Urban Informatics

ORIGINAL ARTICLE

Integrating a non-gridded space representation into a graph neural networks model for citywide short-term crash risk prediction

Gabriel Jurado Martins de Oliveira[1], Patricia Sauri Lavieri[1*] and Andre Luiz Cunha[2]

Abstract

Current applications of Graph Neural Networks in citywide short-term crash risk prediction have been limited by a gridded representation of space, which restricts the network's capability to effectively capture the spatial and temporal dependency of crash occurrences. In addition, a grided representation does not match most geographic units used for administrative purposes, limiting the use of crash risk predictions by practitioners. This paper applies a gated localised diffusion graph neural network (GLDNet) model to compare the use of two alternative geographic units, Mesh Block (MB) and grid, to forecast locations where crashes are likely to occur in a future time window. The GLDNet relies on a graph-based representation of geographic units and a weighted loss function to address the sparsity of crash occurrences. The tests are performed using crash data from the City of Melbourne, Australia, over a period of one year. The predictions are made at six-hour intervals, and the results show that the GLDNet consistently outperforms baseline methods, with differences in prediction accuracy from 10% to 21% in relation to historical average and benchmark deep learning models. In terms of geographic units, the MB-based GLDNet performed better than its grid counterpart, with differences in prediction accuracy of up to 12.3%. The better performance stems from the underlying information attached to the MB units (i.e., land use) and the network properties (i.e., degree of centrality), which enhance the GLDNet capability to identify crash risk in both central and peripherical areas. Regarding its applicability, the MB-based GLDNet directly integrates with other data sources, which provides contextual information about crash hotspots that helps decision-makers develop police patrolling and rescuing strategies.

Keywords Sparse events, Graph Neural Networks, Crash risk modelling, Spatial analysis, Non-gridded data, Australia

1 Introduction

The predictive hotspot mapping of short-term crash risk aims to forecast locations where crashes are likely to occur in a future time window based on historical data of crash occurrences. Identifying locations with a higher probability of crash occurrence provides valuable insights for implementing preventive strategies to improve road safety and can contribute to a more effective allocation of city resources (Mukhopadhyay et al., 2022). For instance, accurate crash forecasting can help transportation agencies to target undesired driver behaviours by supporting police patrolling tasks (Sieveneck & Sutter, 2021). Additionally, effective crash hotspot mapping can assist incident management systems in alleviating complications after crash occurrences by reducing rescue teams' response time and minimising traffic disruptions (Ite & Pande, 2016). Further, navigation

*Correspondence:
Patricia Sauri Lavieri
patricia.lavieri@unimelb.edu.au
[1] The University of Melbourne, Melbourne, Australia
[2] University of Sao Paulo, Sao Carlos, Brazil

applications can utilise dynamic crash risk maps to help drivers select safer routes (Li et al., 2016).

So far, most studies investigate short-term crash prediction at the intersection level (Hu et al., 2020) and corridor level (Basso et al., 2021; Li et al., 2020; Shi & Abdel-Aty, 2015), with only a few studies focusing on macroscopic (spatially aggregated) models for short-term crash risk prediction (Bao et al., 2019; Chen et al., 2016). The lack of studies with the latter focus stems from (1) the difficulty in forecasting sparse events, (2) the non-linear, hierarchical, and complex relationships between the variables that explain crash occurrences, and (3) limitations in the resolution and coverage of urban traffic data.

With the rise of big data and the establishment of deep learning methods in the past few years, there was a significant increase in approaches to address complex problems, including the prediction of spatiotemporal events at a city scale (Wang et al., 2020). For instance, deep learning algorithms have shown promising results compared with traditional statistical methods applied in near-term traffic prediction (Ma et al., 2017), network travel time estimation (Hou & Edara, 2018), safety planning (Cai et al., 2019), and short-term crash risk prediction (Arvin et al., 2021; Bao et al., 2019). However, despite the capability to model complex non-linear relationships through distributed and hierarchical feature representation, standard deep learning methods are constrained by the need to represent spatial data as grids (Bronstein et al., 2017; Ma et al., 2015). This need imposes several limitations to the analysis and the use of model outputs:

1. The shape limitations of grids (square or rectangular units of arbitrary sizes) reduce the possibilities of representing natural and built environments and account for specific geographical information of interest (e.g., lakes, airports, parks, and central business districts). This misrepresentation can distort the spatial correlation within the data and can influence model results (Zhang, 2020).
2. The lack of malleability of grid shapes imposes that grid cells located at the boundary of a study area commonly diverge from their counterparts in shape or size. These differences may result in the loss of information and the deterioration of model performance.
3. The grid-based representation is not very practical in real-world applications, requiring further translation to become useful to decision-makers (Zhang, 2020). For instance, policymaking and police enforcement are commonly based on predefined statistical/administrative areas (e.g., suburbs, local government areas, census tracts, and postal areas).

4. Compared to other geographic units, a gridded-data representation of the space may increase biases associated with the modifiable aerial unit problem in crash data analysis (Ziakopoulos & Yannis, 2020).

In contrast, Graph Neural Network (GNN) models, an extension of deep learning that operates on graphs, are not limited by a grid representation of the space, they rely on a network-based data representation (Wu et al., 2020; Wu et al., 2022; Zhang, 2020). However, the few studies that developed predictions of citywide short-term crash risk using GNN models adopted a grid representation of the space (Huang et al., 2022; Wang et al., 2021a; Wang et al., 2021b; Zhou et al., 2020a; Zhou et al., 2020b), which is a sub-optimal use of the methodological capability and limits the practical applicability of the model, as discussed earlier. In addition, all these studies consider multiple data sources to predict citywide short-term crash risk. While auxiliary data sources improve the quality of risk predictions, they also limit the current feasibility of the real-world application of GNN models. This is because they increase computational power requirements and demand data that may not be available at the city/region level in fine temporal resolution. As a result, GNN crash models are not commonly used by transportation practitioners and governmental agencies, which usually rely on simpler modelling approaches, such as historical averages.

In this sense, a robust citywide short-term crash risk prediction GNN model based on governmental administrative geographic units and solely considering the historical occurrence of crash events would create strong real-world application opportunities. Therefore, the current study has two objectives regarding crash hotspot location predictions. First, to evaluate the prediction accuracy gains that a GNN crash model can provide compared to a benchmark GNN model, the Spatio-Temporal Graph Convolutional Networks (STGCN), and traditional methods, such as Historical Average (HA) and AutoRegressive Integrated Moving Average (ARIMA), when the only data input is the historic crash occurrence information. Second, to evaluate whether a non-gridded space representation based on administrative geographic units improves prediction accuracy compared to a gridded space representation.

