

## Full Length Article

## Machine learning-based real-time crash risk forecasting for pedestrians

Fizza Hussain<sup>a</sup>, Yuefeng Li<sup>b</sup>, Shimul Md Mazharul Haque<sup>a,\*</sup><sup>a</sup> School of Civil & Environment Engineering, Queensland University of Technology, Brisbane, 4001, Australia<sup>b</sup> School of Computer Science, Queensland University of Technology, Brisbane, 4001, Australia

## ARTICLE INFO

## Keywords:

Artificial intelligence (AI)  
Machine learning  
Extreme value theory  
Pedestrian safety  
Forecasting

## ABSTRACT

Recent developments in artificial intelligence (AI) have made significant improvements in understanding and enhancing pedestrian safety—a vulnerable road user group that receives less attention than motorized road users do. Specifically, AI-based video analytics have provided insight into facilitating real-time safety at signalized intersections. However, past studies have not fully realized the essence of real-time analysis, which underpins forecasting pedestrian collision likelihood by analyzing how past extreme events influence future risk over sequential intervals. To this end, we combine extreme value theory and machine learning models for real-time pedestrian collision risk forecasting. Traffic conflicts and their associated variables were identified from 288 h of video footage obtained from three signalized intersections in Queensland, Australia, via computer vision techniques, including YOLO and DeepSORT, to obtain the post encroachment time for vehicle–pedestrian interactions. A Bayesian non-stationary peak over threshold (POT) is developed to obtain real-time pedestrian crash risk at the signal cycle level. The performance of the POT model is compared with observed crashes, and the results demonstrate the reasonable accuracy of the model. The estimated pedestrian crash risk at each signal cycle forms contiguous univariate time series data (which serve as ground truth), which are used as input to develop time series machine learning models (recurrent neural networks (RNNs) and long short-term memory (LSTM)). Both of these models forecast pedestrian crash risk, with the RNN model outperforming the competing model and demonstrating that pedestrian crash risk can be reliably estimated 30–33 min in advance.

## Nomenclature

Abbreviation	Definition
AI	Artificial intelligence
EVT	Extreme value theory
POT	Peak over threshold
RNN	Recurrent neural network
LSTM	Long short-term memory
PET	Post encroachment time
DIC	Deviance information criterion
MAE	Mean absolute error
RAE	Relative absolute error
MAPE	Mean absolute percentage error

## 1. Introduction

Pedestrians are vulnerable and at high risk of collision associated with vehicle interactions in a road traffic environment. Despite walking being considered a sustainable transport mode and possessing several

health benefits, elevated collision risk can deter pedestrians from walking. To this end, road authorities and communities are putting significant efforts into promoting safer walking, but collision statistics indicate an alarming number of pedestrian injuries despite these measures. For example, approximately 769 pedestrians were involved in fatal vehicle–pedestrian crashes in Australia from 2018 to 2022, accounting for 13% of total crashes (Hughes, 2023). Notably, pedestrian crashes are more prevalent in urban areas, where approximately 33% of these fatal crashes occur (ATC, 2020). Signalized intersections play a critical role in managing and channelizing urban traffic while also prioritizing pedestrian safety. However, these locations are not collision risk-free; for example, pedestrian crashes at intersections account for more than one-third of the total number of crashes (DTMR, 2020). These pedestrian crashes at signalized intersections could be attributed to pedestrian and driver characteristics, speeding, red light violations of pedestrians and drivers, poor visibility, especially during nighttime or heavy rainfall, vehicle types, and vehicle maneuvers (Zeng et al., 2023). Given the multitude of factors involved, understanding pedestrian interactions and collision risk becomes highly important in promoting an active lifestyle.

\* Corresponding author.

E-mail address: [m1.haque@qut.edu.au](mailto:m1.haque@qut.edu.au) (S.M.M. Haque).

Pedestrian safety can be assessed via a reactive or proactive approach. The former approach relies on using crash data and developing statistical models to estimate crash risk on the basis of observed pedestrian crashes (Ashraf et al., 2022; Ihssian and Ismail, 2023; Mukherjee and Mitra, 2020). Past studies on crash-based analysis have identified several issues with police-reported crash data, despite being frequently used, such as an imbalance in crash data, underreporting, misclassification (of injury severity), low sample sizes and means, limited behavioral information, and the need for longer waiting periods to obtain desired samples (Ali et al., 2023a). For example, fatal crashes are severely imbalanced compared with serious or slight crashes (1:90, as reported in Ali et al. (2024)), thus significantly impacting model performance. In contrast, the proactive approach bypasses reliance on that data, instead focusing on traffic conflicts that serve as early indicators of potential crashes. As such, the need for a prolonged waiting period is eliminated, enabling real-time analysis of pedestrian crash risk.

Traditional safety evaluation methods rely on crash data and suffer from the aforementioned problem. Furthermore, the application of traditional methods may be limited because (i) ensuring a consistent crash-to-conflict ratio is challenging, particularly when combining conflicts that differ in severity, and (ii) the statistical correlation between crashes and traffic conflicts scarcely captures the physical dynamics underlying crashes (Zheng et al., 2014). Therefore, extreme value theory (EVT) can estimate the likelihood of rare events via limited observational data, thereby partially fulfilling the initial goal of analyzing traffic conflicts. In recent years, there has been a surge in the development of extreme EVT models and promising results (Ali et al., 2023b; Ankunda et al., 2024; Hussain et al., 2023). While EVT models have been applied for several crash types (with a detailed review presented in Ali et al. (2023a)), this study reviews and summarizes EVT models specifically developed for pedestrian safety (see the comprehensive list in Table 1). Several key observations are summarized below.

First, most studies utilize data from autonomous vehicles to estimate either vehicle safety or network-wide safety. For example, Alozi and Hussein (2022) developed peak-over-threshold models using two autonomous vehicle datasets and reported that the anticipated collision count for autonomous vehicles was between 4 and 5.5 per million vehicle kilometers travelled. In another study, a generalized Pareto model was developed from data from three cities, and pedestrian risk

was found to be a function of interaction and environment type, with Singapore showing a greater proportion of riskier autonomous vehicle–pedestrian interactions than two other cities did (Pittsburgh, Las Vegas, and Boston) (Lanzaro et al., 2023). Similarly, Singh et al. (2024) developed extreme value models and estimated network-wide crash risk. This study compared two extreme value models and reported that block maxima models outperform competing models for two reasons. First, the assumption of independent and identically distributed extremes produced more favorable outcomes for the block maxima method. Second, in the peak-over-threshold approach, the process of threshold selection excluded a considerable proportion of the observations, leaving only approximately one-third of the data available for fitting the generalized Pareto model.

Ankunda et al. (2024) compared two conflict metrics for estimating pedestrian crash risk under heterogeneous traffic conditions. Those studies reported that, relative to the gap time, the post encroachment time has superior performance for estimating pedestrian crash risk, which could be attributed to the constant speed assumption of the gap time. Second, developing advanced extreme value models, Nasserreddine et al. (2023) developed bivariate extreme value models to compare vehicle–pedestrian crash risk between sites of different phasing styles (right-turn flashing yellow arrow vs. permissive circular green arrow), and the results revealed that the bivariate models were better than the competing models.

Finally, only a few studies have analyzed real-time pedestrian crash risk at signalized intersections. Ali et al. (2023b) developed a block maxima model for estimating pedestrian crash risk at signalized intersections via signal cycle covariates related to traffic volume and speed and reported that this model outperforms a stationary model in terms of performance. Extending this analysis to estimate pedestrian crash frequency with severity, Bin Tahir and Haque (2024) developed a bivariate block maxima model for estimating signal cycle crash risk. This study reported different determinants for crash frequency (pedestrian conflicts per signal cycle and average pedestrian speed) and severity (pedestrian conflicts per signal cycle and average vehicle speed per cycle). Except for these two applications of real-time crash risk analysis of pedestrians, the literature is scarce, which is clearly disproportionate to its significance.

Most past real-time crash risk studies have considered the signal cycle as a spatiotemporal unit for estimating the crash risk of shorter time windows (Bin Tahir and Haque, 2024; Hussain et al., 2023). The objective of these studies was to quantify the crash risk of past signal cycles, whereas the proactive safety management objective is twofold: (a) measuring the collision risk associated with previous traffic signal cycles, and (b) estimating the collision risk for upcoming signal cycles. Utilizing past and prevailing traffic conditions, these two aspects are interlinked, as the previous signal cycle's collision risk forms the basis for estimating the future signal cycle's crash risk. Existing studies (Ali et al., 2023b; Bin Tahir and Haque, 2024) have limited their analyses for past signal cycles, which could be attributed to extreme value models' inability to predict crash risk in advance, requiring specialized modeling techniques to be integrated with extreme value models to fully realize proactive safety management, thereby estimating real-time crash risk in the future.

Employing past signal cycle crash risk for future crash risk prediction is a classic time series problem, whereby correlations among contiguous signal cycles can be used to estimate the collision risk of future signal cycles. Machine learning models are, therefore, predominantly used in the literature and have shown excellent performance in forecasting several aspects of traffic safety. For example, a real-time crash prediction study used a support vector machine and Bayesian logistic regression to predict crash occurrence and reported that the former model outperformed the latter model (Yu and Abdel-Aty, 2013). Several other studies have also developed crash prediction models (Basso et al., 2021; Mussone et al., 2017); however, many of these studies focused on using crash data, whereas the use of conflict data and the time trend of crash risk has remained largely unexplored, motivating the current study.

