Traffic & Transportation

A Data-Driven Framework for Traffic Crash Risk Prediction – Exploiting Multi-Source Heterogeneous Data

Min GUO[1], Mingxing GAO[2], Haixiao WANG[3]

Original Scientific Paper
Submitted: 5 Jul 2025
Accepted: 29 Oct 2025
Published: 29 June 2026

[1] guomin@imau.edu.cn, School of Energy and Transportation Engineering, Inner Mongolia Agricultural University, Hohhot, China
[2] Corresponding author, 15848943679@163.com, School of Energy and Transportation Engineering, Inner Mongolia Agricultural University, Hohhot, China
[3] wanghx@imau.edu.cn, School of Energy and Transportation Engineering, Inner Mongolia Agricultural University, Hohhot, China

This work is licensed under a Creative Commons Attribution 4.0 International Licence.

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
Frequent traffic crashes on urban roads seriously threaten public safety and traffic operations. Accurate risk prediction is vital for improving management efficiency and developing intervention measures. This paper proposes a traffic crash risk prediction model integrating multi-source heterogeneous data. It constructs a dynamic spatio-temporal graph network (DTGN) based on edge-aware graph convolutional networks (EGCN) and introduces a dynamic threshold risk stratification mechanism and local crash density (LCD) indicators to alleviate the issue of "zero inflation" in low-frequency areas. The model combines graph convolutional and spatio-temporal convolutional networks to extract multi-dimensional spatio-temporal features and enhances the ability to identify high-risk areas through a weighted loss function. The city is partitioned into hexagonal grid units, and a dynamic adjacency matrix is constructed to capture spatial associations and evolutionary features. Experimental results indicate that DTGN performs effectively in processing multi-source data and extracting key risk features, achieving an accuracy rate of 87% in high-risk area predictions, thereby providing more practical early warning support and decision-making basis for urban traffic safety management.

KEYWORDS
traffic crash risk prediction; high-risk area identification; dynamic threshold.

1. INTRODUCTION

With the advancement of urbanisation and the growth of traffic flow, the convenience of residents' travel has increased significantly. However, this progress has also increased the pressure on road traffic safety. *Figure 1* shows the trend of road traffic crashes and fatalities between 2004 and 2022 (Source: National Bureau of Statistics, China Statistical Yearbook https://www.stats.gov.cn/sj/ndsj/). The overall downward trend in traffic crashes and fatalities over the past 20 years reflects the positive effects of traffic management policies and safety measures. However, the risk of crashes remains high in specific regions and periods, indicating the need to enhance current traffic safety management strategies further. Therefore, traffic crash risk prediction is crucial to establishing an efficient traffic safety early warning system, which can provide a scientific basis for traffic safety management and help precise policymaking, risk prevention and control.

Various factors, including traffic flow, road structure, weather conditions, land use and population distribution, influence the frequency of traffic crashes. These elements make it challenging for traditional analytical methods, such as logistic regression and random forests, which struggle to capture spatio-temporal dependencies and environmental variability. Currently, methods such as spatiotemporal data fusion [1], [2] and deep learning models [3, 4] can effectively capture the spatial and temporal patterns of crashes, and have become the primary technical means for traffic flow prediction as well as crash prediction. Traditional machine

learning methods [5], including classification regression trees, negative binomial regression, decision trees, and a regression model between the various factors by clustering the crash rate, predicting the crash rate of the road in the region [6], and verifying its accuracy. However, their effectiveness is often constrained when dealing with large-scale, nonlinear and spatio-temporally correlated data. Recent studies in deep learning show notable improvements in accuracy and robustness, which highlight its potential as a promising path for future studies on traffic safety.

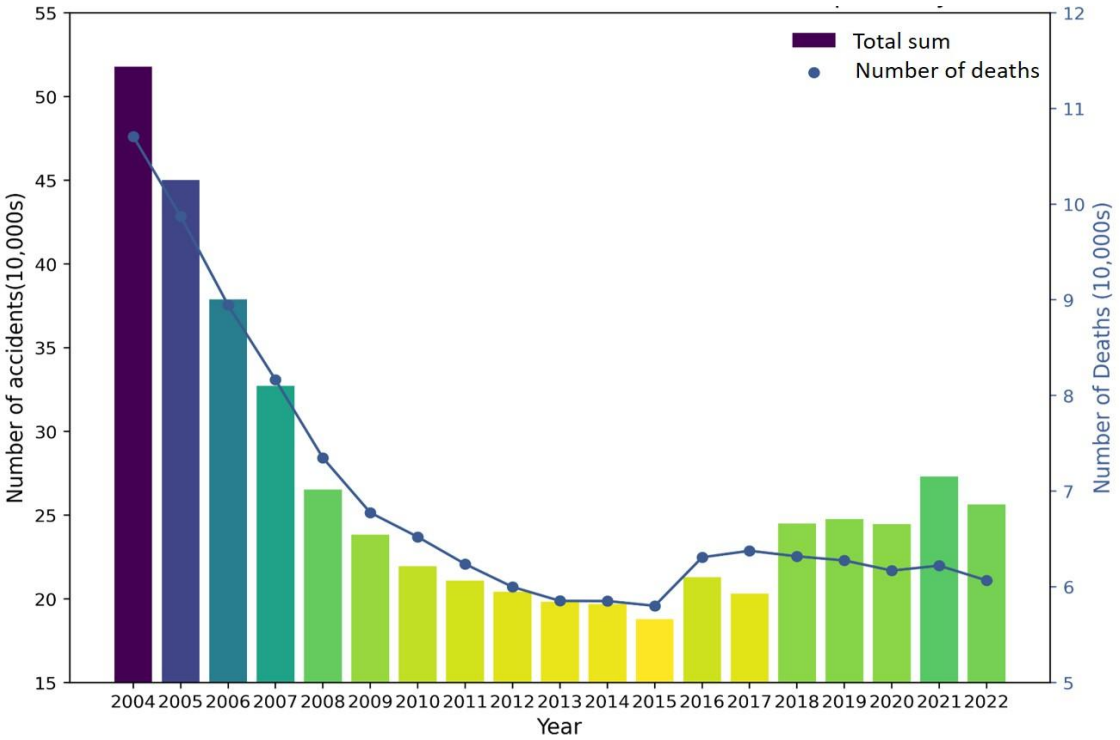

Figure 1 – The number of road traffic crashes and fatalities (recent 20 years)

ConvLSTM integrates the spatial learning capabilities of CNNs with the temporal modelling strength of LSTM to capture the complex nonlinear traffic dynamics [7], [8]. Hu et al. proposed a ConvLSTM-based framework for predicting traffic crashes in Ningbo, which can effectively capture spatial and temporal dependencies in historical crash data, achieving higher predictive accuracy than traditional LSTM models [10]. The end-to-end LSTM [9] can extract hazardous state features from low-quality raw trajectory data; however, it still faces certain limitations, especially in handling complex spatial relationships and high-dimensional data with overfitting or computational inefficiency. These issues have gradually led to graph neural networks (GCNs) becoming a significant research focus. GCNs accurately capture interregional dependencies through their graph structure, demonstrating improved performance during the learning process [2], [10]. For example, STGCN [12] improves spatial-temporal correlation capture and training efficiency through multi-scale graph modelling. Meanwhile, Zhang et al. used random forest, SVM and XGBoost to predict real-time freeway crash risk and analyse the influence of dynamic traffic features [13].

Building on these advances, spatio-temporal graph modelling has become an important line of inquiry in traffic prediction. T-GCN [14] integrates graph structures into LSTM units, enabling dynamic spatiotemporal forecasting. By contrast, Risk Oracle [15] employs multi-task learning combined with deep graph convolutional networks to estimate citywide crash risks. With a different emphasis, Tao et al. [16] proposed a multi-scale spatiotemporal GCN that leverages historical traffic similarity to enhance prediction accuracy through a novel fusion mechanism. In addition, Wang et al. examined the temporal heterogeneity of traffic crash delays and showed that multi-scale temporal factors exert a significant influence on crash frequency [17].

Despite these advances, Choudhary et al. [18] and McCarty et al. [19] emphasise the necessity of integrating road, traffic and urban structural characteristics. Building on this, GSNet [2] further incorporates geographic and semantic information to improve the accuracy of traffic crash risk prediction. Moreover, demographic and behavioural factors, including age, gender and driver behaviour, have been shown to significantly influence crash risk [11, 21]. Future prediction models should integrate socio-economic factors and urban development factors to enhance the comprehensiveness and accuracy of predictions [22–25].

