, 1247–1265. https://doi.org/10.1080/24694452.2017.1352480.

Gedamu, W.T., Plank-Wiedenbeck, U., Wodajo, B.T., 2024. A spatial autocorrelation analysis of road traffic crash by severity using Moran's I spatial statistics: a comparative study of Addis Ababa and Berlin cities. Accid. Anal. Prev. 200, 107535. https://doi.org/10.1016/j.aap.2024.107535.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 36, 121–136. https://doi.org/10.1080/10106049.2019.1595177.

Goel, A., Goel, A.K., Kumar, A., 2023. The role of artificial neural network and machine learning in utilizing spatial information. Spat. Inf. Res. 31, 275–285. https://doi.org/10.1007/s41324-022-00494-x.

Gu, Y., Liu, D., Arvin, R., Khattak, A.J., Han, L.D., 2023. Predicting intersection crash frequency using connected vehicle data: a framework for geographical random forest. Accid. Anal. Prev. 179, 106880. https://doi.org/10.1016/j.aap.2022.106880.

Guo, M., Zhao, X., Yao, Y., Bi, C., Su, Y., 2022. Application of risky driving behavior in crash detection and analysis. Phys. Stat. Mech. Appl. 591, 126808. https://doi.org/10.1016/j.physa.2021.126808.

Guo, M., Zhao, X., Yao, Y., Yan, P., Su, Y., Bi, C., Wu, D., 2021. A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. Accid. Anal. Prev. 160, 106328. https://doi.org/10.1016/j.aap.2021.106328.

Han, L., Abdel-Aty, M., Yu, R., Wang, C., 2024a. LSTM + transformer real-time crash risk evaluation using traffic flow and risky driving behavior data. IEEE Trans. Intell. Transp. Syst. 1–13. https://doi.org/10.1109/TITS.2024.3438616.

Han, L., Yu, R., Wang, C., Abdel-Aty, M., 2024b. Transformer-based modeling of abnormal driving events for freeway crash risk evaluation. Transp. Res. Part C Emerg. Technol. 165, 104727. https://doi.org/10.1016/j.trc.2024.104727.

Hong, J., Shankar, V.N., Venkataraman, N., 2016. A spatially autoregressive and heteroskedastic space-time pedestrian exposure modeling framework with spatial lags and endogenous network topologies. Anal. Methods Accid. Res 10, 26–46. https://doi.org/10.1016/j.amar.2016.05.001.

Hu, J., Huang, M.-C., Yu, X., 2020. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. Accid. Anal. Prev. 144, 105665. https://doi.org/10.1016/j.aap.2020.105665.

Hu, W., Cicchino, J.B., 2020. The effects of left-turn traffic-calming treatments on conflicts and speeds in Washington. DC. J. Safety Res. 75, 233–240. https://doi.org/10.1016/j.jsr.2020.10.001.

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Methods Accid. Res 14, 10–21. https://doi.org/10.1016/j.amar.2017.01.001.

Hunter, M., Saldivar-Carranza, E., Desai, J., Mathew, J.K., Li, H., Bullock, D.M., 2021. A proactive approach to evaluating intersection safety using hard-braking data. J. Big Data Anal. Transp. 3, 81–94. https://doi.org/10.1007/s42421-021-00039-y.

Joshi, M., Bamney, A., Wang, K., Zhao, S., Ivan, J., Jackson, E., 2024. Analyzing the suitability of vehicle telematics data as a surrogate safety measure for short-term crashes. Transp. Res. Rec. 03611981241263341. https://doi.org/10.1177/03611981241263341.

Kabir, R., Remias, S.M., Lavrenz, S.M., Waddell, J., 2021. Assessing the impact of traffic signal performance on crash frequency for signalized intersections along urban arterials: a random parameter modeling approach. Accid. Anal. Prev. 149, 105868. https://doi.org/10.1016/j.aap.2020.105868.

Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting useful information from basic safety message data: an empirical study of driving volatility measures and crash frequency at intersections. Transp. Res. Rec. J. Transp. Res. Board 2672, 290–301. https://doi.org/10.1177/0361198118773869.

Kamrani, M., Wali, B., Khattak, A.J., 2017. Can data generated by connected vehicles enhance safety?: proactive approach to intersection safety management. Transp. Res. Rec. J. Transp. Res. Board 2659, 80–90. https://doi.org/10.3141/2659-09.

Lee, J., Liu, H., Abdel-Aty, M., 2023a. Changes in traffic crash patterns: before and after the outbreak of COVID-19 in Florida. Accid. Anal. Prev. 190, 107187. https://doi.org/10.1016/j.aap.2023.107187.

Lee, T., Cunningham, C., Rouphail, N., 2023b. Movement-based intersection crash frequency modeling. J. Transp. Saf. Secur. 15, 493–514. https://doi.org/10.1080/19439962.2022.2092571.

Li, J., Yang, Y., Hu, Y., Zhu, X., Ma, N., Yuan, X., 2023. Using multidimensional data to analyze freeway real-time traffic crash precursors based on XGBoost-SHAP algorithm. J. Adv. Transp. 2023, e5789573. https://doi.org/10.1155/2023/5789573.

Li, T., Liu, S., Fan, G., Zhao, H., Zhang, M., Fan, J., Li, C., 2025. Spatial heterogeneity effect of built environment on traffic safety using geographically weighted atrous convolutions neural network. Accid. Anal. Prev. 213, 107934. https://doi.org/10.1016/j.aap.2025.107934.

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.

Liu, Y., Chen, T., Chung, H., Jang, K., Xu, P., 2025. Is there an emotional dimension to road safety? a spatial analysis for traffic crashes considering streetscape perception and built environment. Anal. Methods Accid. Res 100374. https://doi.org/10.1016/j.amar.2025.100374.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Mahmoud, N., Abdel-Aty, M., Cai, Q., Zheng, O., 2021. Vulnerable road users' crash hotspot identification on multi-lane arterial roads using estimated exposure and considering context classification. Accid. Anal. Prev. 159, 106294. https://doi.org/10.1016/j.aap.2021.106294.

Mohammadnazar, A., Patwary, A.L., Moradloo, N., Arvin, R., Khattak, A.J., 2022. Incorporating driving volatility measures in safety performance functions: improving safety at signalized intersections. Accid. Anal. Prev. 178, 106872. https://doi.org/10.1016/j.aap.2022.106872.

Nguyen, T.-T., Huang, J.Z., Nguyen, T.T., 2015. Unbiased feature selection in learning random forests for high-dimensional data. Sci. World J. 2015, 471371. https://doi.org/10.1155/2015/471371.

Park, H.-C., Yang, S., Park, P.Y., Kim, D.-K., 2020. Multiple membership multilevel model to estimate intersection crashes. Accid. Anal. Prev. 144, 105589. https://doi.org/10.1016/j.aap.2020.105589.

Patwary, A.L., Haque, A.M., Mahdinia, I., Khattak, A.J., 2024. Investigating transportation safety in disadvantaged communities by integrating crash and