Our model adapts the GNN architecture proposed by Zhang and Cheng (2020) to predict hotspot mapping of crime events, which, similarly to crashes, are sparse events. The prediction of short-term crash risk is conducted for the city of Melbourne at the Mesh Block level (MB) and grid level over a period of one year, starting on January 1st, 2019. The MB unit is chosen because it is widely used by decision-makers to allocate a vast range of city resources as it is the smallest geographic area defined

by the Australian Bureau of Statistics (ABS) that serves as a building block for the larger regions of the Australian Statistical Geography Standard (ASGS). In this sense, we discuss the applicability of the developed GNN model when integrated with other sources of spatial data, including land use and road network, and provide potential implications to decision-makers allocating patrolling and rescue services.

2 Related studies

GNN models commonly combine graph convolution network (GCN) with recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent units (GRU), or gated temporal convolution network (TCN) for spatial and temporal dependency modelling. While the GCN[1] is analogous to the classical convolutional neural network (CNN) used for spatial dependency modelling in gridded data, all the other structures are well-known deep learning approaches for modelling sequential and time-series data. Several GNN models have been developed to predict spatiotemporal events. They were inspired by the early works of Yu et al. (2017) and Li et al. (2017) that modelled spatial and temporal dependency of the traffic flow by combining GCN and time-series deep learning structures to predict traffic speed in a future time window.

Referring to GNN models and citywide short-term crash risk prediction at a spatially aggregated level, Zhou et al., (2020a) adopted a spectral-based GCN with a time-varying affinity matrix to capture the dynamic of traffic conditions and obtain a gridded crash risk map. They proposed a data enhancement method based on the transformation of crash risk into a statistical crash indicator to address the sparsity of crash occurrences. In the same direction, Zhou et al., (2020b) integrated an LSTM structure into Zhou et al., (2020a) network to jointly predict crash risk at multiple spatial scales and temporal time steps. In contrast, Wang et al. (2021a) leveraged a multi-view approach and multi-data sources to capture the spatial and temporal dependency of crashes. The multi-view approach is defined by a module that combines CNN with GRU and a module that integrates spectral-based GCN with GRU. By combining both modules, the multi-view approach leveraged gridded and graph-data representation to modulate crashes' spatial and

temporal dependency and predict crash risk. The sparsity of crashes was addressed via a weighted loss function in this case.

More recently, Wang et al. (2021b) proposed another multi-view and multi-task approach to jointly forecast both fine- and coarse-grained crash risks based on multi-data sources. In this study, the authors captured the spatial and temporal dependency of crashes by leveraging gridded data at different scales and their previous multi-view approach with a LSTM instead of GRU. To address the sparsity of crashes, the authors relied on a weighted loss function again. Finally, Huang et al. (2022) combined a spectral-based GCN with a gated TCN defined by a tanh-style gating mechanism[2] to predict citywide short-term crash risk based on a gridded representation of the space. For that, the author constructed the weighted adjacency matrix considering the cosine similarity of the urban road network. The data enhancement method proposed by Zhou et al., (2020a) was adapted to dynamically consider the global risk of crash events and address the sparsity of crash occurrences.

Several spatial and temporal resolutions were utilised by the abovementioned studies for predicting citywide short-term crash risk. From the temporal side, the resolutions varied between 10 min to 60 min and the prediction horizons from 1 to 6-time steps ahead (10 min up to 3 hours). From the spatial side, grid sizes were not always explicitly reported. For the cases in which they are explicitly reported, the sizes ranged between 2.25 km^2 and 4 km^2. The focus on higher temporal resolutions in contrast to lower spatial resolutions stems from the auxiliary data sources' spatial and temporal resolution constraints rather than practical applicability. This is because auxiliary data sources, such as traffic-related data, are usually not openly available at a fine-grained spatial resolution for large periods of time.

The review above makes evident the focus of previous studies on gridded space representations with large grid sizes and complex frameworks to accommodate large data volumes and multiple data sources. In this context, the current study proposes to take a step back and test how much predictive improvement we can gain simply by adopting a more meaningful space representation without the need for additional data sources. The model used in this study combines a localised spatial-based

---

[1] GCN can be divided into two main streams according to the notion of graph convolution: spectral- and spatial-based approaches. The spectral-based approach defines the convolution in the spectral domain by a graph Fourier transform of the Laplacian matrix. In the spatial-based approach, the convolution is defined by aggregating features of neighbour nodes via an information propagation process. See Wu et al. (2022) and Wu et al. (2020) for a deeper discussion on GCN.

[2] A gated temporal convolution network is a convolutional neural network defined by a gating mechanism without recurrent structure that significantly speeds up the network computation time at the training step in comparison to RNN and the LSTMs-based models. The gate mechanism modulates the nonlinear information projected by a activate function (i.e., hyperbolic tangent activation function (tanh)) to control the information flow between the network layers. See Dauphin et al., 2017 for a discussion of gated convolutional networks.

GCN with a gated TCN for modelling the spatial and temporal dependency of crash occurrences. In contrast to the spectral-based GCN that relies on the eigen decomposition of the graph Laplacian to modulate the dependency between graph nodes, the diffusion process leverages a finite sequence of random walks to achieve the same goal. As a result, the spatial-based GCN is computationally faster and increases model transferability to other applications compared to spectral-based GCN, as any perturbation on a graph leads to a change of eigen basis (Wu et al., 2020). On the temporal side, the gated TCN is defined by a Rectified Linear Unit function (ReLU)-style gating mechanism and includes a base probability of crash occurrences. The ReLU alleviates the vanishing gradient problem compared to other activation functions (i.e., tanh). At the same time, the added base probability of crash occurrences mimics the background intensity of event occurrences in a self-exiting point process and helps the network to learn the temporal dependency of crashes. Like Wang et al. (2021a); Wang et al. (2021b), we utilise a weighted loss function to accommodate the sparsity of crash occurrences. In summary, the model adopted in this study requires lower computational resources than the earlier methods as it relies only on the historical occurrence of crash events, is not constrained by a gridded data representation of space and leverages a network structure that speeds up the modelling process. Table 1 summarises all the abovementioned studies, including ours.

3 Materials and methods

The purpose of a citywide short-term crash risk prediction is to generate probability distributions to indicate locations most likely to observe crash occurrences in the near future. City locations are defined by a spatial representation of space (i.e., geographic units), and the prediction of crash risk is defined as a predictive hotspot mapping problem, which is addressed using GNN models and traditional methods. The objective of the GNN model is to learn a mapping function F that indicates where crashes are more likely to occur in a future time step t (i.e., six-hour interval) based on a network representation of geographic units. Let N be the number of graph nodes (i.e., locations), then the tally of crash occurrences on a specific time step $t-1$ defines a graph signal x_{t-1} that serves as the input data for the GNN model. The prediction of crash risk based on a set of historical graph signals x_M is defined as:

$$y_t = F(x_{t-1} \cdots x_{t-M} \mid G) \tag{1}$$

where $x_{t-M} \in R^N$ is a graph signal defined by a M time window (i.e., the number of time intervals), G is a structured graph (described below) and $y_t \in R^N$ is the estimated graph signal at the time step. Figure 1 shows our proposed analytic framework.