Currently, there are several significant challenges to traffic crash prediction:

1) Complex causality. Various factors influence crashes, including weather, time of day and other related factors. The complex interactions between these factors increase the difficulty of risk prediction.
2) Spatial and temporal dependence. Crashes are affected by temporal and spatial features. Significant variations in travel patterns and traffic conditions over different times and regions complicate the modelling process.
3) Zero inflation problem. The low frequency of crashes leads to a significant number of zero values in the crash data. The high frequency of zero data values causes the model to over-predict the scenario of no crashes, ultimately reducing prediction accuracy.

In this study, we integrate multi-source data and use the hexagonal grid method to evaluate the study area with fine-grained dynamics to improve the spatial resolution. Meanwhile, a dynamic neighbour matrix will be constructed to capture the spatiotemporal correlation and integrate the dynamic threshold risk stratification strategy for crash risk prediction.

2. DATA PROCESSING AND FEATURE SELECTION

2.1 Data acquisition

This study focuses on the main urban area of Hohhot, Inner Mongolia, which is characterised by a complex road network and faces significant challenges related to traffic safety, largely due to harsh winter conditions. To construct a comprehensive data framework, we integrated multiple data sources, including the urban traffic survey data, crash records of the traffic management department, meteorological data and basic road information.

1) Traffic crash data: Obtained from the traffic management department for 2020–2021, including the time, location and casualties of each crash.
2) Traffic flow data: Traffic flow information, including peak hour and off-peak hour traffic flow, was collected from 160 major roads and used to study traffic safety in Hohhot.
3) Road infrastructure data: Include information on road types, intersections and road conditions. Sourced from the Planning Bureau.
4) Socio-economic and land-use data: Cover the 380 transportation districts regarding population density, land-use types.
5) Meteorological data: Access weather variables such as precipitation, temperature, humidity and wind speed on the Meteorological Bureau's official website, https://data.cma.cn/.

2.2 Data visualisation

Figure 2 presents a comprehensive visualisation of traffic trip patterns in the urban study area. *Figure 2(a)* illustrates the distribution of the percentage of traffic trips between weekdays and non-workdays throughout the year. Overall traffic volume on weekdays is higher, with an earlier morning and evening peak due to the commuting demand. While midday travel increases on non-working days, it is likely reflecting leisure and social activities. *Figure 2(b)* illustrates the spatial distribution characteristics of crashes using K-means clustering, revealing a noticeable core-boundary effect, with dense crashes in the central city and fewer crashes in the peripheral areas. *Figure 2(c)* presents traffic flow distribution in the main urban area during off-peak hours. High-traffic flow areas often correspond to higher crash rates, although road type, management measures and environmental conditions influence crash occurrence.

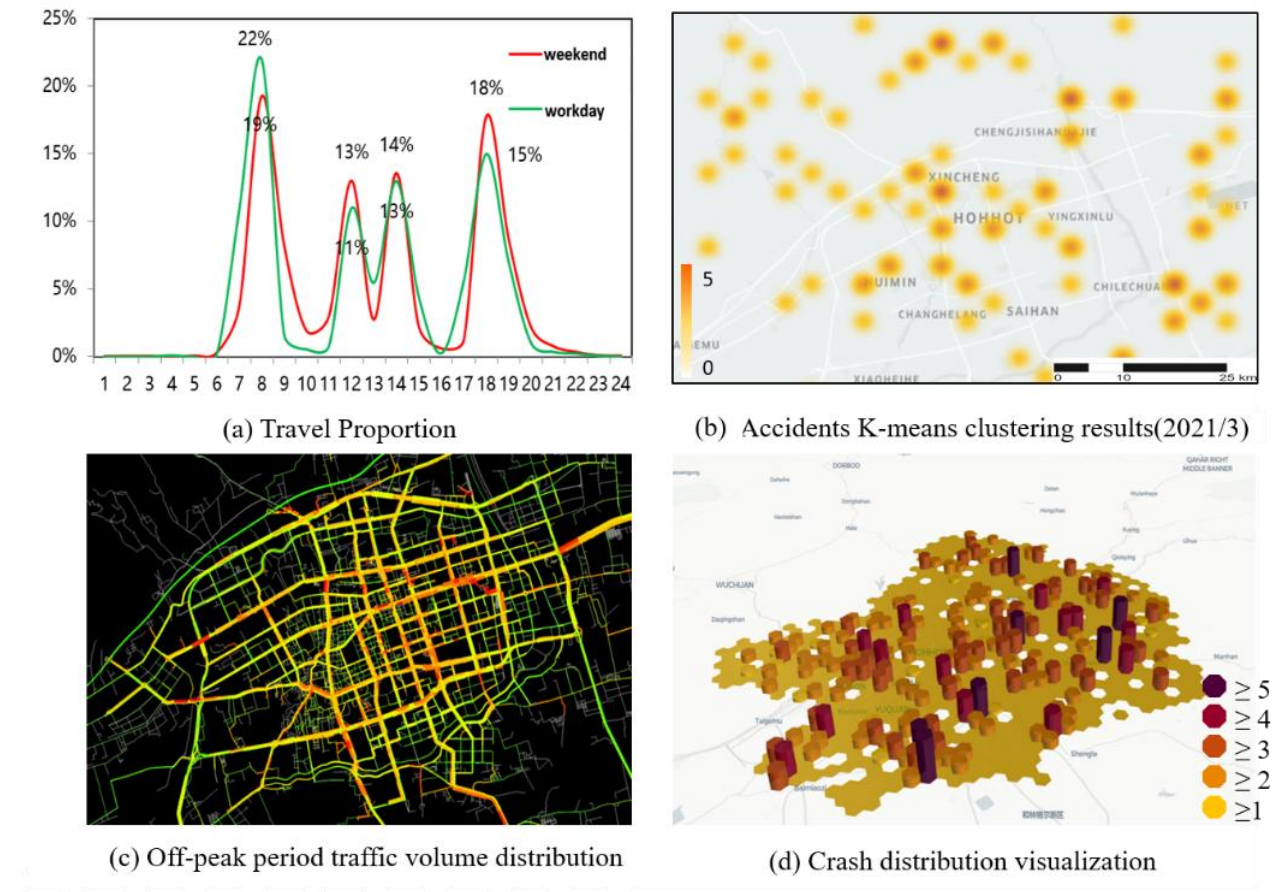

Figure 2 – Data visualisation

To visually and effectively examine the correlation between crash frequency and other factors, we use GIS with data processing techniques to construct a three-dimensional temporal–geospatial model. A hexagonal grid with a spatial resolution of approximately 1 km was employed as the basic analytical unit to minimise the influence of uneven boundary effects on the analysis. Compared to square grids, hexagonal cells provide several advantages: (1) each cell of a hexagonal grid has six equidistant neighbours, more accurately representing real spatial proximity; (2) hexagons exhibit stronger directional isotropy, helping to minimise errors stemming from directional bias during simulations; (3) the use of a hexagonal grid helps reduce the "sawtooth" effect. Owing to these properties, hexagonal grids have been widely applied in transport geography and spatial modelling studies and have been shown to preserve spatial autocorrelation and clustering characteristics better. As shown in *Figure 2(d)*, the study area was divided into 948 cells, capturing variations in the number across different periods.

2.3 Feature selection

This study employed the LightGBM algorithm to evaluate feature importance and select 15 key variables for traffic crash risk prediction. These variables span four categories: crash characteristics, traffic flow indicators, road structure attributes and land-use/environmental factors, to improve the accuracy of crash risk prediction and support urban road safety management.

LightGBM, proposed by Microsoft in 2016 [26], is an optimised gradient boosting decision tree (GBDT) framework. It enhances training efficiency and reduces memory consumption through histogram-based computation and a leaf-wise growth strategy, making it particularly well-suited for large-scale datasets. To ensure consistency and comparability across multiple data sources, the study area used a hexagonal grid and performed spatial matching and data calibration to mitigate boundary effects.

The dataset used in this study is extensive, totalling 2.1 TB. It includes traffic flow information from 160 major roads and crash data covering 380 traffic districts. This large-scale dataset provides a robust foundation for assessing traffic crash risk and identifying high-risk areas and time periods – the final selected variables as listed in *Table 1*.