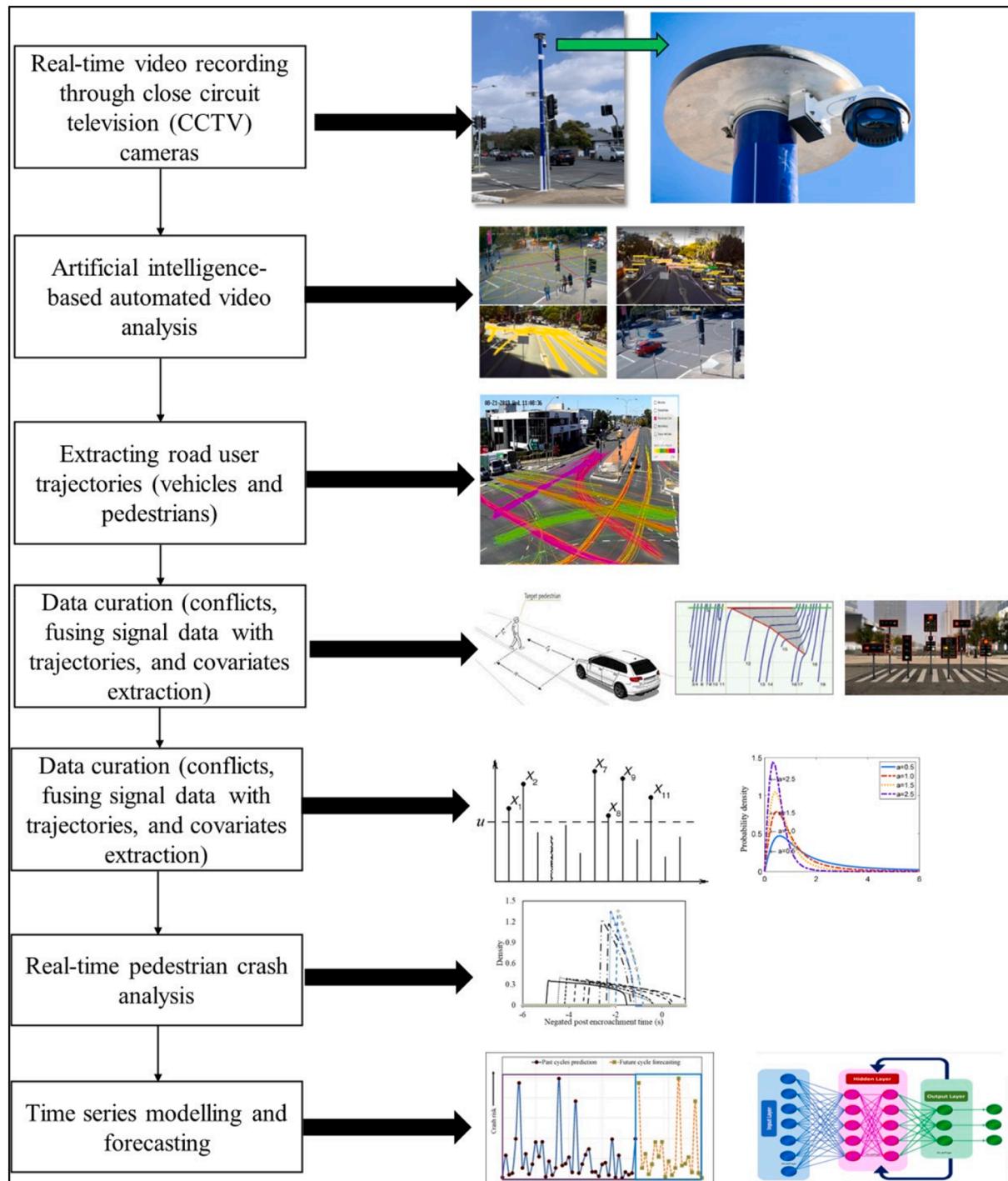
**Table 1**  
Summary of extreme value theory-based pedestrian crash risk models.

Ref.	Real-time (Yes/ No)	Conflict metric	Sampling	Forecasting
Ankunda et al. (2024)	No	PET and gap time	Block maxima and peak over threshold	No
Hewett et al. (2024)	No	PET	Block maxima and peak over threshold	No
Singh et al. (2024)	No	PET	Block maxima and peak over threshold	No
Bin Tahir and Haque (2024)	Yes	PET and Delta-V	Block maxima	No
Ali et al. (2023b)	Yes	PET	Block maxima	No
Arun et al. (2023)	No	TTC	Peak over threshold	No
Hussain et al. (2024b)	No	PET	Unsupervised machine learning	No
Nasserreddine et al. (2023)	No	PET	Block maxima and peak over threshold	No
Alozi and Hussein (2022)	No	TTC and PET	Peak over threshold	No
Lanzaro et al. (2023)	No	PET	Block maxima	No
Guo et al. (2020)	No	PET	Peak over threshold	No
Zhang et al. (2020b)	No	TTC and PET	Peak over threshold	No

Note: PET: post encroachment time; TTC: time-to-collision; Delta-V: predicted post collision change in velocity.

As such, this study develops an integrated model that combines extreme value theory and machine learning to enable proactive estimation of pedestrian collision risk at signalized intersections. The proposed model is evaluated via traffic conflict data extracted from video recordings collected in Brisbane, Queensland, Australia, and makes two contributions. First, focusing on pedestrians—a vulnerable road user group—the integrated model leverages the capabilities of extreme value theory and machine learning, an area for which Mannerling et al. (2020) mentioned that “There is a clear need in the safety field to ground intrinsically predictive models within causal frameworks, while also taking insights from intrinsically predictive models (especially from big

data) to improve upon causal structures through insights from associations involving variables not typically available in traditional safety data. One promising direction for future research would be a hybrid modeling approach of data-driven and statistical methods (with strong consideration of causal elements).” Along this line, an extreme value model quantifies continuous pedestrian collision risk values that exhibit serial correlation over brief intervals, such as signal cycles, but cannot forecast such crash risk in the future. To overcome this shortcoming, time series machine learning models (recurrent neural networks (RNNs) or long short-term memory (LSTM)) are employed to predict pedestrian collision risk by utilizing temporally dependent data derived from



**Fig. 1.** Proposed methodological framework for proactive estimation of pedestrian crash risk.

extreme value modeling. Our integrated model, while novel, lays its foundation for existing studies with some overlap. For example, Howlader and Haque (2025) developed a block maxima model to estimate crash risk, whereas our study develops a peak over threshold model, which has been shown to outperform the former model. Furthermore, using the crash risk from the extreme value model, Howlader and Haque (2025) developed various time series (LSTM and autoregressive integrated moving average models), whereas our study focused on RNNs/LSTMs and demonstrated their efficacy. The variation in performance across different approaches indicates that a one-size-fits-all solution is inadequate, given the distinct crash mechanisms involved in vehicle–pedestrian and opposing-through collisions. This fact underscores the need to develop tailored models for each crash type. Second, while RNNs and LSTMs have been widely compared with each other in several fields, this study compares their performance for pedestrian crash risk forecasting—an underexplored aspect of the crash-conflict literature. Finally, this study also compares two extreme value models, which are rarely compared for real-time pedestrian collision risk prediction in the literature for pedestrian safety, providing an objective evaluation for selecting an appropriate technique for real-time crash risk modeling.

To this end, the rest of the paper is structured as follows. Section 2 details the study methodology outlining model details. Section 3 presents the data collection process and explains AI-based video analytics. The results are presented in Section 4 and extensively discussed in Section 5. Section 6 concludes the study with future research directions.

## 2. Methodology

**Fig. 1** presents the proposed methodology framework of this study. While Sections 2.1 and 2.2 detail each component of the methodology, a brief overview is presented herein. First, vehicle and pedestrian movements are recorded at signalized intersections via closed-circuit television (CCTV) cameras mounted at a height of 6.5 m. These video recordings are then processed through an AI-based video analysis platform (using YOLO for object detection and DeepSORT for vehicle trajectories), converting videos into road user trajectories to identify conflicts. Vehicle–pedestrian interactions are measured by the post encroachment time as a traffic conflict measure. The road user trajectories, conflicts, and loop detector-based signal timing data are then combined to extract the signal cycle-level traffic characteristics (covariates) used for crash risk modeling. For this purpose, a peak-over-threshold approach, corresponding to a generalized Pareto distribution, is employed, whereby covariates are included to account for time-varying crash risk. The developed model is applied to estimate pedestrian collision risk at the signal cycle level, which is then used as input to recurrent neural network/long short-term memory models for crash risk forecasting.

### 2.1. Extreme value model

Consider that  $z_1, z_2, z_3, \dots, z_n$  is an independent and identically distributed random variable, with extreme events  $Z_i$  ( $i = 1, 2, 3, \dots, n$ ) if its value is greater than a prespecified threshold  $u$  (Coles, 2001). With a sufficiently high threshold  $u$ , the threshold excess can be defined with a new variable as  $Y_i = Z_i - u$ . With  $Y > u$ , this newly defined variable  $Y$  follows a generalized Pareto distribution, expressed as

$$G(Y) = 1 - \left(1 + \xi \frac{Y}{\sigma}\right)^{-\frac{1}{\xi}} \quad (1)$$

where  $\sigma (>0)$  refers to the scale parameter and  $\xi$  indicates the shape parameter,  $-\infty < \xi < \infty$ .

Extreme value models suffer from issues such as sample size, non-stationarity, and serial dependency of conflict extremes. Several studies have demonstrated the superiority of peak over threshold models

owing to better utilization of data (Ali et al., 2023b; Nazir et al., 2024), and our study uses this approach to overcome the sample size issue.

As traffic conflicts are non-stationary, this study introduces time-varying covariates to the parameters of a generalized Pareto distribution via an identity link function either in scale (ensuring positivity) or shape parameters. Mathematically, the model can be rewritten as

$$G(Y_{ik} < z | \phi_{ik}, \xi_{ik}) = 1 - \left(1 + \xi_{ik} \frac{Y_{ik}}{\exp(\phi_{ik})}\right)^{-\frac{1}{\xi_{ik}}} \quad (2)$$

where  $Y_{ik}$  indicates the  $k$ th recorded threshold excess at site  $i$ , with  $i = 1, \dots, s$  and  $k = 1, \dots, n$ , with  $n$  representing total threshold excesses;  $\phi_{ik}, \xi_{ik}$  are the two parameters of the generalized Pareto distribution;  $\phi = \log \sigma$  represents the scale parameter transformation to ensure its positiveness, and the covariates are included as

$$\phi_{it} = \beta_{\phi 0} + \beta_\phi X + \epsilon_{\phi i} \quad (3)$$

$$\xi_{ik} = \beta_{\xi 0} + \beta_\xi Y + \epsilon_{\xi i} \quad (4)$$

where  $\beta_{\phi 0}$  indicates the intercept term;  $\beta_\phi$  and  $X$  indicate vectors of estimable parameters and covariates, respectively, for the scale parameter; and  $\epsilon_{\phi i}$  represents additional site-specific heterogeneities that are not directly captured by covariates. Similarly,  $\beta_{\xi 0}$  indicates the intercept term, and  $\beta_\xi$  and  $Y$  indicate vectors of estimable parameters and covariates, respectively, for the shape parameter. Note that the covariates used for the model are explained in Section 3.

The model specified above can be estimated via a Bayesian estimation procedure, whereby the latent process is characterized by priors, allowing estimation of posterior model parameter distributions. In this study, each parameter ( $\beta_{\phi 0}, \beta_\phi, \xi$ ) is assumed to be independent of each other, and owing to the unavailability of information on how these parameters are associated with the generalized Pareto distribution, uninformative priors are used for the scale parameters, with a normal distribution  $N(0, 1 \times 10^6)$ . Information from previous studies is utilized for the shape parameter to properly identify the prior values, as this parameter is sensitive and causes convergence issues (Coles, 2001). Past studies (Nazir et al., 2024; Songchitruksa and Tarko, 2006) have reported shape parameter values between  $-1$  and  $1$ ; therefore, its priors are defined as a normal distribution  $N(0, 0.25)$ . Note that Markov chain Monte Carlo simulation with Gibbs sampling is used to obtain posterior distributions of model parameters. Furthermore, several Bayesian peak-over-threshold models have been developed with different covariate combinations for distributional parameters (scale and shape). The deviance information criterion (DIC) (Spiegelhalter et al., 2002) serves as a metric for model selection in this study, with preference given to models exhibiting lower DIC values.