3.1 Study area and data description

The GLDNet model is applied to the study area of the city of Melbourne, Australia, considering two types of spatial segmentation: MB and grid. The MB serves as the basic building block in the Australian Statistical Geography Standard (ASGC) for aggregating statistics into geographic areas. Therefore, the MB is the baseline geographic unit for allocating a vast range of daily city resources. In total, there are 1340 MBs in the city of Melbourne, where the largest MB has approximately 1.7 km^2, and the smallest has 0.0005 km^2. The MB average size, median size, and standard deviation are equal to 0.028 km^2, 0.009 km^2, and 0.10 km^2, respectively. Most

Table 1 Review summary of GNN models used for citywide short-term crash risk prediction

Authors	GNN framework	Multiple scales	Spatial representation	Prediction horizon	Multiple data sources	Sparsity of crash occurrences
Zhou et al. (2020a)	Time-varying spectral-based GCN	No	Grid units	10 min & 30 min	Yes	Statistical crash indicator
Zhou et al. (2020b)	Time-varying spectral-based GCN + LSTM	Yes	Grid units	10 min	Yes	Statistical crash indicator
Wang et al. (2021a)	CNN with GRU; spectral-based GCN with GRU	Yes	Grid units	1 hour	Yes	Weighted loss function
Wang et al. (2021b)	CNN with GRU; spectral-based GCN with LSTM	Yes	Grid units	2 hours	Yes	Weighted loss function
Huang et al. (2022)	Spectral-based GCN with a gated TCN	No	Grid units	3 hours	Yes	Statistical crash indicator
Ours	Localised spatial-based GCN with a gated TCN	No	Grid units and Mesh Blocks	6 hours	No	Weighted loss function

Data preparation
- Spatial representation of space (i.e., geographic units)
  - ➤ Mesh blocks
  - ➤ Grid units
- Data argumentation
  - ➤ Simple exponential smoothing (SES)

Step 1

The GNN model
- Network data representation of space
  - ➤ Undirect weighted graph $G = (V, E, W)$
- Model architecture
  - ➤ Gated temporal network (GNet)
    - ❖ Temporal layers
  - ➤ Localised graph diffusion network (LDNet)
    - ❖ Spatial layers
  - ➤ Parameter learning
    - ❖ Weighted loss function
- Model implementation
  - ➤ Kernel sizes, number of layers, time window, weighted loss function, batch size, number of epochs, and learning rate

Step 2

Performance measurement
- Prediction metrics
  - ➤ Mean hit rate
  - ➤ Wilcox Signed-Rank test (WSR)
- Impact of the spatial representation of space
  - ➤ Crash risk in central vs peripherical areas
  - ➤ Network properties (i.e., degree of centrality)
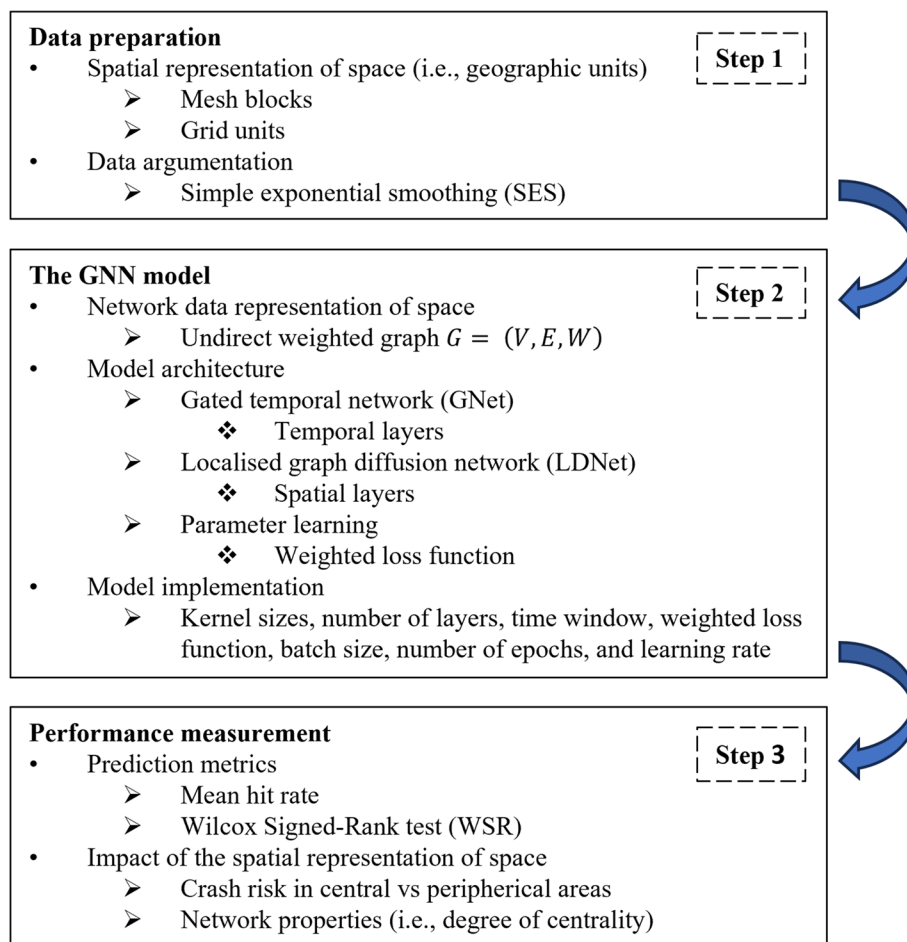
Step 3

Fig. 1 Analytic framework

Mesh Blocks contain between 30 to 60 dwellings and are broadly defined by population size and land use type, including residential, commercial, industrial and others. The larger MB units are broadly assigned to parks, ports, and industrial areas, which reduces the street network density and, thus, the number of locations where crashes may occur. In contrast, the smaller MB units are mostly assigned to residential and commercial areas. The grid units are developed to match the same number of MB geographic units; therefore, 1340 grid cells are created with an average size and standard deviation equal to 0.028 km^2 and 0.06 km^2, respectively. Figures 2 (a) and (c) present the spatial units for the study area, while Figures (b) and (d) present their associated network data representations.

The historical crash data is obtained from the Victoria Department of Transport and Planning and stems from police reports containing several crash attributes, including injury severity, crash type, the hour of the day, latitude and longitude location, mode of transport, and

others. The major street network is also obtained from the Victoria Department of Transport and Planning and contains the main roads for the city of Melbourne according to road users' hierarchy, including pedestrians, bicyclists, motorised vehicles and others.