Table 1 – Selected features for LightGBM-based crash risk analysis

Crash-related features	Socioeconomic and demographic features	Road network structure	Weather-related features
Crash	Population density	Hex grid ID	Weather conditions
Hour		Road type	Rain hour
Weekday	Land use mix degree (LUD)	Road length	Temperature
Day of the week		Joint count	Humidity
			Wind speed

2.4 Data integration and heterogeneous fusion

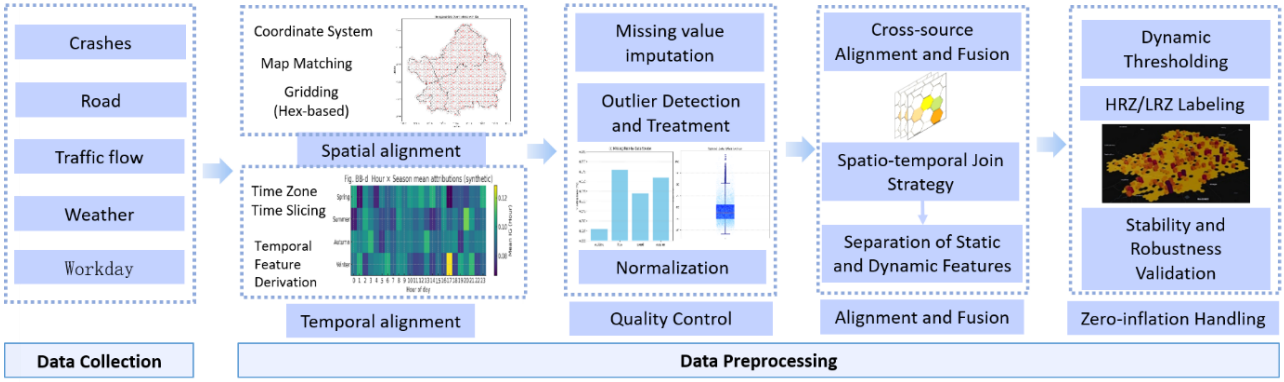

Figure 3 – Framework of multi-source heterogeneous data pre-processing

To unify multi-source heterogeneous data, we designed a pre-processing and fusion framework (*Figure 3*) consisting of three key stages:

1) Spatial and temporal alignment

All datasets were first transformed into a standard coordinate reference system, after which the study area was partitioned into 1 km hexagonal grids. Traffic crash data were mapped to their nearest road segments through map matching, while road attributes, traffic flow and weather observations were aggregated to their corresponding grid cells. All datasets were resampled into 0.5-hour intervals based on time slices to ensure consistency. Dynamic variables – traffic flow, speed and weather data – were summarised as hourly means, while crash events matched their respective grids.

2) Quality control

We implemented several procedures to address missing values and outliers. Missing values within short gaps (under 30 minutes) were imputed via linear interpolation, while longer gaps were corrected using spatio-temporal weighted averages and historical patterns. For grids or intervals with more than 30% missing data, the samples were excluded to avoid noise. To detect outliers, we combined physical constraints, such as traffic speed being limited to a range of 0 to 150 km/h, with statistical methods, including boxplot-based truncation and Winsorization [27]. Lastly, we normalised the data to improve comparability across different features.

3) Cross-source alignment and fusion

Following spatial and temporal alignment, static dynamic attributes were combined into a unified spatio-temporal tensor, indexed by grid ID and time slice. Different data modalities were aligned through a cross-source join strategy, after which the variables were grouped into static and dynamic sets to provide greater flexibility for subsequent modelling tasks.

2.5 Zero-inflated issue

In traffic crash prediction, the crash frequency of some grid areas is extremely low, which leads to sparse data, uneven distribution and a serious zero-inflation problem. However, these zero values do not necessarily mean “no risk at all” but may imply different degrees of potential crash risks. Therefore, this study proposes a dynamic threshold risk stratification method to categorise and process the zero-value grids to enhance the model’s learning ability for the low-crash-frequency region. Calculate the grid crash risk indicator as in *Equation 1* to assess the potential crash risk of each grid.

$$R_i^t = \sum_{j=1}^{3}(N_{ij}^t \delta_j) \quad (1)$$

R_i^t represents the total crash risk for g_i during time period t, N_{ij}^t represents the number of crashes of severity level j occurring in g_i during time period t. δ_j represents the severity level assigned to crash type j, with $\delta_j = 1$, $\delta_j = 2$ and $\delta_j = 3$.

To eliminate the effect of scale, we normalised the crash risk indicators for all grids to a range between [0,1] as in *Equation 2*.

$$R_i^t = \frac{R_i^t - min(R^t)}{max(R^t) - min(R^t)} \quad (2)$$

Since some grids may be free of crashes at a given time, their neighbourhoods may be at high risk. Therefore, we introduce localised crash density (LCD) as *Equation 3* to measure the overall crash risk level around a grid:

$$LCD_i^t = \frac{\sum_{j \in N(i)} S_j^t}{|N(i)|} \quad (3)$$

where: $N(i)$ denotes the neighbourhood of the grid i;

$|N(i)|$ is the number of neighbourhood grids i;

S_j^t is the crash risk indicator for the neighbouring grid j.

We established a dynamic threshold τ using the mean plus standard deviation method to categorise the zero-valued grids more effectively, as in *Equation 4*. This approach allows us to classify the zero-valued grids into high-risk zeros (HRZ) and low-risk zeros (LRZ), and decide the loss function weights w_i, as shown in *Equation 5*.

$$\tau = \mu_{LCD} + \sigma_{LCD} \quad (4)$$

where μ_{LCD}: mean value of localised incident density for all zero-valued grids. σ_{LCD}: standard deviation of localised incident density for all zero-valued grids.

$$w_i = \begin{cases} 1 + \frac{LCD_i^t}{max(LCD)}, LCD_i^t \geq \tau \\ 1 \quad , LCD_i^t < \tau \\ 1 \quad , \text{other} \end{cases} \quad (5)$$

To further evaluate the proposed method's stability and robustness under varying crash density distributions, temporal/spatial subsets and threshold perturbations, we designed systematic comparative experiments in Section 4.3. We summarised the related findings in the discussion section.

3. METHODOLOGY

3.1 DTGN framework

This study proposes a dynamic temporal-spatial graph neural network (DTGN) framework for accurately predicting the risk of traffic crashes on urban road networks. As shown in *Figure 4*, the model first performs unified grid processing and feature selection on multi-source heterogeneous data (traffic flow, crash records, road structure, weather information, etc.) to enhance the compactness and effectiveness of feature representation. A dynamic adjacency matrix integrates temporal synchrony, trend consistency and causal relationships in the spatial modelling layer. Based on this, an enhanced graph convolutional network (EGCN) is employed to extract dynamic spatial association features between regions. In the temporal modelling layer, a temporal convolutional network (TCN) is integrated with gated recurrent units (GRUs) to enable multi-scale learning of crash temporal patterns. Subsequently, a cross-attention fusion module establishes contextual interactions between spatial and temporal features, enhancing the model's ability to perceive complex spatio-temporal dependencies.

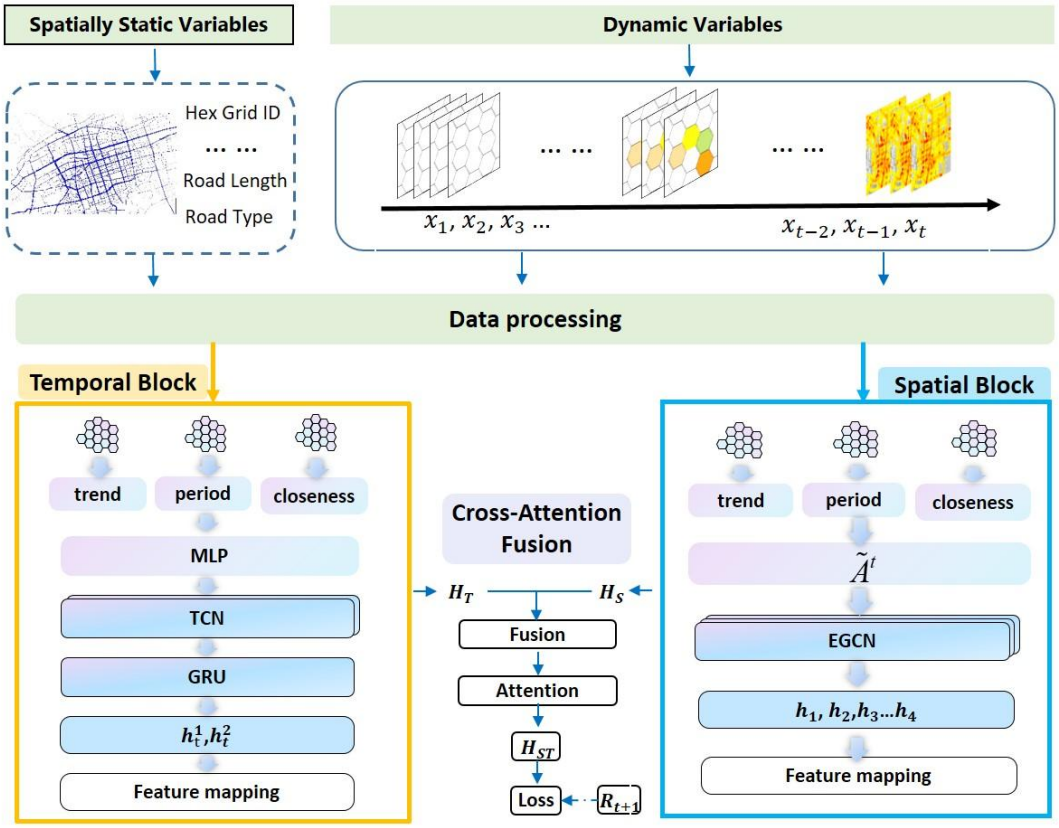

Figure 4 – Framework overview of DTGN

3.2 Temporal block

Time blocks are essential in DTGN models, focusing on temporal features in traffic crash risk prediction. TCN processes time sequential data through convolutional layers, accommodating various temporal dependencies and reducing information leakage by employing causal convolutions that only operate on current and past inputs [14].