The final issue concerns serial dependence in extreme events, which are typically assumed to be independently and identically distributed. This assumption might be valid in theory, but it often fails to hold true in real-world applications. The challenge of dependence is addressed through a declustering approach that isolates dependent data points, resulting in a collection of threshold exceedances that are nearly independent (Coles, 2001). This approach determines the shortest time interval required between consecutive threshold exceedances to consider them independent. Naturally, the signal cycle level offers a unique separation time for identifying independent extremes, as this study utilizes data from signalized intersections where signal cycle-related parameters are likely to influence traffic extremes (Essa and Sayed, 2018). Therefore, this study uses a signal cycle to select only one maximum excess from each cycle used for modeling.

The peak-over-threshold method involves identifying a threshold for separating extremes from normal interactions. Past studies (Arun et al., 2023; Zheng and Sayed, 2019) have adopted quantile regression for this purpose. The use of quantile regression is motivated by findings from Coles (2001), which suggest that the threshold may be influenced by

underlying covariates. For example, traffic volume and composition have been shown to contribute to heterogeneity in post encroachment time (Zheng et al., 2016), thereby necessitating a generalized modeling approach that links thresholds to relevant covariates.

Quantile regression determines the impact of covariates on the quantile of the dependent variable, which is largely different from conventional linear regression, where covariates are found to impact the conditional mean of the dependent variable. Following the basic analogy of threshold excess (a high enough threshold), some high quantiles are likely to ensure a high threshold, thereby making it suitable for estimating peak over threshold models. As such, quantile regression establishes a relationship between the dependent variable and specific quantiles within the interval  $\tau \in (0, 1)$ . For any chosen quantile, the conditional quantile function, denoted as  $q_\tau$ , can be expressed as

$$Q(\tau|X = x) = x'_\tau \beta^\tau \quad (5)$$

where  $Q(\tau|X = x)$  represents the conditional  $\tau$ th quantile of the dependent variable  $q_\tau$  and  $\beta^\tau$  indicates the vector of estimable parameters associated with a vector of covariates,  $x_\tau$ . Model parameters for quantile regression are usually obtained through the following optimization process (Koenker and Bassett, 1978):

$$\min_{\beta \in R^p} \left[ \sum_{\tau: q_\tau > x'_\tau \beta^\tau} \tau |q_\tau - x'_\tau \beta^\tau| + \sum_{\tau: q_\tau < x'_\tau \beta^\tau} (1 - \tau) |q_\tau - x'_\tau \beta^\tau| \right] \quad (6)$$

where the model parameters ( $\beta$ ) are a set of  $p$  rational regression factors minimizing the objective function.

The next step after quantile regression is to obtain a high quantile leading to an appropriate threshold, often found via a threshold stability plot (Coles, 2001). If the generalized Pareto distribution with two parameters  $(\sigma, \xi)$  is valid for excess of the threshold  $u$ , then excesses for a higher threshold  $u$  must follow a generalized Pareto distribution with parameters  $(\sigma_u, \xi)$  with the same shape parameter and modified scale parameter as

$$\sigma_u = \sigma_0 + \xi(u - u_0) \quad (7)$$

As such, the reparametrised scale  $(\sigma^* = \sigma_u - \xi u)$  remains constant with respect to  $u$  following Eq. (7). Hence, estimates of  $\sigma^*$  and  $\xi$  should be constant over the threshold  $u_0$  if  $u_0$  is valid for the excess threshold following the generalized Pareto distribution. Note that the scale parameter estimates of different quantiles will not be exactly constant but rather stable after allowing sampling errors (Zheng and Sayed, 2019). As such, the threshold stability plot can be used to obtain a near-constant region for threshold determination.

To validate the peak-over-threshold model, the crash frequency estimated by the model is compared with historical crash records. For this purpose, the risk of a pedestrian crash can be calculated from the fitted generalized Pareto distribution as

$$R_i = \Pr(Z \geq 0) = 1 - (Z < 0) = 1 - G(Y) \quad (8)$$

where  $R_i$  is the risk of a pedestrian crash for a given extreme  $i$ ,  $Z$  is the negated post encroachment time. Note that negated post encroachment time is used for simplicity, as a post encroachment time greater than or equal to 0 reflects a vehicle–pedestrian crash, as the trajectories of these two road users overlap. With the traffic conflict observation period being  $t$  and representative of a long period  $T$  (in this study, it is five years), the estimated number of crashes,  $N_T$ , can be obtained as

$$N_T = \frac{T}{t} \sum_{i=1}^k R_i \quad (9)$$

where  $k$  represents the total number of extremes. To assess the uncertainty in the model's predicted crash counts, confidence intervals for the average estimated crashes are derived from a simulation study involving

30,000 iterations. Using mean parameter estimates and standard deviations, model parameter distributions are generated, which are employed to determine the 95% confidence intervals for the crashes estimated by the model. To facilitate comparison, Poisson confidence intervals for the observed crash data are calculated via the methodology outlined by Songchitruksa and Tarko (2006), whereby confidence intervals using the true mean,  $\lambda$ , can be obtained as  $\left[ \lambda : \frac{1}{2n} \chi^2_{y_0, 0.975} \leq \lambda \leq \frac{1}{2n} \chi^2_{(y_0+1), 0.025} \right]$ , where  $n = 5$  years and  $y_0$  represents the total number of crashes observed in 5 years for the intersections.

## 2.2. Machine learning models for forecasting crash risk

The developed peak-over-threshold model facilitates the estimation of pedestrian crash risk values at smaller time intervals (signal cycles), forming time series data. For modeling such data, recurrent neural networks are typically used because of their ability to capture serial dependency in time series or sequenced data. Unlike conventional feed-forward neural networks, such as artificial neural networks, RNNs tend to retain information within their hidden state over time, thereby allowing them to mimic patterns in sequence and assisting in predicting the next value of the sequence. The RNNs can be expressed as

$$h(t) = f_H(W_x x(t) + W_h h(t-1)) \quad (10)$$

$$y(t) = f_o(W_y h(t)) \quad (11)$$

where  $x(t)$  and  $y(t)$  are the input and output (pedestrian crash risk) at time  $t$ , respectively, which is the pedestrian crash risk at the next time interval  $t+1$ ;  $h(t)$  is the hidden state of the network at time  $t$ ;  $f_o$  and  $f_H$  indicate the hidden layer and output layer activation functions, respectively; and  $W_x$ ,  $W_h$ , and  $W_y$  are the weight matrices of connections.

RNNs are often found to suffer from vanishing gradient problems, corresponding to the inability to learn typically long data sequences (Ali et al., 2024; Rowan et al., 2025). This problem is addressed via long short-term memory (LSTM) networks, which are a sophisticated extension of traditional RNNs and a powerful method that introduces a memory cell for forgetting certain information. Memory cells are often embedded in hidden layers, serving as memory blocks rather than typical neuron nodes (Olah, 2015). Several self-connected memory cells are found in each block with multiplicative units of input, output, and forget gates, providing continuous information on writing, reading, and resetting operations on the cells.

Notably, our choice of RNN/LSTM was guided by three main considerations. (1) Task requirements and data size: The dataset used in this study is relatively small and sequential in nature, where LSTMs have a well-documented track record of strong performance (see some past studies along the same lines, Hussain et al. (2024a); Howlader and Haque (2025)). Transformer-based methods, while powerful, generally require substantially larger datasets to avoid overfitting and to fully leverage their self-attention mechanisms (Ali et al., 2024). (2) Model interpretability and proven stability: In the traffic safety domain, interpretability and temporal feature analysis are crucial for understanding underlying crash mechanisms. LSTMs provide more straightforward temporal gating mechanisms, making it easier to explain the influence of past states on predictions—an essential factor for our application. (3) Computational efficiency and reproducibility: RNN/LSTM architectures offer lower computational costs than transformer models do, enabling faster training and easier deployment on standard hardware.

In this study, RNNs and long short-term memory (LSTM) are used for pedestrian crash risk prediction by capturing the temporal correlation among crash risks into consecutive signal cycles, forming the basis for pedestrian crash risk forecasting for future signal cycles.

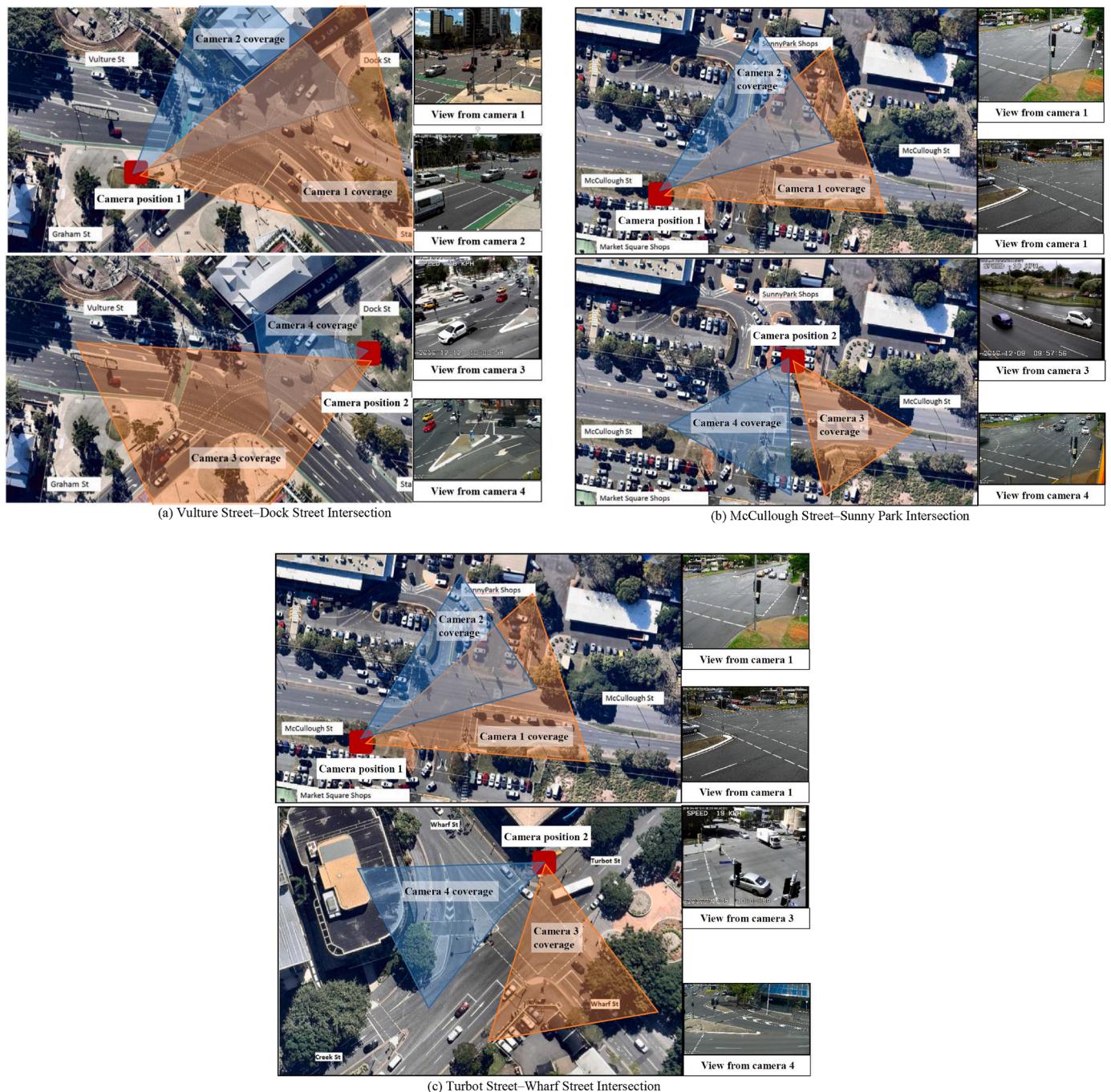
For each signalized intersection, two typical weekdays of video data