After removing inconsistencies (missing crash location data), the historical crash dataset has a total of 714 crash occurrences (all types of crashes) observed from January 1st, 2019, to December 31st, 2019. The hourly crash occurrences are sparse, with a mean and standard deviation in the entire city equal to 0.08 and 0.29 crashes, respectively (minimum of 0 and maximum of 2 crashes in an hour). In this paper, crash occurrences are aggregated in six-hour intervals per geographic unit, and the predictions are performed one interval ahead. The six-hour intervals cover four periods of the day commonly associated with different human mobility patterns, including late night (i.e., 0h to 6h), morning (i.e., 6h to 12h), afternoon (i.e., 12h to 18h) and evening (i.e., 18h to 0h). The aggregation in six-hour intervals not only reduces the sparsity
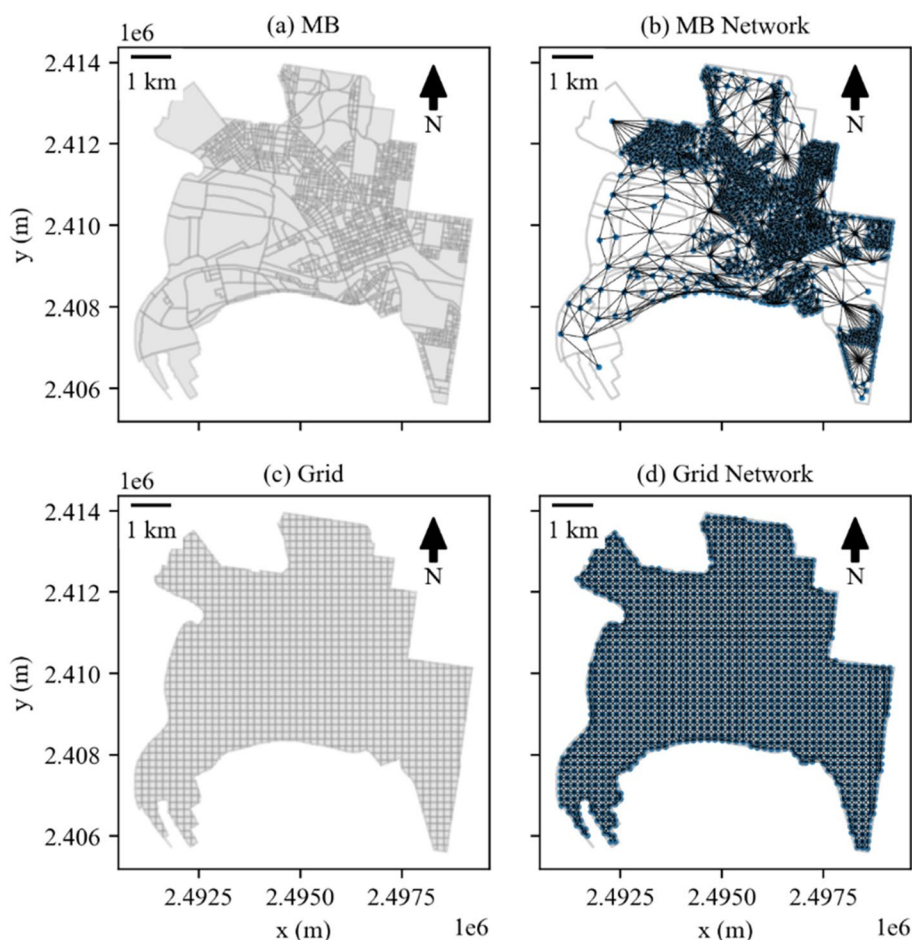
Fig. 2 The City of Melbourne and network data representation based on the two geographic units

of crash occurrences, but also is in alignment with several city services, including police patrolling, traffic management, and rescue system, as resource allocations are usually decided within hours in advance (i.e., based on periods of the day). The proportion of crash occurrences for the four intervals is 5.7%, 33.8%, 34.3% and 26.2%, respectively.

Because the occurrence of crashes per unit of area is sparse, we use a simple exponential smoothing (SES) technique to augment the data in the temporal domain for model training and validation (Zhang & Cheng, 2020). The SES exponentially decreases the weight of past observations based on the equation $s_t = a_s x_t + (1 - a_s)s_{t-1}$, $t > 0$, where a_s is the smoothing factor and is defined between 0 and 1. Smaller a_s leads to smoother augmented data. In this paper, we set a_s equal to 0.5.

3.2 Model
The model adapts the framework proposed by Zhang and Cheng (2020) to predict sparse events. In the following sub-sections, we describe the network representation of

geographic units, the model architecture, and its components, including the spatial and temporal structures. Lastly, we discuss the parameter learning process.

3.2.1 *Network Data Representation*
Compared to standard deep learning methods, GNN models operate on a structured graph $G = (V, E, W)$, where V, E, and W are the graph nodes, edges, and weight matrix, respectively. An undirect graph G is utilised to represent the data and predict the probability distributions of crash occurrences in a future time step. The undirect graph represents the geographic units (e.g., mesh blocks or statistical areas) as a set of graph nodes, while the edge represents whether two geographic units i and j are neighbours. Two geographic units are considered neighbours if they share borders by one or more points. The weight matrix $w_{ij} \in W$ represents the relationship between two neighbour geographic units e_{ij} and is defined based on the idea that the similarity between objects decays with the increase of the spatial distance. Although several decay

functions may be used to represent the spatial similarity between objects, the Gaussian kernel function is commonly used in GNN frameworks. For this reason, an edge weight w_{ij} is defined to be inversely proportional to the Euclidian distance between the centroids of two neighbour areas according to a Gaussian kernel function (Anselin & Rey, 2014):

$$w_{ij} = \begin{cases} \dfrac{\exp\left(-\left(\dfrac{dist_{i,j}}{h_i}\right)^2\right)}{\sqrt{2\pi}}, if \quad e_{ij} \; is \; an \; edge \\ 0 \; otherwise, \end{cases} \quad (2)$$

where h_i is the kernel bandwidth and $dist_{ij}$ is the distance between the centroids of two neighbour geographic units i and j.

3.2.2 Architecture

The model GLDNet is defined by integrating a localised graph diffusion network (LDNet) layer, a gated temporal convolution network (GNet) layer and a fully connected layer (Zhang & Cheng, 2020). The gated network processes the input sequence through L GNet hidden layers to learn how the influence of historical crash occurrences spread through time. Then, the processed information is passed to the K LDNet hidden layers to learn how crashes propagated across the space. Next, the information processed by both GNet and LDNet is fed into a fully connected layer, which transforms it into a predictive mapping. The mapping indicates the probability of crash occurrence in each region of the city in a future time step. The crash model is formulated as:

$$GLDNet\,(X) = g_K\,W_{fc} + b_{fc} \quad (3)$$

where GLDNet is the gated localised diffusion network, $g_K \in \mathbb{R}^{N \times m^k}$ is the k-th hidden layer of the diffusion network, $W_{fc} \in \mathbb{R}^{m^k}$ and $b_{fc} \in R^N$ are the fully connected layer learnable parameters. Figure 3 presents the proposed crash model framework.