Nevertheless, TCNs may encounter difficulties in the variability of dependency strengths and adapting to dynamic temporal data, such as irregular patterns in traffic crash data. To address this limitation, this study integrates TCN with gated recurrent units (GRUs). TCNs capture both short- and long-term variations through one-dimensional convolutions, whereas the GRU models temporal dependencies using recurrent hidden states.

Integration of TCN and GRU through time blocks leverages the strengths of both models to improve the identification of temporal correlations and enhance the prediction accuracy for time series data. As shown in *Figure 5*, input features are $X = \{x_1, x_2, ..., x_T\}$. The process h_t begins with the TCN, as in *Equations 6–7,* which utilises extended convolution and residual connections to extract multi-scale temporal features, effectively capturing both short-term and long-term dependencies. The sequences of features $h_t^{(l)}$ extracted by the TCN are then input into the GRU, which further captures temporal dependencies through the computations of update and reset gates. In the end, H_T produced by the GRU are provided as the output of the time block for subsequent prediction tasks. As shown in *Equation 8*, where h_{t-1} represents the hidden state from the previous time step, and H_T denotes the memory unit in the GRU.

$$h_t = \sum_{i=0}^{k-1} W_i x_{t-i \cdot d} + b \; W \in \mathbb{R}^{k \times d_x \times d_{out}}, b \in \mathbb{R}^{d_{out}} \tag{6}$$

$$h_t^{(l)} = \sum_{i=0}^{k-1} W_i^{(l)} x_{t-i \cdot d}^{(l-1)} + b^{(l)} \tag{7}$$

$$H_T = (1 - h_t^{(l)}) \odot h_{t-1} + h_t^{(l)} \odot \tilde{h}_t \tag{8}$$

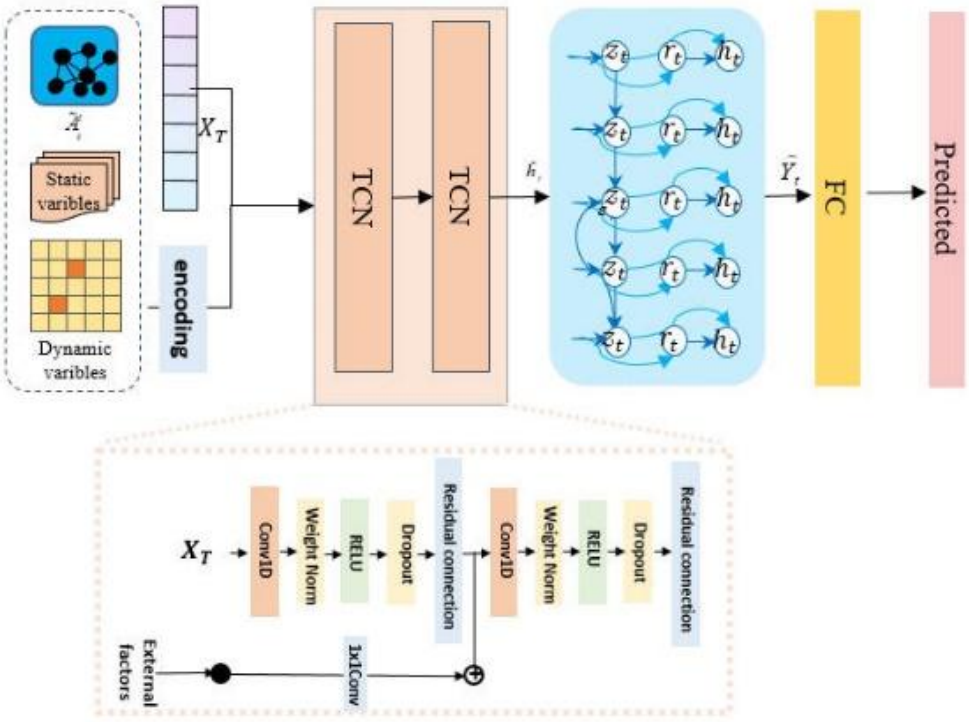

Figure 5 – Flow of temporal block

3.3 Spatial block

1) Dynamic adjacency matrix

After pre-processing the data to extract time synchronisation, trend consistency and collision risk correlation features, this study combines static spatial topology with time-varying factors and introduces a dynamic adjacency matrix construction strategy.

Specifically, three complementary metrics are introduced to model inter-regional dynamic relationships: time synchronisation (TS), trend consistency (TC) and crash risk correlation (ARC) to model dynamic regional relationships and develop a more precise dynamic adjacency matrix. First, the temporal synchronisation of traffic flows and crash events is calculated based on dynamic temporal regularisation (DTW), which measures the synchronisation of their dynamic change patterns, as shown in *Equation 9*. Next, mutual information (MI), as presented in *Equation 10*, is used to evaluate the consistency between historical crash rates and weather variables, thereby revealing potential regional associations at the trend level. Subsequently, the Granger causality test, described in *Equation 11*, is applied to historical crash data, holiday information and road network characteristics to identify inter-regional causal relationships, ensuring that the defined adjacencies reflect both correlation and causal influence. The above three metrics are fused through a multiplicative perceptron (MLP) as in *Equation 12*, with Softmax normalisation applied to generate the dynamic adaptive adjacency matrix.

$$TS(i,j) = \frac{1}{t}\sum_{t=1}^{T} DTW(X_i^t, X_j^t) \tag{9}$$

$$TC(i,j) = \frac{\sum_{t=1}^{T-1}(\Delta X_i^t, \Delta X_j^t)}{\sqrt{\sum_{t=1}^{T-1}(\Delta X_i^t)^2}\sqrt{\sum_{t=1}^{T-1}(\Delta X_j^t)^2}} \tag{10}$$

$$ARC(i,j) = \frac{P(Y_i \cap Y_j)}{P(Y_i)P(Y_j)} \tag{11}$$

$$\tilde{A}^t = \alpha^t TS^t + \beta^t TC^t + \gamma^t ARC^t \tag{12}$$

where: $P(Y_i \cap Y_j)$ denotes the probability of a crash occurring at both zones i and j at the same time.

2) Enhanced GCN

The spatial module employs a multi-layer edge-aware graph convolutional network (EGCN) to effectively capture the complex spatial dependencies within the urban traffic network. Unlike traditional GCNs that rely solely on adjacency structures, the proposed method incorporates both node and edge-level attributes, thereby enriching the representation of inter-regional interactions. Specifically, a dynamic adaptive neighbourhood matrix $\tilde{A}^t$, constructed from temporal synchronisation, trend consistency and crash causality, serves as the backbone for spatial message propagation.

At each layer l, node features are updated via edge-conditioned graph convolution operations as shown in *Equations 13–14*:

$$H_i^{(l+1)} = \sigma(\sum_{j \in N} \tilde{A}^t [(W^{(l)} H_j^{(l)} + \phi(f(E_{ij};\theta)))] \tag{13}$$

$$f(E_{ij};\theta) = W_e E_{ij} + b_e \tag{14}$$

where $H_i^{(l)}$ denotes the feature vector of node i at layer l, and E_{ij} represents the feature vector of the edge crash co-occurrence frequency between nodes i and j. The learnable function $f(\cdot)$ encodes edge semantics into adaptive weights, allowing spatial propagation to reflect heterogeneous and directional interactions more precisely.