are recorded only in the daytime (6 a.m.–6 p.m.), leading to non-continuous data from different days across different intersections. As such, as an input to RNN/LSTM, the combined data are not used, which violates the regularly spaced time interval assumption of RNN/LSTM models. Similarly, several RNN and LSTM models have been developed for each intersection and each day separately via different combinations of activation functions, optimizers, and numbers of neurons. To select the best model, the mean absolute error (MAE) and relative absolute error are used to quantify the error associated with model predictions. The best RNN model from different hyperparameter combinations is obtained using the smallest MAE and RAE values, which consists of two layers, namely, the recurrent layer with 60 neurons and the connected output layer, yielding pedestrian crash risk. The rectified linear unit activation function and Adam optimizer are used for the back-

propagation process, with an input sequence length of 10. An LSTM model with a two-layered architecture is developed for comparison. Note that the activation function, optimizer, number of neurons and sequence length are the same as those in the RNN model.

### 3. Data collection and AI-based video analytics

For this study, video footage capturing traffic flow was collected from three signalized intersections. These intersections are located in Brisbane, Queensland, Australia (see names of these intersections in Table 4). As Fig. 2 shows some overlapping areas between cameras at intersections, step-by-step postprocessing was performed to avoid duplications. First, trajectories found in both cameras were identified and sorted on the basis of their length. A trajectory with a longer time in a



**Fig. 2.** Study sites, camera positions, and camera coverages.

given camera was selected, providing more information, which was manually verified. Second, conflicts obtained from multiple cameras were meticulously checked, and duplicate conflicts were omitted. These intersections possess high pedestrian volume (see Table 2) and represent risky vehicle–pedestrian interactions. For each intersection, two overhead cameras were mounted on top of the trailer mast—approximately 6.5 m high—to capture all vehicle–pedestrian interactions. For the Vulture Street intersection, data were collected over four days, whereas for the other two intersections, data were collected over two different days, resulting in 288 h of video data.

At the study intersections, actuated traffic signals are implemented, leading to variations in the signal cycle length over the course of the day in response to changing traffic demand. As such, signal cycle data, which include information on signal cycle durations, are integrated with raw video data processed via the Advanced Mobility Analytics Group's (AMAG) automated video analysis platform (AMAG, 2024), as detailed below.

Leveraging AI embedded into AMAG's automated platform, which utilizes computer vision techniques, conflicts are automatically obtained via four steps, namely, camera calibration and validation, object detection and tracking, trajectory generation, and traffic conflict extraction (Fig. 3). During the camera calibration, three-dimensional video footage was converted to two-dimensional image space, whereby road users are monitored as separate images and each is mapped according to their real-world positions. Road users in the platform are detected through the YOLO (“You only look once”) algorithm (Yang and Deng, 2020), and user trajectories are obtained through the Deep ORT (“Simple online and real-time tracking”) algorithm (Wojke et al., 2017). These detections and tracking algorithms rely on trained convolutional neural networks to deliver precise and reliable results. Several studies have demonstrated and confirmed the high accuracy of road user detection through YOLO (Feng et al., 2020); hence, YOLO was adopted in this study. Similarly, DeepSORT has also shown high accuracy in generating trajectories of several objects detected in subsequent video frames through the YOLO algorithm (Chan and Suandi, 2019; Zhang et al., 2020a), whereby linking Euclidean coordinates generates continuous trajectories of road users. These trajectories are then used to compute the conflict between road users, which is the post encroachment time in this context, as described below.

The post encroachment time (PET) is “the time difference between the moment an “offending” vehicle passes out of the area of a potential collision and the moment of arrival at the potential collision point by the “offended” vehicle possessing the right-of-way” (Allen et al., 1978). Considering interacting road users as pedestrians and turning vehicles, the PET for vehicle–pedestrian interactions can be obtained as

$$\text{PET} = t_2 - t_1 \quad (12)$$

**Table 2**  
Summary of road user volumes.

Intersection	Vehicle volume (av. per hour)				Pedestrian	
	Direction	Left	Through	Right	Crossing #	Volume (av. per hour)
Vulture	South-East	133	957	162	1	120
	North-East	69	17	80	2	170
	West	44	1064	73	3	135
	South-West	102	8	134	4	223
McCullough	South	60	11	69	1	5
	East	95	861	60	2	21
	North	84	11	59	3	20
	West	69	750	54	4	39
Turbot	South	60	11	69	1	5
	East	95	861	60	2	21
	North	84	11	59	3	20
	West	69	750	54	4	39

where  $t_1$  and  $t_2$  indicate the moment pedestrian exits the encroachment or conflict area and the time at which the vehicle enters that same zone, respectively.

Several covariates, including the proportion of turning vehicles per cycle, pedestrian count per cycle, average speed of vehicles in a cycle and average speed of pedestrians in a cycle, were calculated via road user trajectories and signal cycle timing, facilitating post hoc real-time analysis. A detailed statistical summary of covariates and traffic conflicts can be found in Ali et al. (2023b). Five years of crash data sourced from the Department of Transport and Main Roads(TMR), Queensland, are utilized to validate the model. Several filters were applied to the original crash dataset to isolate pedestrian crashes exclusively. These included criteria based on the crash timing, location, severity, details about the road users involved, and the specific type of collision.

## 4. Results

This section first outlines the findings from the quantile regression model used for threshold determination, then proceeds to the estimation results of the peak-over-threshold model and finally covers real-time crash risk prediction by employing a recurrent neural network. Detailed explanations of each model are provided in Sections 4.1 and 4.2.

### 4.1. Quantile regression for threshold determination

Quantile regression models are developed from 80% to 95% at a regular interval of 2.5% with four covariates. Table 3 summarizes the modeling results, and it is evident that the proportion of turning vehicles per cycle, pedestrian count per cycle, and average turning vehicle speed are statistically significant in the model at the 5% significance level, whereas the average pedestrian speed is not statistically significant. For all quantiles, all model parameters have positive signs, which has implications for extreme conflicts. For example, the quantile regression results indicate that an increase in turning vehicle count, speed, or pedestrian speed leads to more extreme conflicts.

For each quantile, several non-stationary peak-over-threshold models are estimated (see more details about model estimation and related intricacies), and the deviance information criterion is used to choose the best model. The scale and shape parameters of the best-fit model are then used to develop a threshold stability plot, as shown in Fig. 4. It appears that the shape parameter remains stable from 87.5% to 90%, whereas the scale parameter also appears to be stable in the same range. As such, the final threshold is 87.5%, and the corresponding values are then used to fit the peak-over-threshold model described in Section 4.2.

### 4.2. Peak-over-threshold model estimation results

Several Bayesian peak-over-threshold models are estimated with the negated post encroachment time as the dependent variable and a combination of covariates. All the models are estimated to have two chains and different starting values, with a total number of iterations of 50,000, with the first 40,000 iterations considered burn-in for model convergence and the next 10,000 considered posterior distributions. Model convergence is assessed through trace plots for each model parameter, revealing that both chains are well mixed, which is further confirmed by the Gelman-Rubin statistic of two chains for each parameter, with the value being smaller than 1.1, indicating model convergence.

Table 4 presents the Bayesian peak over threshold model estimates for pedestrian conflicts. Several alternative models were also developed by varying the combinations of covariates in the scale and shape parameters; however, none demonstrated a statistically better fit. Throughout the model development process, multiple covariates were evaluated, with the final model selection guided by two key criteria: (a) The parameters should be statistically significant (vehicle volume

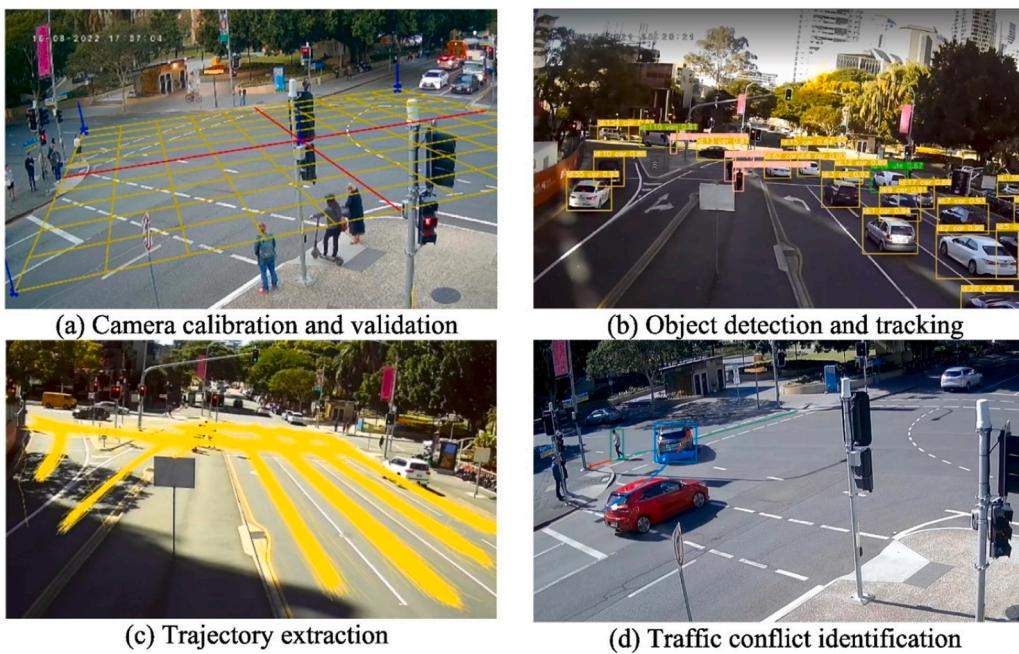


Fig. 3. AI-based automated video analytic process.