3.2.3 The GNet Component

This component is a gated Temporal Convolution Network (TCN) for modelling the temporal propagation of crash occurrences. The network utilises a ReLU-style gating mechanism, has only one gate (defined by a sigmoid activation function), and consists of L hidden layers, denoted $h_0, \cdots, h_L$. The ReLU allows the gradient to easily propagate in comparison to other activation functions (i.e., tanh) (Dauphin et al., 2017; Jahan et al., 2022), while the single gate significantly speeds up the network computation time at the training step, particularly in the case of deeper networks (i.e., large number of hidden layers). Let $X^l \in \mathbb{R}^{N \times n^{l-1}}$ be the input of the layer h_L and $X^{l+1} \in \mathbb{R}^{N \times n^l}$ be the output, where N and n^l are respectively the number of graph nodes and the dimension of the nodes features of the l-th hidden layer. Then, the l-th is formulated as:

$$X^{l+1} = h_l\left(X^l\right) = ReLU\left(X^l W^l + b^l\right) \bigodot \sigma\left(X^l V^l + c^l\right) + d^l \quad (4)$$

where $W^l, V^l \in \mathbb{R}^{n^{l-1} \times n^l}$, $b^l, c^l \in \mathbb{R}^{n^l}$ and $d^l \in R^N$ are learnable parameters, $ReLU = \max(0, x)$, and $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. The output X^{l+1} of a layer h_l is the non-linear projection $ReLU(X^l W^l + b^l)$ modulated by the gate $\sigma(X^l V^l + c^l)$ with an added base probability of future crash occurrences for each node d^l. The base probability mimics the background intensity of event occurrences in a self-exiting point process (Zhang & Cheng, 2020). The gated network is defined by stacking multiples layers as:

$$GNet(X) = h_L(h_{L-1}(\cdots h_2(h_1(X))\cdots)) \quad (5)$$

3.2.4 The LDNet Component

The propagation of crashes across space can be considered to follow a diffusion process that reaches a stationary distribution after a finite sequence of random walks in a graph G (Teng, 2016). The stationary distribution of the graph nodes is obtained by a k-th random walks
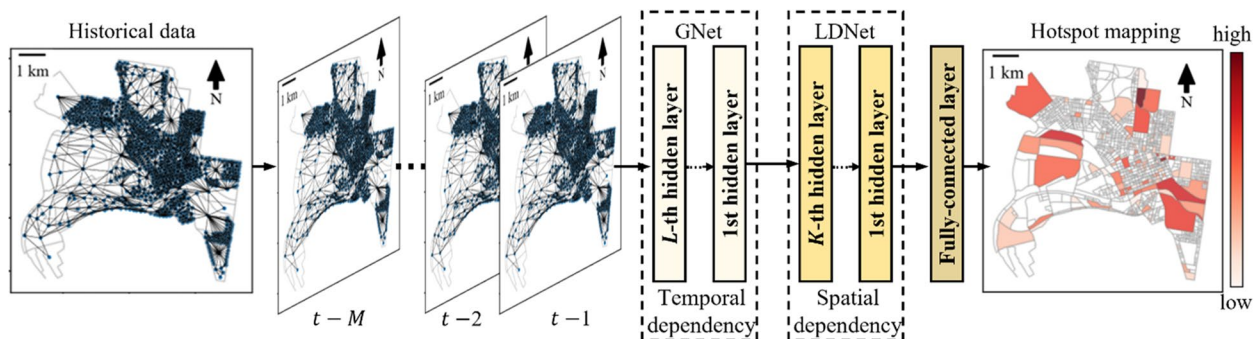

Fig. 3 GLDNet framework adapted from Zhang and Cheng (2020)

defined by a transition matrix $P = D^{-1}W$, where, p_{ij} in $P \in R^{N \times N}$ correspond to the probability of crash occurrences at node i influence the occurrence of crashes at node j, $W \in R^{N \times N}$ is the weight matrix of the graph G, and $D \in R^{N \times N}$ is the diagonal of W. More precisely, the k-th step random walks measure the extent a graph node i is affected by its k-order adjacent neighbours. The one-step walk captures the spatial dependency between a node i-th and its immediate neighbours, while the k-step walk captures the spatial dependency between the node i and its k-th adjacent neighbours.

A localised graph diffusion network with K hidden diffusion convolutional layers, denoted $g_0, \cdots, g_K$, is utilised for modelling the spatial dependency of crashes. Let $X^k \in \mathbb{R}^{N \times m^{k-1}}$ be the input of the layer g_K and $X^{k+1} \in \mathbb{R}^{N \times m^k}$ the output, where N and m^k are respectively the number of graph nodes and the dimension of the nodes features of the k-th hidden layer. Then, the k-th hidden layer is formulated as:

$$X^{k+1} = g_k\left(X^k\right) = ReLU\left(X^k * \theta^k + PX^k * \eta^k\right) \quad (6)$$

where $ReLU = \max(0, x)$ is the activation function, $X^k * \theta^k$ captures the dependency of each node itself and PX^k represents the one-step random walk. θ^k and $\eta^k \in \mathbb{R}^{N \times m^{k-1} \times m^k}$ are learnable parameters in k-th hidden layer. The localised graph diffusion network is constructed by stacking multiple hidden layers and uses a localised parameter-sharing scheme to capture the heterogeneity of crashes over space. Each added layer allows the network to capture the spatial dependency between a graph node i and its k-order adjacent neighbours. For an input X, the diffusion network is defined as:

$$LDNet(X) = g_K\left(g_{K-1}\left(\cdots g_2\left(g(X)\right)\cdots\right)\right) \quad (7)$$

3.2.5 Parameter Learning
A weighted loss function is used to address the sparsity of crash occurrences and train the GLDNet model, as realised in Wang et al. (2021a). Compared to traditionally mean-square error (MSE) and mean absolute error (MAE), the weighted loss function addresses the unbalanced regression problem by assigning a higher misprediction cost to graph nodes with crash occurrences, which enables the crash model to learn the locations with a higher probability of crash occurrences. The model can be trained via backpropagation and the weighted loss function is defined as:

$$loss = \frac{1}{N} \sum_{i=0}^{N} \omega_i\left(\hat{y}_i - y_i\right)^2 \quad (8)$$

where N is the number of nodes, the $\hat{y}_i$, y_i, are, respectively, the predicted and observed values at the i-th node, and ω_i is the weight assigned to each square error $\left(\hat{y}_i - y_i\right)^2$. Let $\rho \in [0, 1)$ be a predefined coefficient, and then the weight ω_i is defined as y_i, if $y_i > 0$ (crashes occurred at the i-th node) and ρ otherwise. Therefore, smaller ρ leads to lower misprediction costs at graph nodes without crash occurrence in comparison to nodes with crash occurrence.