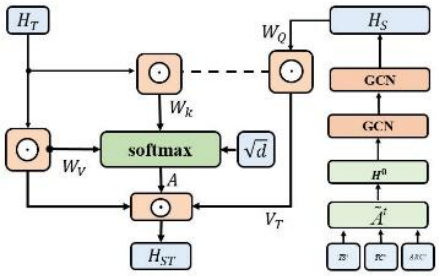

Figure 6 – Flow of spatial block

To further enhance spatial representation, the output of the spatial block, as expressed in *Equation 15,* is fed into a cross-attention fusion module. The final output spatial feature representation H_{ST} is then passed into a cross-attention fusion module, where temporal features H_T are integrated through a query-key-value attention mechanism as shown in *Equation 16*:

$$H_S = H^{(L)} \tag{15}$$

$$R_{ST} = soft\max(\frac{Q_S K}{\sqrt{d}})V_T \tag{16}$$

where $Q_S = H_S W_Q$, $K_T = H_T W_K$, $V_T = H_T W_V$.

3.4 Loss function

We propose a loss function for the zero-inflation problem, as detailed in *Equation 17*. The issue is mitigated by introducing dynamic weights in Section 2.5. The weight assigned larger values in high-risk areas, enabling the model to focus more on these low-frequency regions. Conversely, the weights are reduced in low-risk areas to minimise their influence on training. In this way, this model focuses more on high-risk regions, enhances its ability to learn from rare crash events, while diminishing the adverse effects of zero inflation on prediction performance.

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} w_i [R_i \log(\hat{R}_i) + (1 - R_i)\log(1 - \hat{R}_i)] \tag{17}$$

4. EXPERIMENT

4.1 Experimental setup and baseline methods

To effectively validate the DTGN model, we employed a multidimensional traffic dataset covering the period from 02/01/2020 to 01/01/2021, in the Saihan District of Hohhot. The dataset comprises 3,150 reported traffic crashes, among which 424 involved casualties. In addition, it includes approximately 45 million taxi trip records, 275 million taxi order records, 9,654 points of interest (POIs), 7,524 hourly weather records, nearly 9.8 k road segments and 869 land-use degree (LUD) entries. The entire study area was partitioned into 948 hexagonal grids (about 1 km × 1 km each), and traffic as well as environmental features were aggregated at an hourly temporal resolution. For model development, the data were split chronologically into training, validation and test sets with a ratio of 6:2:2.

To benchmark the effectiveness of the proposed DTGN framework, we compared it against several representative approaches widely used in spatio-temporal prediction, as follows:

1) SARIMA [28]: The seasonal autoregressive integrated moving average model, which is effective for handling time series data characterised by periodicity and trends.
2) GCN-GRU [29]: An intelligent network traffic prediction model that utilises a joint attention mechanism with GCN-GRU, merging spatio-temporal properties to enhance prediction accuracy.
3) Hetero-ConvLSTM [3]: A hybrid model that integrates convolutional and long short-term memory networks.
4) GraphWaveNet [30]: A deep learning model that stacks graph convolutional layers to predict spatio-temporal graph data with long-term time series.
5) AGCRN [31]: A wavelet transform-based multi-scale graph convolutional recurrent network that captures dynamic spatio-temporal features across multiple scales.

4.2 Evaluation metrics

We used a variety of key evaluation metrics to comprehensively evaluate the DTGN model's performance in traffic crash risk prediction.

1) Acc (Accuracy): Measures how correct the overall predictions are.
2) RMSE (root mean square error): Quantifies the prediction error, with smaller values indicating lesser differences between predicted and actual values.
3) Recall: Evaluates the model's ability to identify crashes. In sparse data, a high recall rate suggests that the model effectively captures the occurrence of crashes, thereby minimising underreporting.
4) F1-score is used to measure the comprehensive performance of the model in weighing false alarms and underreporting.
5) Zero inflation rate: Measures the model's ability to handle sparse data, helping avoid over-predicting "no-crash" scenarios, improving the prediction accuracy in low-frequency crash regions.

4.3 Experiment results and analysis

Table 2 shows that DTGN achieves the best performance across all evaluation metrics by jointly modelling temporal dynamics, extracting spatial features and learning an adaptive adjacency matrix. It attains the lowest numerical error, accuracy, recall, F1-score, as well as the lowest zero inflation rate, thereby demonstrating DTGN can simultaneously adapt to complex spatiotemporal dependencies and data sparsity. In contrast, assuming a linear and stable model, SARIMA, it is challenging to capture nonlinearity and cross-node dependencies, with the highest error and lowest classification metrics performance. This validates the inherent shortcomings of traditional statistical paradigms in such tasks; among the deep learning models, GCN-GRU relies on first-order neighbourhood aggregation and single-channel GRU for temporal modelling. Although it has significantly improved compared to the baseline (Acc=75.5%, Rec=72.1%), due to limited capture of high-order topology and multi-scale dependencies, the RMSE is still high (19.127), and the F1-score is not dominant (0.796). Hetero ConvLSTM improves local spatiotemporal interactions through convolutional gating, but lacks explicit dynamic graph modelling, which results in only moderate recall and F1 (80.48%, 0.867). Nevertheless, its RMSE remains high (19.6584), and the zero-inflation rate is also elevated (ZIR = 33), reflecting insufficient sensitivity to structural evolution and data sparsity.

Table 2 – Validation results

Model	RMSE	Acc	Rec	F1-score	ZIR
SARIMA	28.2243	65.27	56.35	0.627	38
GCN-GRU	19.127	75.5	72.1	0.796	29
Hetero-ConvLSTM	19.6584	79.37	80.48	0.867	33
GraphWaveNet	15.9	85.31	83.25	0.894	27
AGCRN	20.7	84.5	84.42	0.901	21
DTGN	14.86	87.25	86.33	0.916	15

In more complex graph neural network models, GraphWaveNet relies on diffusion convolution and gated and dilated temporal convolution to effectively capture high-order topologies and long-term dependencies, thus performing well in regression accuracy and overall consistency. However, it relies on the static diffusion path and noise amplification, making the calibration of sparse zero values inferior to AGCGN (ZIR=27). AGCRN captures potential spatial relationships through adaptive graph learning and performs outstandingly in accuracy and F1-score, while outperforming most models in zero inflation control. Nevertheless, due to the absence of explicit temporal dynamic modelling, the results in the RMSE remain considerably higher than for DTGN.

In summary, GraphWaveNet and AGCRN exhibit complementary but limited strengths, with the former excelling in regression accuracy and long-term dependency modelling, and the latter in spatial adaptation and zero-inflation control. GCN-GRU and Hetero-ConvLSTM deliver moderate gains but remain constrained in capturing complex dynamics. By integrating dynamic temporal modelling with spatial representation fusion, DTGN achieves balanced improvements in error reduction and sparsity robustness, consistently outperforming all baselines across performance metrics.

4.4 Robustness analysis of the dynamic threshold

To further assess the robustness of the proposed dynamic thresholding method, we designed experiments from three aspects: data subset partitioning, time dimension rolling window analysis and threshold parameter sensitivity.

As shown in *Figure 7*, under different subsets of time and space (weekdays/weekends, day/night, city centre/suburbs), the proportion of HRZ remains in the range of 11%–20%, indicating that DTGN can stably identify a reasonable share of high-risk areas under varying conditions.

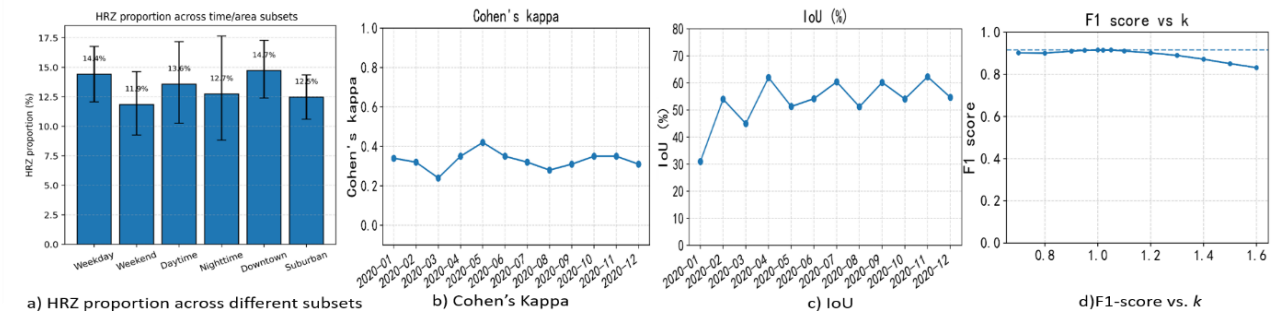

Figure 7 – Robustness evaluation of the dynamic threshold method

In *Figure 7(a–c)*, a monthly rolling window was applied to compute the HRZ proportion, IoU and Kappa coefficient. The results indicate that the overall HRZ proportion exhibited no significant fluctuations, with IoU values generally above 0.6 and Kappa coefficients mostly exceeding 0.4, suggesting that the spatial distribution of high-risk areas remained stable across different time periods. Furthermore, *Figure 7(c)* shows that Cohen's Kappa coefficient was predominantly above 0.4, reflecting moderate or higher consistency between adjacent time slices, while IoU values mostly exceeded 0.6, indicating a high degree of spatial overlap and consistency in the distribution of high-risk areas.