**Table 3**

Quantile regression model estimation summary.

Parameter	Quantile	Estimate	Standard error	z-statistic	p value	95% confidence interval	
						Lower	Upper
Proportion of turning vehicles	80%	0.2947	0.1070	2.7552	0.0059	0.0851	0.5044
Pedestrian count		0.1635	0.0167	9.7898	< 0.001	0.1308	0.1962
Turning vehicle speed		0.3165	0.1032	3.0659	0.0022	0.1141	0.5188
Pedestrian speed		-0.1446	0.3322	-0.4353	0.3317	-0.7957	0.5065
Proportion of turning vehicles	82.5%	0.2713	0.1148	2.3644	0.0181	0.0464	0.4962
Pedestrian count		0.1602	0.0178	8.9888	< 0.001	0.1253	0.1951
Turning vehicle speed		0.2445	0.1115	2.1921	0.0284	0.0259	0.4631
Pedestrian speed		-0.7460	0.6532	-1.1420	0.1267	-2.0263	0.5343
Proportion of turning vehicles	85%	0.2311	0.1109	2.0839	0.0712	0.0137	0.4485
Pedestrian count		0.1484	0.0200	7.4296	< 0.001	0.1092	0.1875
Turning vehicle speed		0.2137	0.1055	2.0244	0.021	0.0068	0.4205
Pedestrian speed		-0.2587	0.4235	-0.6109	0.2709	-1.0888	0.5714
Proportion of turning vehicles	87.5%	0.1178	0.0454	2.5959	0.0047	0.0288	0.2067
Pedestrian count		0.1351	0.0215	6.2914	< 0.001	0.0930	0.1772
Turning vehicle speed		0.1320	0.0622	2.1219	0.0169	0.0101	0.2539
Pedestrian speed		-0.1440	0.7890	-0.1825	0.4277	-1.6904	1.4024
Proportion of turning vehicles	90%	0.0800	0.0333	2.4018	0.0081	0.0147	0.1452
Pedestrian count		0.1229	0.0262	4.6979	< 0.001	0.0716	0.1741
Turning vehicle speed		0.1245	0.0450	2.7647	< 0.001	0.0362	0.2128
Pedestrian speed		-0.2789	0.2569	-1.0856	0.1389	-0.7824	0.2246
Proportion of turning vehicles	92.5%	0.0445	0.0158	2.8146	0.0024	0.0135	0.0754
Pedestrian count		0.1065	0.0267	3.9932	0.0001	0.0542	0.1587
Turning vehicle speed		0.0699	0.0170	4.1124	< 0.001	0.0366	0.1032
Pedestrian speed		-0.1789	0.1557	-1.1490	0.1252	-0.4841	0.1263
Proportion of turning vehicles	95%	1.0149	0.0795	12.7600	< 0.001	0.8590	1.1708
Pedestrian count		0.1971	0.0101	19.5934	< 0.001	0.1774	0.2168
Turning vehicle speed		1.1871	0.0641	18.5341	< 0.001	1.0616	1.3126
Pedestrian speed		-0.4589	0.3789	-1.2111	0.1129	-1.2015	0.2837

should not be significant, whereas pedestrian volume should be significant), and (b) model fit should improve. For example, we tested several geometric design characteristics, such as lane configurations and the number of lanes, but these parameters were not retained in the parsimonious model because of the lack of statistical significance, which could be attributed to small intersection numbers and limited variability.

This study also developed a stationary model, but the result of this model is not presented, as non-stationarity models are found to

consistently outperform stationarity models in past studies and in this study. Table 4 indicates that two and one parameters are significant for the scale and shape parameters, respectively. Note that the statistical significance of the model parameters in Bayesian estimation is evaluated via Bayesian credible intervals. The deviance information criterion for the model is 241.43, which is relatively smaller than those of the other estimated models. For example, a model that includes pedestrian volume, the proportion of turning vehicles, and pedestrian speed in the scale parameter and turning vehicle speed in the shape parameter yields

**Table 4**  
Bayesian peak-over-threshold model estimation summary.

Parameter	Description	Mean	Standard deviation	Bayesian credible intervals	
				2.5%	97.5%
$\mu_0$	Scale (Intercept)	0.5291	0.0463	0.216	0.4653
$\mu_{\text{Prop\_Turn}}$	Scale (Proportion of turning vehicles)	0.0679	0.0121	0.041	0.1009
$\mu_{\text{Ped vol}}$	Scale (Pedestrian volume)	0.0083	0.0032	0.0116	0.0193
$\xi_0$	Shape (Intercept)	-0.1723	0.0562	-0.4849	-0.2634
$\xi_{\text{Speed\_Turn}}$	Shape (Speed of turning vehicles)	0.0057	0.0045	0.0120	0.0015

Deviance information criterion = 215.7  
Widely Applicable Information Criterion [WAIC] = 213.2  
Leave-one-out cross-validation [LOO-CV] = 214.1  
5-year observed crashes (95% confidence intervals) = 7 (2.81, 14.78)  
Model estimated 5-year crashes (95% confidence intervals) = 9.3 (2.7, 16.21)

a deviance information criterion (DIC) value of 257.23. Notably, pedestrian speed did not emerge as a statistically significant factor within the model.

The parameter for the proportion of turning vehicles that interact with pedestrians is significant in the model and is positively associated with pedestrian crash risk. This relationship suggests that pedestrian crash risk increases when the proportion of turning vehicles increases. Given that pedestrians are right-of-way and that vehicles are required to wait for pedestrians to clear the intersection region, a greater number of vehicles passing through the intersection in the same signal cycle implies that vehicles immediately start moving as soon as pedestrians leave the encroachment zone, thereby increasing their crash risk. Another plausible reason for this association could be group and dispersed movements. Ali et al. (2023b) reported that pedestrians at signalized intersections tend to move in groups, which may move together or at a distance from each other. This movement pattern increases vehicle waiting times, as one pedestrian group is likely to clear the intersection, whereas another pedestrian group may still be crossing the intersection. In such situations, vehicles often seek gaps between groups, significantly increasing pedestrian crash risk.

Pedestrian volume is another key parameter in the model and is positively associated with pedestrian crash risk. An increased pedestrian count suggests a longer clearance time for pedestrians, which increases vehicle waiting times. Longer waiting times have been shown to significantly affect driving behavior, and in turn, drivers tend to make

risky decisions. A study on merging behavior revealed that drivers rejecting gaps inherently increase their waiting times, thereby resulting in the selection of gaps that are smaller than the critical gaps, negating gap acceptance theory (Marczak et al., 2013).

The model also considers the speed of turning vehicles as a significant parameter that is positively related to pedestrian crash risk. Higher vehicle speeds increase the risk to pedestrians, as drivers must apply abrupt braking to yield, particularly in contexts where pedestrians have the right-of-way. Several studies have confirmed the positive association between hard braking and increased crash risk for pedestrians (Ali et al., 2022).

To estimate the vehicle–pedestrian crash frequency, the peak-over-threshold model is applied, following the approach outlined in Section 2. The resulting crash estimates, along with their corresponding confidence intervals, are summarized in Table 4. The mean number of estimated crashes is 9.3, as opposed to the 7 observed crashes. The estimated 95% confidence intervals (lower and upper) are 3.7 and 16.21, whereas the observed confidence intervals are 2.81 and 14.78. The range, calculated as the difference between the upper and lower bounds of the confidence interval, is calculated to compare the model-predicted crash estimates and historical crashes. The observed range is 11.97 (14.78–2.81), whereas this range for the model-estimated crashes is 12.51. Notably, the mean number of crashes and upper bounds of the peak-over-threshold model are within the observed confidence intervals, suggesting good performance of the model.

Following the evaluation of the model's performance, it is subsequently applied to estimate pedestrian crash risk at the level of individual signal cycles. Using signal cycle level covariates, generalized Pareto distributions for each signal cycle for different intersections are generated and presented in Fig. 5. In this context, the form of the generalized Pareto distribution is particularly important, as the tail portion of the distribution depicts the riskiness of a signal cycle. Specifically, a signal cycle is considered risky if the tail of the distribution for the same cycle surpasses the point where PET = 0, thereby indicating a positive pedestrian crash risk.

Fig. 5 displays the generalized Pareto distributions generated for a sample of five typical signal cycles of each of the three intersections. The red dotted line shows the threshold for the negative post encroachment time. Similarly, the red signal cycle number indicates signal cycles with positive pedestrian crash risk, whereas black signal cycle numbers demonstrate safe cycles. Notably, 2, 2, and 3 signal cycles are risky for Vulture Street, McCullough Street, and Turbot Street intersections, respectively, as their tails end beyond the 0 s point.

#### 4.3. Machine learning models for crash risk forecasting

From the fitted generalized Pareto distributions, each signal cycle crash risk is estimated via the methodology described in Section 2.1. Sequences of crash risk across consecutive signal cycles are employed as inputs to recurrent neural network (RNN) and long short-term memory (LSTM) models, which are then used to forecast pedestrian crash risk in upcoming cycles. Model training was conducted using 70% of the dataset, with 20% allocated for validation and the remaining 10% reserved for forecasting. To mitigate the risk of overfitting, early stopping is implemented by tracking the loss between the training and validation sets. Training is terminated when the error metric, measured by the mean absolute error, fails to improve over successive epochs. The training vs validation loss plot (not presented herein for brevity) indicates that both curves are close to each other and that the distance between them does not increase with increasing number of epochs, implying no overfitting. After ensuring that the models are free from overfitting, they are used for pedestrian crash risk forecasting, and model performance is assessed via three metrics: the mean absolute error (MAE), mean absolute percentage error (MAPE), and relative absolute error (RAE), with the latter defined as the ratio between the MAE and the average crash risk estimated through the peak-over-threshold