Environmental Justice data. Accid. Anal. Prev. 194, 107366. https://doi.org/10.1016/j.aap.2023.107366.
Potts, I.B., Bauer, K.M., Torbic, D.J., Ringert, J.F., 2013. Safety of channelized right-turn lanes for motor vehicles and pedestrians. Transp. Res. Rec. 2398, 93–100. https://doi.org/10.3141/2398-11.
Pulugurtha, S.S., Sambhara, V.R., 2011. Pedestrian crash estimation models for signalized intersections. Accid. Anal. Prev. 43, 439–446. https://doi.org/10.1016/j.aap.2010.09.014.
Qu, Y., Lin, Z., Li, H., Zhang, X., 2019. Feature recognition of urban road traffic accidents based on GA-XGBoost in the context of big data. IEEE Access 7, 170106–170115. https://doi.org/10.1109/ACCESS.2019.2952655.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. Doi: 10.1145/2939672.2939778.
Sander, U., 2017. Opportunities and limitations for intersection collision intervention—A study of real world 'left turn across path' accidents. Accid. Anal. Prev. 99, 342–355. https://doi.org/10.1016/j.aap.2016.12.011.
Shirazi, M.S., Morris, B.T., 2017. Looking at Intersections: a Survey of Intersection monitoring, Behavior and Safety Analysis of recent Studies. IEEE Trans. Intell. Transp. Syst. 18, 4–24. https://doi.org/10.1109/TITS.2016.2568920.
Sigrist, F., 2023. Mixed Effects Machine Learning with GPBoost for Grouped and Areal Spatial Econometric Data [WWW Document]. Medium. URL https://towardsdatascience.com/mixed-effects-machine-learning-with-gpboost-for-grouped-and-areal-spatial-econometric-data-b26f8bddd385 (accessed 7.31.24).
Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x.
Sun, K., Zhou, R.Z., Kim, J., Hu, Y., 2024. PyGRF: An Improved Python Geographical Random Forest Model and Case Studies in Public Health and Natural Disasters. Trans. GIS n/a. Doi: 10.1111/tgis.13248.
Tang, J., Gao, F., Liu, F., Han, C., Lee, J., 2020. Spatial heterogeneity analysis of macro-level crashes using geographically weighted Poisson quantile regression. Accid. Anal. Prev. 148, 105833. https://doi.org/10.1016/j.aap.2020.105833.
Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018. How is driving volatility related to intersection safety? a Bayesian heterogeneity-based analysis of instrumented vehicles data. Transp. Res. Part C Emerg. Technol. 92, 504–524. https://doi.org/10.1016/j.trc.2018.05.017.
Wang, C., Ijaz, M., Chen, F., Easa, S.M., Zhang, Y., Cheng, J., Zahid, M., 2024a. Temporal assessment of injury severities of two types of pedestrian-vehicle crashes using unobserved-heterogeneity models. J. Transp. Saf. Secur. 16 (8), 820–869.
Wang, C., Xie, Y., Huang, H., Liu, P., 2021. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accid. Anal. Prev. 157, 106157. https://doi.org/10.1016/j.aap.2021.106157.
Wang, C., Abdel-Aty, M., Cui, P., Han, L., 2024b. Effects of helmet usage on moped riders' injury severity in moped-vehicle crashes: Insights from partially temporal constrained random parameters bivariate probit models. Accid. Anal. Prev. 208, 107800.
Wang, C., Abdel-Aty, M., Han, L., 2024c. Effects of speed difference on injury severity of freeway rear-end crashes: Insights from correlated joint random parameters bivariate probit models and temporal instability. Analytic methods in accident research 42, 100320.
Wang, S., Gao, K., Zhang, L., Yu, B., Easa, S.M., 2024. Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US. Accid. Anal. Prev. 199, 107528. https://doi.org/10.1016/j.aap.2024.107528.
Wang, X., Zhang, Q., Yang, X., Pei, Y., Yuan, J., 2023. Traffic safety analysis and model updating for freeways using Bayesian method. J. Transp. Saf. Secur. 15, 737–759. https://doi.org/10.1080/19439962.2022.2128957.
Wang, C., Abdel-Aty, M., Han, L., 2025. Grouped random parameters Poisson-Lindley model with spatial effects addressing crashes at intersections: Insights from visual environment features and spatiotemporal instability. Anal. Methods Accid. Res., 100387