In addition, *Figure 7(d)* presents the sensitivity analysis of the threshold parameter k. The HRZ proportion and IoU exhibited no notable fluctuations, and the F1-score consistently remained above 0.90 within the range of $k = 0.9 - 1.1$, reaching its peak at $k = 1.0$. These results demonstrate that this method is robust to threshold perturbations and exhibits strong parameter stability.

To further verify the rationality of the adaptive weighting mechanism, we conducted weighted sensitivity experiments and visualised weight evolution to further verify the rationality of weight learning for the TS, TC and ARC indicators in constructing dynamic adjacency matrices.

The model was first evaluated under different fixed weight combinations (α, β, γ), and the dynamic weight changes were tracked throughout training. As shown in *Figure 8*, during the early stages of training, the weights fluctuate considerably but gradually converge and stabilise as training progresses. Among the three indicators, TS achieves the highest adaptive weight, followed by TC and ARC, suggesting that time synchronisation contributes most to performance improvement. The stability of the final curve further confirms the alignment between sensitivity analysis and adaptive weight learning.

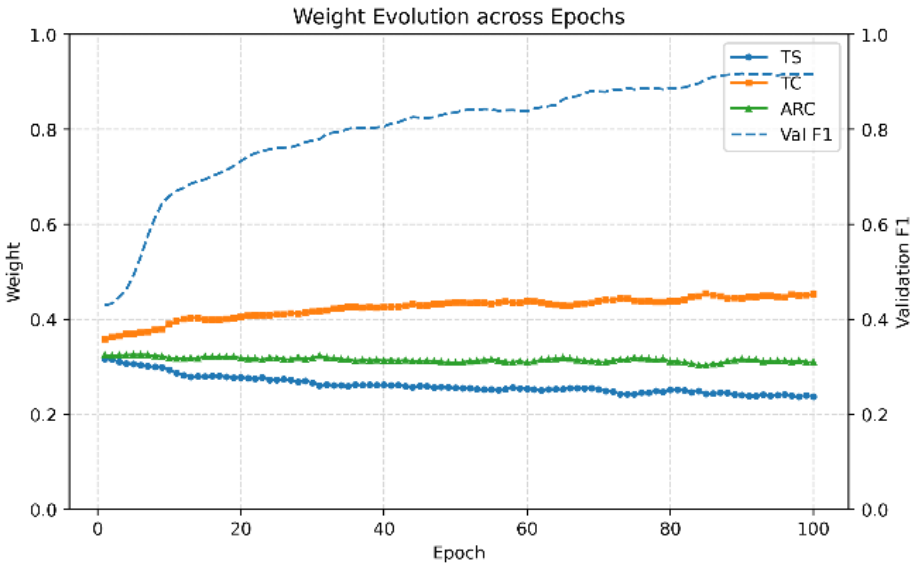

Figure 8 – Evolution of TS/TC/ARC weight learning during training

5. DISCUSSION

5.1 Contribution of different components

To further illustrate the effectiveness of the different components, we designed four variants for ablation experiments: (1) TEP block: This variant uses only the temporal module, removing the spatial feature modelling. (2) SPA block: Only the spatial module is retained, while the temporal feature modelling is removed. (3) No dynamic neighbourhood matrix: This variant combines features without introducing the dynamic neighbourhood matrix, only concatenating. (4) Complete DTGN: This variant incorporates temporal, spatial and dynamic neighbourhood matrix features to validate the effectiveness of the overall framework.

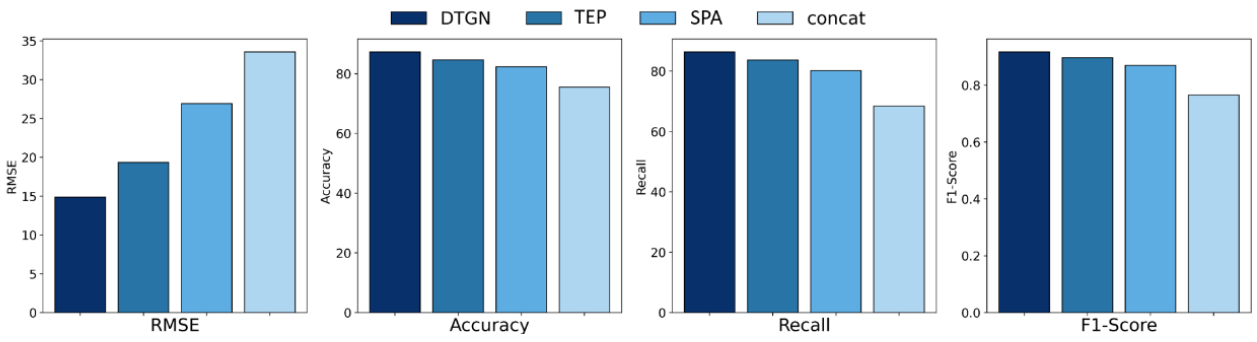

Figure 9 – Performance comparison of the complete model against ablations on all metrics

As can be seen from *Figure 9*, the complete model performs best on all indicators, indicating that the temporal block, spatial block extraction and fusion mechanisms synergise and work together to improve the model's prediction capability effectively. Notably, the performance of temporal blocks is better than that of spatial blocks, suggesting that temporal features have more influence on crash risk prediction, which may be related to the temporal patterns of traffic crashes, such as peak hours, weather factors, etc. Feature splicing alone performs poorly, and the dynamic adjacency matrix construction is sound. The DTGN model suggests that capturing both temporal and spatial blocks is essential for predicting the risk of traffic crashes.

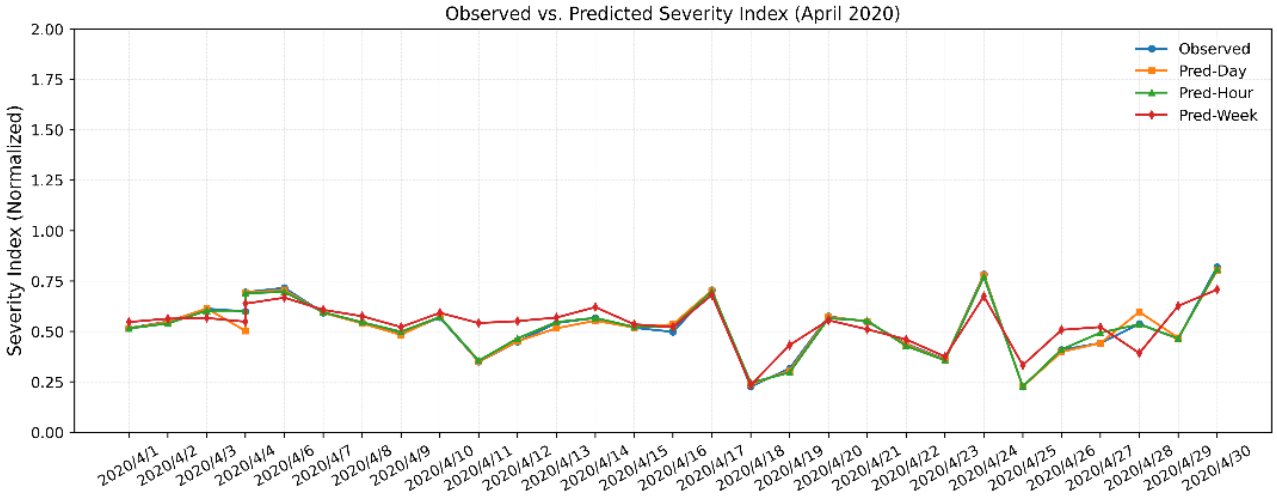

Figure 10 – Comparison of observed values and model predictions at different temporal resolutions

In addition, we further analysed the perspectives from temporal resolution and seasonal effects. As shown in *Figure 10*, the comparison between observed values and predicted results at different time resolutions highlights the importance of fine-grained temporal features for the model. The hourly and day level features can effectively capture the short-term fluctuations and cyclical patterns of traffic accident risks, such as morning and evening peak hours, differences between workdays and weekends, etc., allowing the prediction curves to closely align with observed values. By contrast, weekly features oversmooth the data, weakening peak–valley variations and causing the model to lose critical short-term dynamics and extreme-value information, thereby only reflecting long-term trends. Since traffic accident prediction's core goal is to identify short-term risk fluctuations, the higher accuracy and explanatory power of hourly and daily predictions are of greater practical significance.