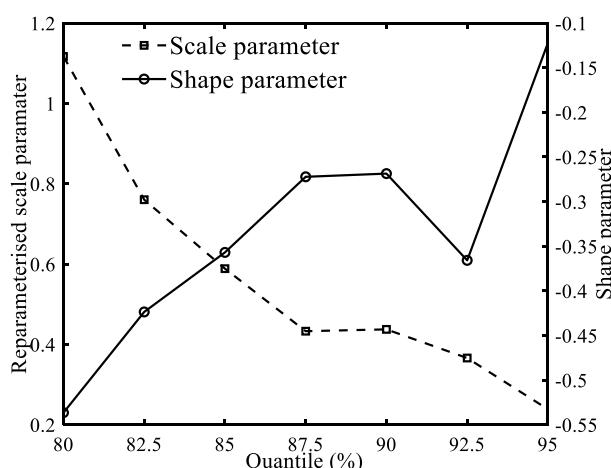
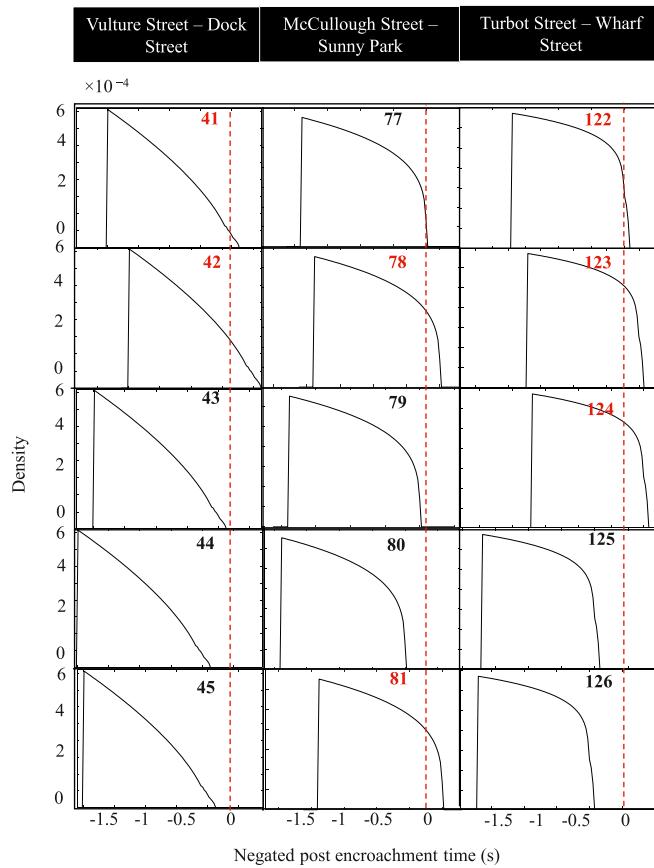


Fig. 4. Schematic of threshold stability plots.



**Fig. 5.** Signal cycle-level generalized Pareto distribution for real-time pedestrian crash risk analysis.

model. Table 5 presents the forecasting performance of the models, and it can be noted that daywise models are developed for the reasons mentioned in Section 2.2. A typical illustration of model performance for a few signal cycles for the McCullough Street intersection is presented in Fig. 6.

Table 5 suggests that the RNN and LSTM models can predict pedestrian crash risk for the immediate future cycles well. For example, the RAEs for the RNN model for the Vulture Street intersection for 4 days are  $7.33 \times 10^{-2}$ ,  $1.33 \times 10^{-1}$ ,  $1.21 \times 10^{-1}$ , and  $1.10 \times 10^{-2}$ , whereas the mean RAE for this intersection is  $8.46 \times 10^{-2}$ .

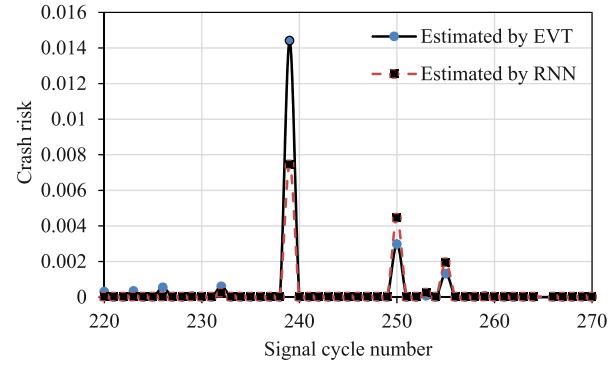
Furthermore, given that RNN/LSTM models are likely to be used for real-time applications, their computational performances are also assessed and compared. The results indicate that the RNN model is relatively faster than the LSTM model because of its simple architecture.

**Table 5**

Forecasting performance of the machine learning models.

Intersection	Day	RNN			LSTM		
		MAE	RAE	MAPE	MAE	RAE	MAPE
Vulture Street	1	$1.79 \times 10^{-4}$	$7.33 \times 10^{-2}$	7.03	$1.87 \times 10^{-4}$	$7.33 \times 10^{-2}$	7.11
	2	$7.11 \times 10^{-4}$	$1.33 \times 10^{-1}$	9.33	$7.89 \times 10^{-4}$	$1.42 \times 10^{-1}$	9.45
	3	$2.99 \times 10^{-4}$	$1.21 \times 10^{-1}$	7.78	$3.05 \times 10^{-4}$	$1.37 \times 10^{-1}$	7.91
	4	$1.13 \times 10^{-3}$	$1.10 \times 10^{-2}$	6.91	$1.78 \times 10^{-3}$	$1.28 \times 10^{-2}$	7.01
	Mean	$5.80 \times 10^{-4}$	$8.46 \times 10^{-2}$	7.76	$7.65 \times 10^{-4}$	$9.13 \times 10^{-2}$	7.87
McCullough Street	1	$1.01 \times 10^{-4}$	$1.81 \times 10^{-3}$	6.77	$1.02 \times 10^{-4}$	$1.93 \times 10^{-3}$	6.79
	2	$1.31 \times 10^{-4}$	$6.91 \times 10^{-3}$	7.02	$1.36 \times 10^{-4}$	$6.96 \times 10^{-3}$	7.05
	Mean	$1.16 \times 10^{-4}$	$4.36 \times 10^{-3}$	6.90	$1.19 \times 10^{-4}$	$4.45 \times 10^{-3}$	6.92
Turbot Street	1	$2.38 \times 10^{-4}$	$5.93 \times 10^{-2}$	7.35	$2.48 \times 10^{-4}$	$5.99 \times 10^{-2}$	7.45
	2	$1.33 \times 10^{-4}$	$1.23 \times 10^{-3}$	7.03	$1.47 \times 10^{-4}$	$1.45 \times 10^{-3}$	7.11
	Mean	$1.86 \times 10^{-4}$	$3.03 \times 10^{-2}$	7.19	$1.98 \times 10^{-4}$	$3.07 \times 10^{-2}$	7.28

Note: RNN: recurrent neural network; LSTM: long short-term memory; MAE: mean absolute error; RAE: relative absolute error; MAPE: mean absolute percentage error.



**Fig. 6.** Pedestrian crash risk prediction for the McCullough Street intersection.

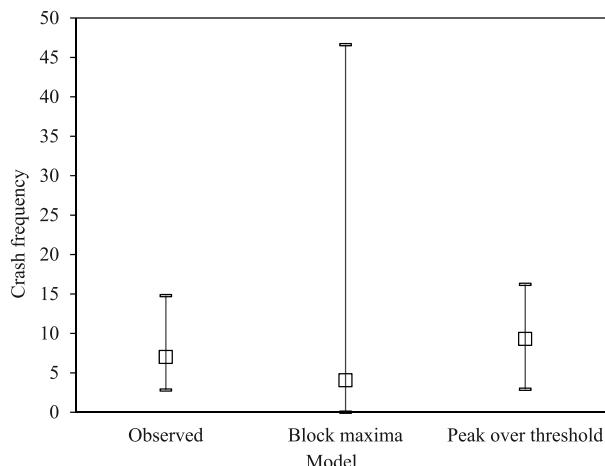
For example, for one day of Vulture Street intersection, the RNN model takes approximately 7 min, whereas the LSTM model converges in 10.5 min. Similar results are also found for other days and intersections.

## 5. Discussion

This study developed a pedestrian crash risk forecasting model by leveraging extreme value models and machine learning algorithms. Section 5.1 compares both of these models with their counterparts.

### 5.1. Model comparisons

This study developed a peak-over-threshold model for signal-cycle level crash risk estimation. To shed more light on the performance of the peak-over-threshold model, a subsequent analysis is performed to compare the performances of the peak-over-threshold method and the block maxima approach. In the literature, it is common to evaluate real-time extreme value models by comparing predicted crashes with aggregated crash counts. Nonetheless, these models also need to be examined at the signal cycle scale. An important avenue for future research would be to validate cycle-level crash risk estimates via detailed crash data across a broader range of traffic sites. Fig. 7 compares pedestrian crash frequencies estimated through the block maxima (where the block is considered a signal cycle following earlier studies, Ali et al. (2023b) and Bin Tahir and Haque (2024)) and peak over threshold models. The confidence intervals associated with the block maxima model are considerably wider than those associated with the peak-over-threshold model, indicating greater uncertainty in the predictions generated by the former. Ankunda et al. (2024) compared block maxima and peak over threshold models for pedestrian crash frequency for heterogeneous traffic and reported that the latter model outperformed the former model because of a smaller threshold, resulting in true conflict extremes that can mimic the crash mechanism. Similarly, a



**Fig. 7.** Comparison of pedestrian crash risk prediction methods.

study (Hussain et al., 2022) demonstrated that true conflict events detected by anomaly detection techniques lead to significantly better performance of extreme value models, and peak-over-threshold models are likely to benefit more than block maxima models are. Our comparative analysis further confirms the superior efficacy of the peak-over-threshold approach in accurately estimating the crash frequency for pedestrians.

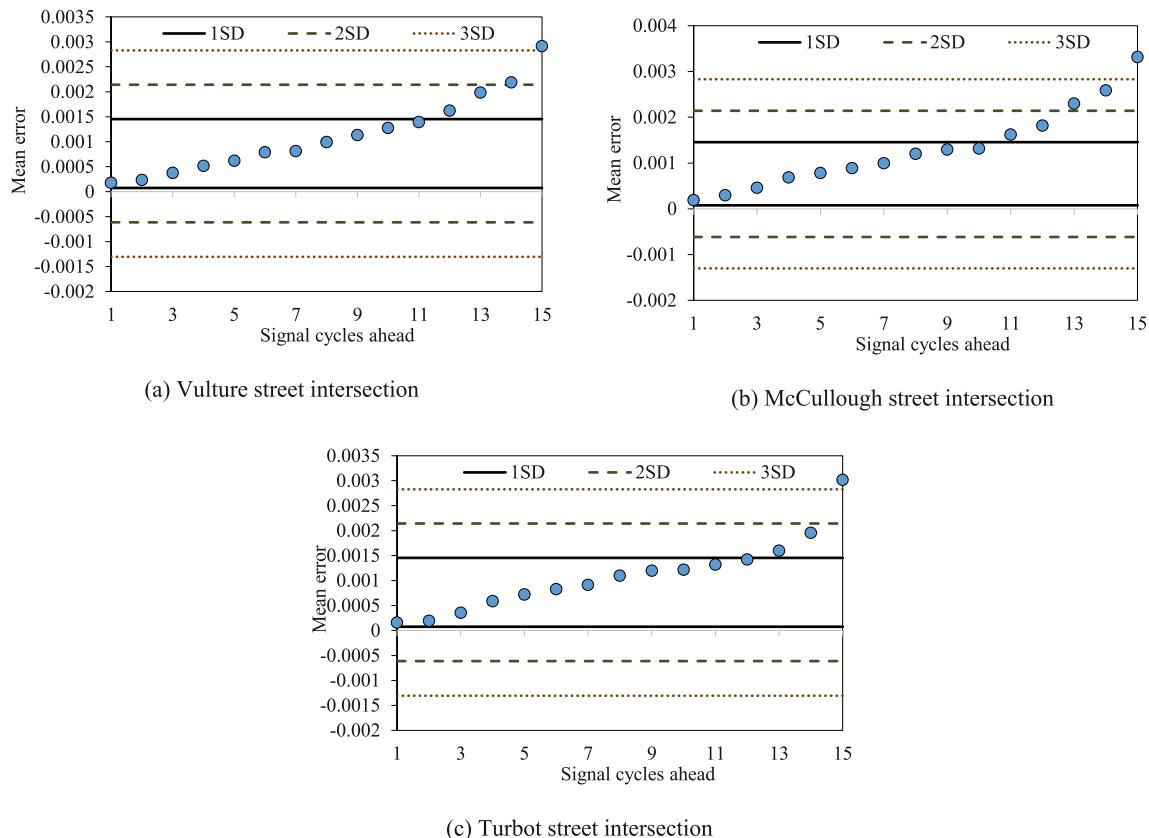
As presented in Table 5, the RAEs for the LSTM model are consistently higher for all intersections than those for the RNN model, indicating relatively better performance. The percentage increase in RAEs for the LSTM model with respect to the RNN model is calculated as  $\frac{RAE_{LSTM} - RAE_{RNN}}{RAE_{RNN}} \times 100$ , where a value of zero indicates that the LSTM model error is the same as the RNN model error, and a positive value