3.3 Performance Measurement
To meet our objectives, we first compare the GLDNet performance against baseline methods, then we investigate the effect of the spatial unit choice on the prediction success.

3.3.1 GLDNet comparison against baseline methods
The GLDNet model is compared with four baseline methods commonly used for predicting spatiotemporal events, including short-term crash risk (Bao et al., 2019, Wang, S. et al., 2021a, b). The baseline methods include the Spatio-Temporal Graph Convolutional Networks (STGCN), Historical Average (HA), AutoRegressive Integrated Moving Average (ARIMA) and Gradient Boosting Regression Tree (GBRT). The STGCN is a benchmark GNN model widely used for predicting spatiotemporal events (Yu et al., 2017). HA is an approach widely used by transportation practitioners and governmental agencies in the absence of more robust models. ARIMA is a classical time series regression model that integrates the autoregressive, difference and moving average components of a time series (Box & Pierce, 1970). GBRT is a tree-based ensemble method that combines the prediction of multiple sequential tree-based models to obtain better prediction performance (Zhang & Haghani, 2015). The STGCN parameters are defined following the authors' recommendations. We consider a time window of sixteen observations (i.e., four days), as the model computational costs significantly increase for large time windows. The ARIMA and GBRT models' parameters are defined using a grid search strategy and fine-tuned with the same input data to conduct a fair comparison. The HA, ARIMA and GBRT models are applied individually to each geographic unit.

Although a comparison between the GLDNet and other GNN crash models (such as the ones described in Section 2) would be of great value, the replication, implementation and calibration of these models is not a trivial task. This is because the codes and required data sets are often not available, and the replication of

Jurado Martins de Oliveira *et al. Urban Informatics* (2023) 2:7

Page 9 of 17

the models would be extremely time-consuming. Furthermore, such models were developed for specific network structures and removing or adapting fundamental components of their architectures would likely jeopardise their usability and predictive capability. Finally, the models described in our review rely on multiple data sources and thus would not lead to a fair comparison with the model used in this paper, which relies only on historic crash occurrence data.

The mean hit rate is used to compare models directly, and the Wilcoxson Signed-Rank (WSR) test is performed to evaluate whether the prediction results of distinct models are statistically different at a 90% confidence level. While standard metrics, such as mean squared error (MSE) and mean absolute error (MAE), are commonly applied to measure the prediction of dense spatiotemporal events (e.g., traffic speed and flow), these metrics are poorly suited for predictive hotspot mapping of sparse events (Adepeju et al., 2016). The standard metrics are strongly influenced by the large number of zero crash counts (i.e., they are not designed for zero-inflated problems) as they are designed to evaluate the mean performance over all space and time, which may lead to an underrepresentation of crash risk in the more relevant hotspot locations.

The hit rate is defined as the number of events accurately captured by the hotspot locations divided by the total number of events and thus is not influenced by a large number of zeros. Therefore, the hit rate is a straightforward interpretable metric used to evaluate the performance of predictive hotspot mapping of sparse events (Bowers, 2004, Zhang & Cheng, 2020). The hit rate is calculated by sorting all regions in descending order by their predicted values and by tailing the proportion of events that fall on the sorted regions. In this paper, a maximum of 30% coverage level (i.e., 30% of the geographic units) is considered for computing the hit rate. This is because, at higher coverage levels, any model tends to have high performance as most locations of the study area are likely to observe a crash occurrence, and therefore, the prediction becomes non-informative for decision-making (i.e., at a 100% coverage level, any model has 100% of accuracy). The mean and standard deviation of the hit rate aggregated over all consecutive testing periods are considered to evaluate the performance of the prediction. The hit rate is formulated as:

$$HR = \frac{n_{si}}{N_i} \tag{9}$$

where HR is the hit rate, n_{si} and N_i are, respectively, the number of crashes that fall within a coverage area s and the total number of crash occurrences within the entire study area during a time window i.

Although the mean hit rate can be used to directly compare the results of different prediction methods, the statistical significance of the results is unknown (Adepeju et al., 2016). Therefore, to assess the significance of the results, the Wilcox Signed-Rank (WSR) test is used to evaluate whether the predictions obtained with different methods are statistically different. For that, it is assumed that the underlying distributions of crash occurrences do not change over time and the hit rates of two methods at a given coverage level are treated as paired samples. Although the temporal instability of crash-related factors should be considered in crash data analysis (Mannering, 2018), short periods (such as the testing sample) are less likely to be affected. Therefore, it is reasonable to assume that the underlying distribution of crash occurrences does not change. The WSR is a distribution-free test that assesses whether the mean population rank of two related samples differs. The WSR test statistic is given by:

$$W_{WSR} = \sum_{i=1}^{N} \left(\text{sgn}(y_{1,i} - y_{2,i}) \cdot R_i \right) \tag{10}$$

where N is the sample size, *sgn* is the sign function used to extract the sign of a real number, $y_{1,i}$ and $y_{2,i}$ are the hit rate on test time interval i from models 1 and 2, respectively. R_i is the rank of the difference $y_{1,i} - y_{2,i}$. The statistical significance of W_{WSR} is obtained using a single-tailed lookup table.

3.3.2 *The impacts of spatial representation on GLDNet performance*

The impacts of spatial unit choice on the GLDNet performance are discussed based on considering the effects of variability in crash occurrence across space and time on the model's predictive capability. Across space, the GLDNet prediction performance based on each spatial unit is evaluated for central and peripherical areas and explained in terms of spatial information (i.e., land use, population) and network properties (i.e., network density and degree of centrality). On the other hand, to investigate the effects of variability in crash occurrences across time, the GLDNet performance is evaluated according to the number of crash occurrences per time interval.

3.4 Implementation
The GLDNet model is implemented using a GPU-version PyTorch Geometric Temporal 0.51.0 (Rozemberczki et al., 2021). A grid search strategy is executed for tuning four model hyper-parameters, the parameter ρ

in the loss function and the time window M. For training the model, we set the batch size equal to 50, learning rate equal to 0.001 and ran 30 epochs with the Adam optimiser (Kingma & Ba, 2015). Table 2 shows the grid search space and, in bold, the optimal parameters for the MB model and in italics for the grid counterpart (in most cases, they are the same). The optimal parameters are defined based on the historical data and are discussed in detail in Section 5.3.

The weight matrix representing the similarity between two neighbours' geographic units is constructed by setting the Gaussian kernel function with a fixed bandwidth h equal to the maximum distance among all the geographic unit's nearest neighbours, 0.80 km, and 0.17 km for MB and grid, respectively. The training, validation, and testing samples are set to 60% (876 intervals), 15% (219 intervals) and 25% (365 intervals), respectively.