5.2 Impact of different variables and feature groups

An intuitive analysis is conducted to examine the dependence of the deep graph neural network (DTGN) on different features and to reveal the influence patterns of each feature in crash-risk prediction. Accordingly, it is also necessary to assess the importance of both individual features and feature groups. We introduce the integrated gradients (IG) method, which measures the contribution of each variable by calculating the cumulative value of the gradient of the feature input along the path. The IG method effectively captures the variation in model response to different variables and provides a more explanatory ranking of feature importance. Meanwhile, to examine the impact of various categories of variables (e.g. traffic flow, weather, road structure) on the model's overall performance, we further analyse the importance of the feature groups.

To visually present the IG method's results regarding each feature's contribution, an importance plot for the IG is created based on the SHAP summary plot. The X-axis represents the importance value of the feature, the Y-axis represents the corresponding feature, and the colour maps the size of the feature value to reveal the mechanism of the influence of different features on the decision-making of the model and its mode of action. *Figure 11* displays the results.

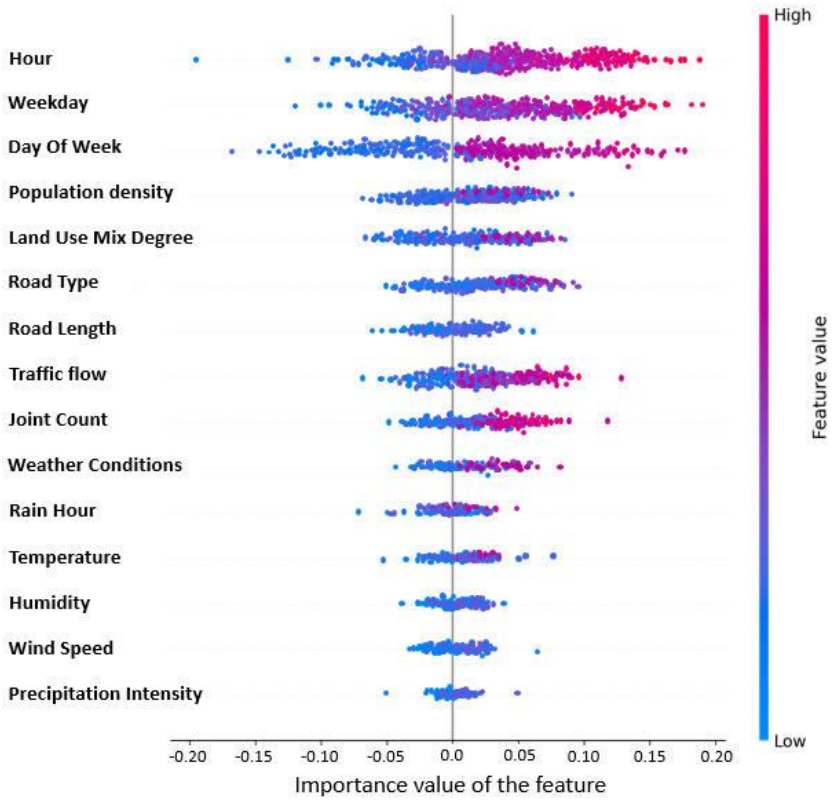

Figure 11 – Importance plot for the IG

Figure 11 illustrates that key factors influencing traffic crashes include time variables, traffic flow and the number of joints. The results show a specific temporal pattern in the occurrence of crashes, which contributes to the prediction at a high level. Day of the week has a relatively small effect, but distinguishes the difference between weekends and weekdays. Traffic flow ranks high, showing the significant effect of flow levels on crash risk, with either a flow rate too high or too low, increasing the likelihood of a crash.

Roadway network characteristics (e.g. roadway type and number of intersections) also play an important role, reflecting differences in risk across roadway configurations. This conclusion is consistent with Roland's paper's findings.

Weather factors significantly impact crash risk, with higher SHAP significance for rain hours and weather conditions. This suggests that inclement weather (e.g., heavy rain, dense fog) may reduce visibility and increase the risk of slippery road surfaces. In addition, temperature, humidity and wind speed may further increase crash risk during extreme weather conditions by influencing driving behaviour and traffic flow patterns.

Population density and land use mix degree (LUD) ranked low in the SHAP variable significance analysis. This is because there are no significant differences in the distribution of jobs and housing in the study area, resulting in a more limited impact on crash risk.

Although feature attribution analysis reveals the contribution of individual variables, it does not directly capture the robustness of the model when an entire feature modality is missing. We further evaluated the DTGN's performance under modality-missing scenarios by removing specific types of variables.

Table 3 – Validation results

Setting	Accuracy	Recall	F1-score	RMSE	ZIR	ΔF1 vs Full
All features	0.872	0.863	0.916	14.9	15	–
– Weather	0.82	0.785	0.858	19.8	24	−6.3%
– Road network	0.832	0.798	0.872	18.5	22	−4.8%
– POI / Taxi flow	0.844	0.815	0.884	17.2	20	−3.4%
– Socio-economic	0.865	0.849	0.903	15.9	16	−1.4%

As shown in *Table 3*, DTGN exhibits varying degrees of performance degradation under missing-modality conditions. The absence of weather variables has the most significant impact (F1 decreased by -6.3%), indicating that extreme weather information is crucial in predicting traffic accident risks. Road network variables have the second-largest effect (–4.8%), underscoring the importance of road structure in risk modelling. The absence of POI variables leads to a moderate decline (–3.4%), whereas socio-economic variables have little impact on overall performance (–1.4%). These results are consistent with the feature attribution analysis in *Figure 11*, further indicating that DTGN has a high dependence on key modalities, but still maintains strong robustness when some data sources are missing.

5.3 Small-sample adaptability

First, we selected the Yuquan District of Hohhot City as the testing area. The region was divided into multiple spatial units, and 10%, 30%, 50% and 100% of the dataset were used to conduct evaluations under different data scales.

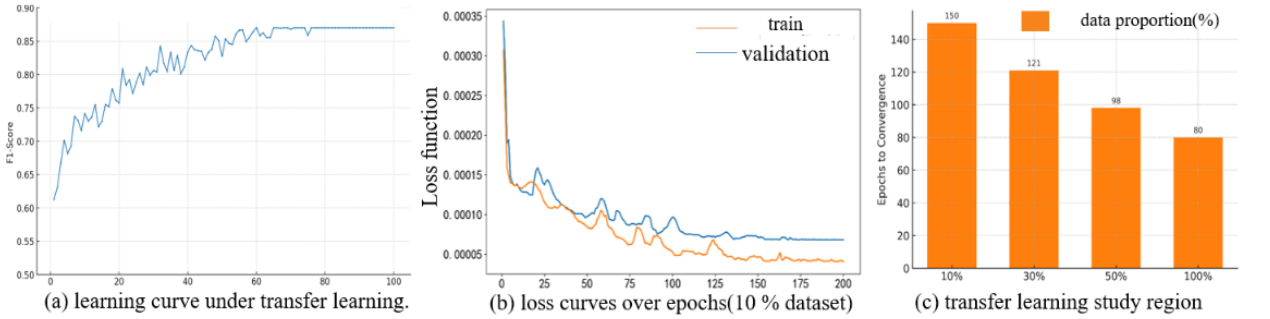

Figure 12 – Results of cross-district transfer learning experiments with DTGN

From the perspective of performance, *Figure 12(a)* shows that the F1-score of DTGN in the target area increases steadily with higher annotation ratios. At dataset ratios (10%–30%), the curve exhibits certain fluctuations, indicating model instability under small-sample conditions. When the dataset ratio exceeds 50%, the performance stabilises and converges, reaching about 0.86 at 100% annotation. This indicates that transfer learning can quickly restore performance under a limited dataset and reach near saturation levels in high annotation contexts.

The loss curve *Figure 12(b)* of the training and validation sets reveals the convergence characteristics of the model on a 20% subset of data. As the number of epochs increases, both training and validation losses decrease and eventually stabilise, following similar trajectories without signs of overfitting. It suggests that the model maintains good generalisation ability under small sample conditions.

In terms of convergence speed, *Figure 12(c)* shows that as the annotation ratio increases, the number of epochs required for convergence decreases substantially: the model requires approximately 150 epochs to converge with 10% annotation, whereas only about 80 epochs are needed under the 100% annotation condition.

In summary, the results of the cross-domain transfer learning experiment indicate that DTGN exhibits strong adaptability across domains: even under low annotation or partial data loss conditions, it can progressively recover performance.