indicates that the LSTM model error is greater than the RNN model error. For the three intersections, the percentage increases are 7.92%, 7.29%, and 1.38%, implying that the LSTM model has a relatively large error compared with the RNN model. While in the literature, LSTM models (Zhang et al., 2020a) are often found to outperform RNN models, our study shows contrasting results, which can be explained as follows. LSTM models exhibit superior capabilities in modeling extended temporal dependencies, whereas this study predicts the pedestrian crash risk of future signal cycles from a smaller number of previous cycles; therefore, long-term dependency may not be applicable here. From a practical perspective, the pedestrian crash risk in the immediate future signal cycle is a function of the previous risk or a couple of cycles, reflecting short-term dependency, leading to the RNN model's better performance. A study on identifying anomalous driving behavior in routine tasks revealed that an RNN model was better than an LSTM model because of the local contextual influence of prior actions on future actions (Matousek et al., 2019).

## 5.2. Real-time pedestrian crash forecasting outreach

The next natural—and important—step is to determine how far in advance pedestrian crash risk can be reliably estimated. This forecasting horizon is critical for enabling real-time (a) safety assessments, (b) development of risk-responsive countermeasures, and (c) assessment of countermeasure effectiveness. To this end, the RNN model is used for pedestrian crash risk forecasting, and the model error (average mean absolute error of all 4 days for a given intersection) is computed and shown in Fig. 8.

Fig. 8 shows the RNN model performance for the next 15 signal cycles, with corresponding standard deviations. The RNN model is capable of predicting a minimum of 10 cycles within one standard deviation. Except for the Vulture Street intersection, the prediction of the 10th cycle is outside of one standard deviation for the other two intersections,



**Fig. 8.** Typical real-time pedestrian crash risk forecasting results from the RNN model.

whereas 12 signal cycles can be predicted within two standard deviations. These results imply a reasonable pedestrian crash risk forecasting performance of the RNN model. To summarize, the RNN model forecasting outreach within two standard deviations for Vulture Street, McCullough Street, and Turbot Street intersections is 13 cycles (approximately 33 min), 12 cycles (approximately 30 min), and 12 cycles (approximately 30 min), respectively.

### 5.3. Pedestrian crash risk variations across different times of the day

To fully leverage the capabilities of a real-time crash forecasting system, it is essential to consider crash risk variations on a typical day. For example, morning and evening peak hours may present significantly greater exposure, leading to greater pedestrian crash risk, whereas off-peak hours may present varying risk because of dispersed activities such as school pick-up times and routine tasks such as groceries and other appointments. To this end, the developed RNN model is employed to predict pedestrian crash risk across different times of a typical day for 4 days for all intersections, and the results are summarized in [Table 6](#). The RNN model demonstrates satisfactory performance in predicting pedestrian crash risk at various times of the day. While the prediction error (measured in mean absolute error) remains similar across all four days for a given intersection, it varies across different times. For example, prediction errors are greater during off-peak periods and smaller during morning and evening peak periods. A plausible reason for the higher error for off-peak events could be the monotonous nature of the crash risk during these times, which may offer limited informative patterns for the RNN model to learn from and make accurate predictions.

## 6. Conclusions

This study combined extreme value theory and machine learning models for forecasting pedestrian crash risk at signalized intersections. For extreme value theory, a Bayesian nonstationary peak-over-threshold model was developed, which considers the post encroachment time as a traffic conflict measure for characterizing vehicle–pedestrian safety-critical interactions. Several covariates were identified at the signal cycle and used as inputs to the peak-over-threshold model to handle the non-stationarity of conflict extremes. To address serial dependency in conflict extremes, a decluttering approach was applied, whereby only one conflict per signal cycle was considered. Several covariates were introduced in the scale and shape parameters, and the parsimonious model contained two covariates in the scale parameter and one in the shape parameter. As a requisite for generalized Pareto (or peak-over-threshold) model development, a threshold was needed, which was obtained via quantile regression, suggesting the use of the 87.5% quantile. The model was estimated via the Bayesian estimation procedure, and the best-fit model was used for crash frequency analysis. The developed model showed reasonable predictions (measured by mean crashes and confidence intervals) that were comparable to historical crash records. The model was subsequently employed to estimate

pedestrian crash risk at the signal–cycle level, which was then used as an input for machine learning models to forecast pedestrian crash risk.

Recurrent neural network (RNN) and long short-term memory (LSTM) models were evaluated for their effectiveness in forecasting pedestrian crash risk. When the optimal hyperparameters were used, both models achieved reasonable prediction accuracy, with the RNN model outperforming the LSTM model, suggesting the absence of long dependencies in the data. The RNN model was then applied to forecast pedestrian crash risk, and the model was able to forecast crash risk for 30–33 min. Furthermore, the RNN model was tested to capture within-day crash risk variation and demonstrated excellent performance in capturing time-varying crash risk.

Notably, our study developed separate models for each intersection and each day, which may limit practical applications. However, our choice of developing separate models for intersections and days was primarily based on data limitations (smaller sample sizes with less variation). If the dataset contains 24 h of continuous data for each day, the model will be generalizable, whereby the training sequence will update easily from the last data point of one day to the starting data point of another day. Future research, where continuous data from multiple days are available, could verify the generalizability of the model. Furthermore, given that the focus of the current study was on crash risk forecasting, generalizability capabilities were not assessed. Future research could address this by designing training and testing setups that evaluate the performance on unseen intersections. For example, data from the Turbot Street intersection could be used for training, and the model could then be applied to forecast crash risk at the McCullough Street intersection. Such analysis could be conducted via strategies such as partial model retention or transfer learning.

The findings of this study are useful for road authorities, transport operators, and stakeholders/policymakers in understanding pedestrian crash risk in real time, devising countermeasures, and testing them in real time. The methodology presented in this study leverages existing CCTV cameras and can be considered cost-effective. Finally, this study provides a pathway to a paradigm shift to proactive safety management, whereby road authorities do not need to wait unnecessarily long to collect crash data but utilize crash precursors to make decisions in real time and enhance pedestrian safety.

This study offers several avenues for further extension. First, this study considered EVT-estimated risk values as ground truths, which may contain uncertainty, which is likely to propagate to RNN/LSTM models. To understand its impact and then capture it, a two-stage modeling procedure can be used to quantify the error propagation (see [Ali et al. \(2021\)](#) for more information). Second, although this study compared block maxima and peak over threshold models, several advancements prevail in the literature, meriting a rigorous analysis/comparison. For example, using a dynamic updating process to include covariates in the extreme value could be worth investigating. Third, this study only evaluated crash risk forecasting without severity, which provides partial information about safety. Future studies can develop bivariate models to forecast pedestrian crash risk by severity level. Fourth, this study used RNN/LSTM models, but their latest advancements, such as deep

**Table 6**  
RNN model prediction performance across different times of the day.

Intersection	Time of the day	Day			
		1	2	3	4
Vulture Street	Morning peak	$1.81 \times 10^{-4}$	$6.33 \times 10^{-4}$	$2.77 \times 10^{-4}$	$1.15 \times 10^{-4}$
	Off-peak	$2.01 \times 10^{-4}$	$7.32 \times 10^{-4}$	$2.01 \times 10^{-4}$	$1.01 \times 10^{-4}$
	Evening peak	$1.85 \times 10^{-4}$	$7.11 \times 10^{-4}$	$2.88 \times 10^{-4}$	$1.21 \times 10^{-4}$
McCullough Street	Morning peak	$1.87 \times 10^{-4}$	$1.98 \times 10^{-4}$	—	—
	Off-peak	$1.60 \times 10^{-4}$	$1.27 \times 10^{-4}$	—	—
	Evening peak	$2.00 \times 10^{-4}$	$2.10 \times 10^{-4}$	—	—
Turbot Street	Morning peak	$6.34 \times 10^{-4}$	$4.34 \times 10^{-4}$	—	—
	Off-peak	$3.72 \times 10^{-4}$	$3.13 \times 10^{-4}$	—	—
	Evening peak	$7.86 \times 10^{-4}$	$6.96 \times 10^{-4}$	—	—

learning, sparse adversarial frameworks, and transformers (Liu et al., 2025), were not used predominantly because of data limitations. Future studies can collect data for longer periods, allowing for the use of deep learning models. Fifth, this study only used the post encroachment time as a conflict metric. However, as newer physics-based metrics, such as risk force (Sahu et al., 2025), a comparison of post encroachment time with risk force needs to be performed to ascertain the best-performing metric for vehicle–pedestrian interactions. Sixth, this study used RNN/LSTM models, but their latest advancements, such as deep learning, sparse adversarial frameworks, gated recurrent units, and transformers (Liu et al., 2025), were not used predominantly because of data size limitations. Future studies can collect data for longer periods, allowing a rigorous comparison of these advanced models. Finally, building on our previous studies (Hussain et al., 2022, 2024b) on replacing conventional sampling (block maxima and peak over threshold) with anomaly detection, our next work combines anomaly detection with machine learning-based forecasting models such as RNN/LSTM.

### CRediT authorship contribution statement

**Fizza Hussain:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yue-feng Li:** Writing – review & editing, Validation, Conceptualization. **Shimul Md Mazharul Haque:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

### Replication and data sharing

The data and code used in this study can be accessed at <https://github.com/Fizza-nn/Pedestrian-crash-risk-forecasting>.

### 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 research is partially funded by the Queensland University of Technology, iMOVE CRC, and supported by the Cooperative Research Centres program, an Australian Government initiative. The authors would also like to acknowledge Transport and Mains Roads (TMR), Queensland, for providing video data.