4 Results and discussion

The results are presented as follows. First, the GLDNet model is compared against baseline methods. Second, the impacts of spatial unit choice on the GLDNet

performance are discussed in detail, including the effects of variability in crash occurrence across space and time on the model's predictive capability. Lastly, a sensitivity analysis of the GLDNet parameters is presented.

4.1 Comparison of the GLDNet model against baseline methods

Table 3 presents the mean hit rate for the baseline methods and the GLDNet model for the two geographic units. In regard to traditional methods, the results show that the GLDNet performs better than the HA, ARIMA and GBRT for most cases, except at a 10% coverage level for the grid units. Interestingly, even for lower coverage levels (5% and 10%), where a small number of spatial units consistently have crashes, the MB-based GLDNet model outperforms all traditional methods. At higher coverage levels, the differences between simpler and more robust methods increase as more locations with a relatively lower frequency of crash occurrences or greater crash risk variability are included in the analysis. In this scenario, the models must be able to capture variability to obtain an improved performance, which is the case of the GLDNet. The best relative performance of the GLDNet is observed for coverage levels of 20% and 25%.

Concerning more robust methods, the results show that the GLDNet performs better than the STGCN in all coverage levels for the MB units, with the best relative performance observed at a 15% coverage level. The difference between the GLDNet and STGCN performance for grid units is lower, with the GLDNet outperforming its STGC counterpart at coverage levels higher than 15%. In this case, the best relative performance is observed at a 30% coverage level.

Table 2 Details of the grid search for GLDNet

Model	Searching range of hyperparameters
GLDNet	number of L Gnet layers: [1,***2***,3,4,5] number of K Dnet layers: [1,***2***,3,4,5] dimension of the node features n^l: [2, 4, ***8***, 16, 32] dimension of the node features m^k: [2, ***4***, 8, 16, 32] ρ: [0.001, *0.05*, 0.01, **0.02**, 0.025] time window M: [30, 60, 90, ***120***, 150]

Table 3 Mean hit rate for the GLDNet and baseline methods for the two geographic units

Coverage level	Spatial unit	GLDNet	STGCN	HA	ARIMA	GBRT	GLDNet vs. STGCN	
							Difference[a]	*p*-value
5%	MB	36.6%	32.0%	31.2%	28.6%	28.3%	4.5%	0.0421
10%		52.8%	43.4%	44.7%	39.5%	39.8%	9.4%	0.0004
15%		59.6%	49.1%	47.9%	49.3%	49.3%	10.5%	0.0003
20%		66.8%	62.9%	49.7%	51.7%	51.7%	4.0%	0.0142
25%		69.8%	64.9%	53.4%	54.3%	54.3%	4.9%	0.0196
30%		71.3%	67.8%	57.4%	56.8%	56.8%	3.5%	0.0826
5%	Grid	26.1%	27.4%	20.2%	24.0%	25.1%	-1.2%	0.6815
10%		40.5%	38.4%	38.5%	39.5%	39.1%	2.1%	0.2314
15%		53.3%	51.1%	43.2%	52.0%	52.0%	2.1%	0.2228
20%		62.0%	57.1%	43.2%	53.5%	53.5%	4.9%	0.0569
25%		66.6%	61.0%	45.7%	55.4%	55.4%	5.6%	0.0264
30%		67.4%	61.6%	52.2%	57.4%	57.4%	5.8%	0.023

[a] Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

Overall, the GLDNet better performance is explained by its capability to simultaneously capture the spatial and temporal dependency of crash occurrences, which shows that leveraging a graph-neural network framework designed to account for the sparsity of crash events provides significant accuracy gains compared to traditional baseline methods for predicting citywide short-term crash risk based only on historical crash occurrences, including standard graph neural network models.

4.2　Impacts of spatial unit choice on the GLDNet performance

Table 4 presents the comparison between the mean hit rate for the GLDNet model using the two geographic. The results show that the MB-based model performs better than its grid counterpart, with a higher mean hit rate for all coverage levels. The differences are statistically significant at a 90% confidence level for up to a 20% coverage level. It is expected that the differences between both models decrease when the coverage levels increase as prediction improvements become more marginal because more areas are considered. This explains the statistically insignificant difference in the model's performance at the 25% and 30% coverage levels. These results

Table 4 Comparison of the mean hit rate for the GLDNet model based on the two geographic units

Coverage level	MB	Grid	Difference	*p*-value
5%	36.6%	26.1%	10.5%	0.0118
10%	52.8%	40.5%	12.3%	0.0046
15%	59.6%	53.3%	6.3%	0.0710
20%	66.8%	62.0%	4.9%	0.0803
25%	69.8%	66.6%	3.2%	0.1612
30%	71.3%	67.4%	3.9%	0.1139

point to the potential benefits of adopting administrative geographic units compared to gridded space representations for predicting citywide short-term crash risk, as we further discuss in Sections 4.2.1 and 4.2.2.

4.2.1　Effects of variability in crash frequency across space

As the spatial sparsity of crashes may influence the prediction performance of GLDNet, we examine the model performance according to the city region. To illustrate the effects of spatial sparsity on the model performance, we use the 12h-18h interval of a random day in the test sample. Figure 4 presents the predictive mapping of crash risk at a 10% coverage level (the green and blue dots represent crash occurrences captured and not captured, respectively), as this is the level with the greatest difference in predictive accuracy, as shown in Table 4. Comparing both subfigures, the MB-based spatial distribution of the crash risk is spread all over the city of Melbourne, while the grid-based model leads to a spatial distribution of crash risk that is concentrated around the city business district (CBD). The CBD has an area of 2.4 km^2 (only 6.3% of the city of Melbourne's total area) and concentrates a large portion of the crash occurrences over the study period (30%). In this sense, this example illustrates how the grid-based network may reduce the GLDNet capability to capture crash risk in areas with lower crash density (outside the CBD). On the other hand, the MB-based model identifies the crash risk in both central and peripherical areas.