6. CONCLUSION

Accurate prediction of traffic accident risk is essential for improving urban traffic safety, as it directly influences the effectiveness of management interventions. This study proposes a traffic accident risk prediction model (DTGN) that can achieve high-precision identification of accident-prone areas, providing a practical tool for urban traffic safety management. First, it integrates multi-source data, including road structure, traffic flow and historical crash records, into a unified framework for more complete risk characterisation. Second, it introduces dynamic threshold layering to improve prediction accuracy in low-frequency regions. Finally, it incorporates dynamic graph modelling with cross-attention mechanisms to address the shortcomings of static graph structures.

Despite the above advances, several limitations remain. The available data do not capture micro-level details such as driving behaviour or vehicle attributes, which restricts the depth of risk assessment. Feature

interpretability is limited, and causal mechanisms are challenging to quantify. Moreover, the model's online updating mechanism is still incomplete. Future work will seek to include external mobility and POI data, together with multi-scale spatial representations and real-time updating strategies, to enhance the model's precision, interpretability and adaptability.

ACKNOWLEDGEMENTS

This work was partly supported by the Program for Inner Mongolia Natural Science Foundation 2024LHMS05024 and 2025QN05122.

REFERENCES

[1] Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the 27th International Joint Conference on Artificial Intelligence;2018 July 13-19; Stockholm, Sweden.*2018. p.3634–3640. DOI: 10.48550/arXiv.1709.04875.

[2] Wang B, et al. GSNet: Learning Spatial-Temporal Correlations from Geographical and Semantic Aspects for Traffic Crash Risk Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence. 2021, Feb 2–9, virtually.* 2021;35(5):4402-4409. DOI: 10.1609/aaai.v35i5.16566.

[3] Yuan Z, Zhou X, Yang T. Hetero-Convlstm: A deep learning approach to traffic crash prediction on heterogeneous spatio-temporal data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018, 19-23 Aug, London, United Kingdom.* 2018. p.984–992. DOI: 10.1145/3219819.3219922.

[4] Yang D, et al. Urban rail transit passenger flow forecast based on LSTM with enhanced long-term features. *IET Intelligent Transport Systems.* 2019;13(10):1475–1482. DOI: 10.1049/iet-its.2018.5511.

[5] Chang LY, Chen WC. Data mining of tree-based models to analyze freeway crash frequency. *Journal of Safety Research*. 2005;36(4):365–375. DOI: 10.1016/j.jsr.2005.06.013.

[6] Kasatkina EV, Ketova KV, Vavilova DD. Development of analysis and forecast technologies for road crashes in the region and its application, *Proceedings Of The Iii International Conference On Advanced Technologies In Materials Science, Mechanical And Automation Engineering(Mip: Engineering-Iii – 2021).2021.29–30 April; Krasnoyarsk, Russian Federation.*2021,pp. 2402(1):070005. DOI: 10.1063/5.0071291.

[7] Ai Y, et al. A deep learning approach to predict the spatial and temporal distribution of flight delay in network. *Journal of Intelligent & Fuzzy Systems*. 2019;37(5):6029–6037. DOI: 10.3233/JIFS-179185.

[8] Zheng HF, et al. A hybrid deep learning model with attention-based Conv-LSTM networks for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems.* 2021;22(11): 6910–6920. DOI: 10.1109/TITS.2020.2997352.

[9] Hu Z, Zhou J, Huang K, Zhang E. A data-driven approach for traffic crash prediction: A case study in Ningbo, China. International Journal of Intelligent Transportation Systems Research. 2022;20(4):709–719. DOI: 10.1007/s13177-022-00307-3.

[10] Li H, Yu L. Prediction of traffic crash risk based on vehicle trajectory data. *Traffic Injury Prevention*. 2024;26(2),164–171. DOI: 10.1080/15389588.2024.2402936.

[11] Bao J, Liu P, Ukkusuri SV. A spatio-temporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention*.2019;122:239–254. DOI: 10.1016/j.aap.2018.10.015.

[12] Ma C, Zhang Y, Wang Q, Liu, X. Point-of-interest recommendation: Exploiting self-attentive auto-encoders with neighbor-aware influence. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.2018,22-26Oct.Torino, Italy*.2018.pp:697-706. DOI: 10.1145/3269206.3271733.

[13] Zhang Z, et al. Machine learning based real-time prediction of freeway crash risk using crowd sourced probe vehicle data. *Journal of Intelligent Transportation Systems*. 2022;28(1):92–106. DOI: 10.1080/15472450.2022.2061093.

[14] Zhao L. et al. T-GCN: A temporal graph convolutional network for traffic prediction .*IEEE Transactions on Intelligent Transportation Systems*.2020;21(9):3848-3858. DOI: 10.1109/TITS.2019.2935152.

[15] Zhou ZY, et al. Risk Oracle: A minute-level citywide traffic crash forecasting framework. *The Thirty-Fourth AAAI Conference on Artificial Intelligence. 2020 7-12 Feb. New York, USA.*2020; 34(1):1258–1266. DOI: 10.48550/arXiv.2003.00819.

[16] Tao SM, et al. Multiple information spatial–temporal attention-based graph convolution network for traffic prediction. *Applied Soft Computing.* 2023;136:110052. DOI: 10.1016/j.asoc.2023.110052.

[17] Wang J, et al. Temporal heterogeneity in traffic crash delays: causal inference from multi-scale time factors and sample-wise structural decomposition. *Accident Analysis & Prevention*. 2025;210:108220. DOI: 10.1016/j.aap.2025.108220.

[18] Choudhary A, Garg RD, Jain SS, Khan AB. Impact of traffic and road infrastructural design variables on road user safety – a systematic literature review. *International Journal of Crashworthiness,* 2023;29(4),583–596. DOI: 10.1080/13588265.2023.2274641.

[19] McCarty D, Lee D, Park Y, Kim HW. Exploring road safety through urban fabric characteristics and theory-driven prediction modeling with SEM-XGBoost. *Environment and Planning B: Urban Analytics and City Science.* 2024:52(2):2399-8083. DOI: 10.1177/23998083241259069.

[20] McCarty D, Kim HW. Risky behaviors and road safety: An exploration of age and gender influences on road crash rates. *PLoS One*, 2024;19(1):e0296663. DOI: 10.1371/journal.pone.0296663.

[21] Adenan ST, Lubis SRH, Mardiana D. Analysis of risk factor traffic crashes and implementation of road safety: A Systematic literature review. *Jurnal Kesehatan.* 2024;17(2):161–175. DOI: 10.23917/jk.v17i2.5354.

[22] Intini P, et al. Predicting traffic volumes on road infrastructures in the context of multi-risk assessment frameworks. *International Journal of Disaster Risk Reduction,* 2025;117:105139. DOI: 10.1016/j.ijdrr.2024.105139.

[23] Pavlou D, Christodoulou G, Yannis G. The impact of weather conditions and driver characteristics on road safety on rural roads, *Transportation Research Procedia,*2023;72:4081-4088. DOI: 10.1016/j.trpro.2023.11.369.

[24] Faria MV, et al. Assessing the impacts of driving environment on driving behavior patterns. *Transportation*. 2020;47:1311–1337. DOI: 10.1007/s11116-018-9965-5.

[25] Eltemasi M, Behtooiey, H. Examining the relationship between wind speed, climatic conditions, and road crashes in Iran. *Heliyon*. 2024;10(13):e33228. DOI: 10.1016/j.heliyon.2024.e33228.

[26] Guo LK, et al. LightGBM: A highly efficient gradient boosting decision tree. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*. *2017.4-9 Dec.2017,NY USA.2017.* p. 3149–3157.

[27] Wilcox, RR. *Introduction to robust estimation and hypothesis testing(Fourth Edition).* San Diego, CA, USA: *Academic Press, 2017*

[28] Deretić N, et al. SARIMA modelling approach for forecasting of traffic crashes. *Sustainability*. 2022;14(8):4403. DOI: 10.3390/su14084403.

[29] Shi HF, et al. AGG: A novel intelligent network traffic prediction method based on joint attention and GCN-GRU. *Security and Communication Networks*. 2021;9:1-11. DOI: 10.1155/2021/7751484.

[30] Qian QP, Mallick T. Wavelet-inspired multiscale graph convolutional recurrent network for traffic forecasting. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, 14-19 Apr; Seoul, Korea*.2024. pp. 5680-5684.DOI: 10.1109/ICASSP48485.2024.10446847.

[31] Xu Y, et al. Adaptive graph fusion convolutional recurrent network for traffic forecasting. *IEEE Internet of Things Journal.*2023;10(13):11465-11475. DOI: 10.1109/JIOT.2023.3244182.