### References

- Ali, Y., Haque, M.M., Mannerling, F., 2023a. Assessing traffic conflict/crash relationships with extreme value theory: recent developments and future directions for connected and autonomous vehicle and highway safety research. *Anal. Meth. Accid. Res.* 39, 100276.
- Ali, Y., Haque, M.M., Mannerling, F., 2023b. A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. *Anal. Meth. Accid. Res.* 38, 100264.
- Ali, Y., Haque, M.M., Zheng, Z., Afghari, A.P., 2022. A Bayesian correlated grouped random parameters duration model with heterogeneity in the means for understanding braking behaviour in a connected environment. *Anal. Meth. Accid. Res.* 35, 100221.
- Ali, Y., Hussain, F., Haque, M.M., 2024. Advances, challenges, and future research needs in machine learning-based crash prediction models: a systematic review. *Accid. Anal. Prev.* 194, 107378.
- Ali, Y., Zheng, Z., Haque, M.M., Yildirimoglu, M., Washington, S., 2021. CLACD: a complete Lane-Changing decision modeling framework for the connected and traditional environments. *Transport. Res. C Emerg. Technol.* 128, 103162.
- Allen, B.L., Shin, B.T., Cooper, P.J., 1978. Analysis of traffic conflicts and collisions. *Transp. Res. Rec.* 667, 67–74.
- Alzoli, A.R., Hussein, M., 2022. Evaluating the safety of autonomous vehicle–pedestrian interactions: an extreme value theory approach. *Anal. Meth. Accid. Res.* 35, 100230.
- AMAG SMART safety. <https://amagroup.io/smart-safety>, 2024.
- Ankunda, A., Ali, Y., Mohanty, M., 2024. Pedestrian crash risk analysis using extreme value models: new insights and evidence. *Accid. Anal. Prev.* 203, 107633.
- Arun, A., Lyon, C., Sayed, T., Washington, S., Loewenherz, F., Akers, D., et al., 2023. Leading pedestrian intervals—Yay or Nay? A Before-After evaluation of multiple conflict types using an enhanced Non-Stationary framework integrating quantile regression into Bayesian hierarchical extreme value analysis. *Accid. Anal. Prev.* 181, 106929.
- Ashraf, M.T., Dey, K., Pyrialakou, D., 2022. Investigation of pedestrian and bicyclist safety in public transportation systems. *J. Transport Health* 27, 101529.
- ATC, 2020. National Road Safety Strategy 2011–2020. [https://www.roadsafety.gov.au/sites/default/files/2019-11/nrss\\_2011\\_2020.pdf](https://www.roadsafety.gov.au/sites/default/files/2019-11/nrss_2011_2020.pdf).
- Basso, F., Pezoa, R., Varas, M., Villalobos, M., 2021. A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. *Accid. Anal. Prev.* 162, 106409.
- Bin Tahir, H., Haque, M.M., 2024. A non-stationary bivariate extreme value model to estimate real-time pedestrian crash risk by severity at signalized intersections using artificial intelligence-based video analytics. *Anal. Meth. Accid. Res.* 43, 100339.
- Chan, Z.Y., Suandi, S.A., 2019. City tracker: multiple object tracking in urban mixed traffic scenes. In: 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 335–339.
- Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, London.
- DTMR, 2020. Summary Road Crash Report Queensland Road Fatalities. Queensland Transport, Australia. <https://www.qld.gov.au/transport/safety/road-safety/statistics>.
- Essa, M., Sayed, T., 2018. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. *Transport. Res. C Emerg. Technol.* 89, 289–302.
- Feng, R., Fan, C., Li, Z., Chen, X., 2020. Mixed road user trajectory extraction from moving aerial videos based on convolution neural network detection. *IEEE Access* 8, 43508–43519.
- Guo, Y., Sayed, T., Zheng, L., 2020. A hierarchical Bayesian peak over threshold approach for conflict-based before-after safety evaluation of leading pedestrian intervals. *Accid. Anal. Prev.* 147, 105772.
- Hewett, N., Fawcett, L., Golightly, A., Thorpe, N., 2024. Using extreme value theory to evaluate the leading pedestrian interval road safety intervention. *Stat* 13, e676.
- Howlader, M.M., Haque, M.M., 2025. Opposing-through crash risk forecasting using artificial intelligence-based video analytics for real-time application: integrating generalized extreme value theory and time series forecasting models. *Accid. Anal. Prev.* 218, 108073.
- Hughes, C., 2023. Vehicles & Road Traffic. <https://www.statista.com/statistics/992074/total-number-pedestrian-deaths-australia>.
- Hussain, F., Ali, Y., Li, Y., Haque, M.M., 2023. Real-time crash risk forecasting using Artificial-Intelligence based video analytics: a unified framework of generalised extreme value theory and autoregressive integrated moving average model. *Anal. Meth. Accid. Res.* 40, 100302.
- Hussain, F., Ali, Y., Li, Y., Haque, M.M., 2024a. A bi-level framework for real-time crash risk forecasting using artificial intelligence-based video analytics. *Sci. Rep.* 14, 4121.
- Hussain, F., Ali, Y., Li, Y., Haque, M.M., 2024b. Revisiting the hybrid approach of anomaly detection and extreme value theory for estimating pedestrian crashes using traffic conflicts obtained from artificial intelligence-based video analytics. *Accid. Anal. Prev.* 199, 107517.
- Hussain, F., Li, Y., Arun, A., Haque, M.M., 2022. A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts. *Anal. Meth. Accid. Res.* 36, 100248.
- Ihsian, A., Ismail, K., 2023. Modelling pedestrian safety at urban intersections using user perception. *Accid. Anal. Prev.* 180, 106912.
- Koenker, R., Bassett, G., 1978. Regression quantiles. *Econometrica* 46, 33–50.
- Lanzaro, G., Sayed, T., Fu, C., 2023. A comparison of pedestrian behavior in interactions with autonomous and human-driven vehicles: an extreme value theory approach. *Transp. Res. Part F Traffic Psychol. Beyond Behav.* 99, 1–18.
- Liu, Y., Jia, C., Rasouli, S., Gong, J., Feng, T., Wong, M., et al., 2025. SAFER-predictor: sparse adversarial training framework for robust traffic prediction under missing and noisy data. *Commun. Transp. Res.* 5, 100192.
- Mannerling, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Anal. Meth. Accid. Res.* 25, 100113.
- Marczak, F., Daamen, W., Buisson, C., 2013. Merging behaviour: empirical comparison between two sites and new theory development. *Transport. Res. C Emerg. Technol.* 36, 530–546.
- Matousek, M., El-Zohairy, M., Al-Momani, A., Kargl, F., Bösch, C., 2019. Detecting anomalous driving behavior using neural networks. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2229–2235.
- Mukherjee, D., Mitra, S., 2020. Identification of pedestrian risk factors using negative binomial model. *Transport. Dev. Econ.* 6, 4.
- Musrone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. *Accid. Anal. Prev.* 103, 112–122.
- Nasserreddine, H., Santiago-Chaparro, K.R., Noyce, D.A., 2023. Safety estimation of vehicle–pedestrian interactions using extreme value theory at intersections with and without right-turn flashing yellow arrow indication. In: International Conference on Transportation and Development, vol 2023, pp. 237–249.
- Nazir, F., Ali, Y., Haque, M.M., 2024. Effects of sample size on pedestrian crash risk estimation from traffic conflicts using extreme value models. *Anal. Meth. Accid. Res.* 44, 100353.
- Olah, C., 2015. Understanding lstm networks. <https://research.google/pubs/understanding-lstm-networks>.

- Rowan, D., He, H., Hui, F., Yasir, A., Mohammed, Q., 2025. A systematic review of machine learning-based microscopic traffic flow models and simulations. *Commun. Transp. Res.* 5, 100164.
- Sahu, S., Ali, Y., Glaser, S., Haque, M.M., 2025. A physics-informed risk force theory for estimating pedestrian crash risk by severity using artificial intelligence-based video analytics. *Anal. Meth. Accid. Res.* 46, 100382.
- Singh, S., Ali, Y., Haque, M.M., 2024. A Bayesian extreme value theory modelling framework to assess corridor-wide pedestrian safety using autonomous vehicle sensor data. *Accid. Anal. Prev.* 195, 107416.
- Songchitruksa, P., Tarko, A.P., 2006. The extreme value theory approach to safety estimation. *Accid. Anal. Prev.* 38, 811–822.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64, 583–639.
- Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649.
- Yang, Y., Deng, H., 2020. GC-YOLOv3: you only look once with global context block. *Electronics* 9, 1235.
- Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. *Accid. Anal. Prev.* 51, 252–259.
- Zeng, Q., Wang, Q., Zhang, K., Wong, S.C., Xu, P., 2023. Analysis of the injury severity of motor vehicle–pedestrian crashes at urban intersections using spatiotemporal logistic regression models. *Accid. Anal. Prev.* 189, 107119.
- Zhang, S., Abdel-Aty, M., Cai, Q., Li, P., Ugan, J., 2020a. Prediction of pedestrian-vehicle conflicts at signalized intersections based on long short-term memory neural network. *Accid. Anal. Prev.* 148, 105799.
- Zhang, S., Abdel-Aty, M., Wu, Y., Zheng, O., 2020b. Modeling pedestrians' near-accident events at signalized intersections using gated recurrent unit (GRU). *Accid. Anal. Prev.* 148, 105844.
- Zheng, L., Ismail, K., Meng, X., 2014. Freeway safety estimation using extreme value theory approaches: a comparative study. *Accid. Anal. Prev.* 62, 32–41.
- Zheng, L., Ismail, K., Meng, X., 2016. Investigating the heterogeneity of postencroachment time thresholds determined by peak over threshold approach. *Transp. Res. Rec.* 2601, 17–23.
- Zheng, L., Sayed, T., 2019. Bayesian hierarchical modeling of traffic conflict extremes for crash estimation: a non-stationary peak over threshold approach. *Anal. Meth. Accid. Res.* 24, 100106.



**Fizza Hussain** received the M.S. degree in telecommunication and computer networking from the National University of Science and Technology, Islamabad, Pakistan. She is currently a Ph.D. candidate at Queensland University of Technology. Her research interests include artificial intelligence for crash risk estimation in real time.



**Yuefeng Li** is a Professor at the School of Computer Science, Queensland University of Technology. He is recognized internationally in knowledge & data engineering, artificial intelligence, and information systems. He performs research on data mining and knowledge discovery, semi and unsupervised learning, deep learning, information modeling, management and ontologies, knowledge representation and data structure and algorithms.



**Shimul Md Mazharul Haque** is the Head of School and a Professor of transport engineering at the School of Civil and Environmental Engineering, Queensland University of Technology, Australia. He is a specialist in econometrics and artificial intelligence applications in transport engineering and traffic safety. His research interests include road safety, traffic conflict analysis, evaluation of engineering treatments and connected and automated vehicles.