Wen, X., Xie, Y., Jiang, L., Li, Y., Ge, T., 2022. On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. Accid. Anal. Prev. 168, 106617. https://doi.org/10.1016/j.aap.2022.106617.
Wen, X., Xie, Y., Jiang, L., Pu, Z., Ge, T., 2021. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. Transp. Rev. 41, 855–879. https://doi.org/10.1080/01441647.2021.1954108.
Wu, D., Zhang, Y., Xiang, Q., 2024. Geographically weighted random forests for macro-level crash frequency prediction. Accid. Anal. Prev. 194, 107370. https://doi.org/10.1016/j.aap.2023.107370.
Wu, P., Chen, T., Diew Wong, Y., Meng, X., Wang, X., Liu, W., 2023. Exploring key spatio-temporal features of crash risk hot spots on urban road network: a machine learning approach. Transp. Res. Part Policy Pract. 173, 103717. https://doi.org/10.1016/j.tra.2023.103717.
Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Anal. Methods Accid. Res 2, 39–51. https://doi.org/10.1016/j.amar.2014.06.001.
Xue, H., Guo, P., Li, Y., Ma, J., 2024. Integrating visual factors in crash rate analysis at Intersections: an AutoML and SHAP approach towards cycling safety. Accid. Anal. Prev. 200, 107544. https://doi.org/10.1016/j.aap.2024.107544.
Yang, C., Chen, M., Yuan, Q., 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: an exploratory analysis. Accid. Anal. Prev. 158, 106153. https://doi.org/10.1016/j.aap.2021.106153.
Yu, R., Han, L., Zhang, H., 2021. Trajectory data based freeway high-risk events prediction and its influencing factors analyses. Accid. Anal. Prev. 154, 106085. https://doi.org/10.1016/j.aap.2021.106085.
Yu, R., He, Y., Li, H., Li, S., Jian, B., 2024. RiskFormer: Exploring the temporal associations between multi-type aberrant driving events and crash occurrence. Accid. Anal. Prev. 206, 107698. https://doi.org/10.1016/j.aap.2024.107698.
Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119, 274–289. https://doi.org/10.1016/j.aap.2018.07.031.
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.
Zhai, X., Jiang, J., Dejl, A., Rago, A., Guo, F., Toni, F., Sivakumar, A., 2025. Heterogeneous graph neural networks with post-hoc explanations for multi-modal and explainable land use inference. Inf. Fusion 120, 103057. https://doi.org/10.1016/j.inffus.2025.103057.
Zhang, S., Abdel-Aty, M., 2022. Real-time crash potential prediction on freeways using connected vehicle data. Anal. Methods Accid. Res 36, 100239. https://doi.org/10.1016/j.amar.2022.100239.
Zhang, Z., Xu, N., Liu, J., Jones, S., 2024. Exploring spatial heterogeneity in factors associated with injury severity in speeding-related crashes: an integrated machine learning and spatial modeling approach. Accid. Anal. Prev. 206, 107697. https://doi.org/10.1016/j.aap.2024.107697.
Zhao, J., Li, Z., Liu, P., Chen, K., Zhai, X., 2025. Impact of road network patterns on traffic crash prediction: An interpretable geography machine learning approach. Journal of Transportation Safety & Security 1–29.
Zhao, J., Liu, P., Li, Z., 2024. Exploring the impact of trip patterns on spatially aggregated crashes using floating vehicle trajectory data and graph Convolutional Networks. Accid. Anal. Prev. 194, 107340. https://doi.org/10.1016/j.aap.2023.107340.
Zheng, M., Xie, X., Jiang, Y., Shen, Q., Geng, X., Zhao, L., Jia, F., 2024. Optimizing kernel density estimation bandwidth for road traffic accident hazard identification: a case study of the city of London. Sustainability 16, 6969. https://doi.org/10.3390/su16166969.
Zhou, D., Gayah, V.V., Wood, J.S., 2023. Integration of machine learning and statistical models for crash frequency modeling. Transp. Lett. 15, 1408–1419. https://doi.org/10.1080/19427867.2022.2158257.