Table 5 compares the GLDNet prediction performance based on each spatial unit for CBD and outside CBD areas. Concerning the CBD area, we observe that the grid-based model has a higher mean hit rate and statistically different results from its MB counterpart for all coverage levels. On the other hand, for outside CBD areas,
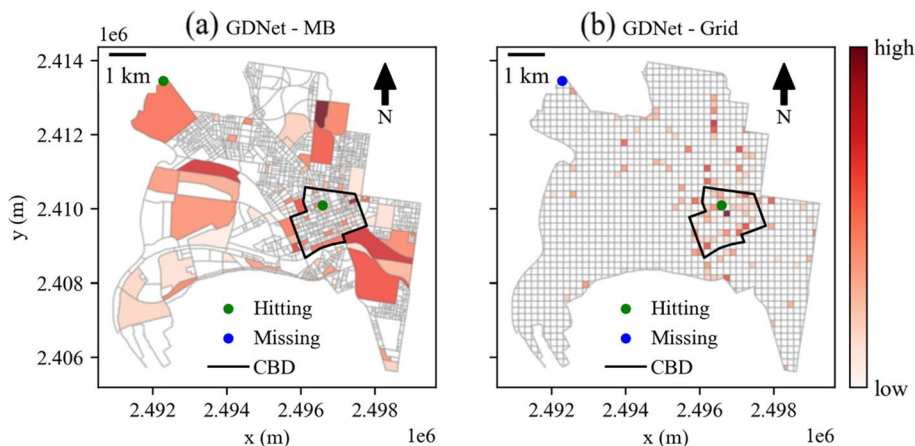

Fig. 4 GLDNet predictive hotspot mapping of crash risk for a random day at a 10% coverage level based on the two geographic units

Table 5 GLDNet mean hit rate based on city regions for the two geographic units

Coverage level	CBD				Outside CBD			
	MB	Grid	Difference*	*p*-value	MB	Grid	Difference[a]	*p*-value
5%	11.6%	14.9%	-3.3%	0.0924	24.9%	11.2%	13.7%	0.0006
10%	13.5%	19.5%	-6.0%	0.0030	39.3%	21.0%	18.3%	0.0000
15%	15.8%	21.9%	-6.1%	0.0051	43.8%	31.3%	12.4%	0.0013
20%	18.4%	24.7%	-6.3%	0.0081	48.4%	37.2%	11.2%	0.0008
25%	19.4%	25.5%	-6.1%	0.0051	50.4%	41.1%	9.3%	0.0089
30%	19.8%	25.5%	-5.7%	0.0091	51.5%	41.9%	9.6%	0.0053

[a] Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

we observe the exact opposite. However, the mean hit rate differences between MB- and grid-based models for the CBD and non-CBD areas vary greatly. For instance, for CBD areas, the differences between the MB- and grid-based models are up to 6.3%, while for the non-CBD areas, the differences are up to 18.3%. As a result, for the entire study area, the MB-based model outperforms its grid counterpart with a higher mean hit rate for all coverage levels, as shown in Table 4.

The MB-based GLDNet capability to capture crash risk outside central areas is explained by two main factors: (1) the underlying information associated with the definition of the boundaries of MB units and (2) network properties. First, spatial information, such as land use, number of dwellings, and road networks, are utilised for developing MB units. Such variables are also associated (correlated) with the likelihood of crash occurrences, and thus, affect the prediction of short-term crash risk. Second, regarding network properties (Newman, 2018), the MB has a higher average degree of centrality, closeness centrality, K-core number and clustering coefficient compared to the grid representation. In addition, the distribution of these MB and grid network indicators are significantly different at a 99% confidence level based on the non-parametric statistical Kolmogorov-Smirnov (KS) test. Further, the MB representation also has a higher network density. These properties are global network metrics commonly used for classifying how information propagates in social networks (i.e., Twitter). Higher values are associated with networks where the information is more likely to be widely spread through its nodes, while lower values tend to reflect a network structure that hinders the spread of information (Conover et al., 2012; Pierri et al., 2020). In this sense, the MB-based network facilitates the spread of information (crash occurrences) among its nodes in comparison to the grid-based network, thus enhancing the GDNet capability to capture crash risk outside central areas.

While underlying factors associated with the definition of MB boundaries improve the model performance in

low density areas, they are also probably associated with the lower performance of this spatial unit in high density areas. The CBD is homogeneous in terms of land use and road networks but is subdivided into multiple units because each MB unit is designed to host a similar number of dwellings. In other words, there is high granularity in space representation but low variance in crash occurrence, which increases prediction error. For instance, the total number of grid units that cover at least a portion of the CBD is 115, while for the MB units, this number more than doubles (283). In this sense, strategies to merge some of the geographic units in high density areas or alternative geographic units that are not highly influenced by the number of dwellings are likely to improve the performance of GNN models in crash hotspot mapping applications.

4.2.2 Effects of variability in crash frequency across days
As the temporal sparsity of crashes may also influence the prediction performance of the GLDNet, we examine the model performance according to the number of crash occurrences by time interval. For that, we split the test sample into two groups. The first group (324 intervals) comprises intervals with less than two crash occurrences in the entire study area, while the second group (41 intervals) is defined by intervals with two or more crash occurrences.

Table 6 shows the GLDNet mean hit rate by level of crash occurrences by time interval. For intervals with fewer than two crash occurrences, the MB-based model outperforms its grid counterpart, with higher mean hit rates and statistically significant differences for all coverage levels. On the contrary, the grid-based model presents higher mean hit rates for intervals with two or more crash occurrences at higher coverage levels, but the results are not statistically different. For lower coverage levels (5% and 10%), the MB representation presents the best performance, with higher mean hit rates and statistically significantly different results at a

Table 6 Mean hit for the GLDNet model by daily crash count level for two geographic units

Coverage level	0 ≤ crash count < 2				2 ≤ crash count < 6			
	MB	Grid	Difference[a]	p-value	MB	Grid	Difference[a]	*p*-value
5%	38.9%	27.4%	11.6%	0.0315	31.1%	23.3%	7.8%	0.0797
10%	57.9%	42.1%	15.8%	0.0068	41.1%	36.8%	4.2%	0.2557
15%	65.3%	54.7%	10.5%	0.0385	46.5%	49.8%	-3.3%	0.5883
20%	71.6%	60.0%	11.6%	0.0205	55.9%	66.5%	-10.7%	0.8772
25%	74.7%	66.3%	8.4%	0.0721	58.4%	67.3%	-8.9%	0.8774
30%	75.8%	67.4%	8.4%	0.0653	60.8%	67.3%	-6.5%	0.8028

[a] Positive values indicate that MB has a higher mean hit rate, while negative values indicate the opposite

5% coverage level. These results corroborate the discussion in the previous section that points to the particular advantage of using the MB and potentially other non-gridded space representations to predict sparser events (both in space and time).

4.3 Sensitivity analysis

As described in Table 2, a sensitivity analysis of the GLDNet was conducted to investigate the effects of parameter settings on the model's performance and define the optimal model configuration. In this section, we discuss the effects of parameter settings at a 5% and 15% coverage level. Additionally, we evaluate the models' computational time. Figure 5 presents the changes in the mean rate at 5% and 15% coverage levels based on the two geographic units and for four parameters of interest, including the number of L GNet (i.e., temporal) and K DNet (i.e., spatial) hidden layers, the parameter ρ in the loss function and the time window M. Regarding the time window M, Figure 5(a) shows that the highest mean hit rate is observed for a time window of 120 six-hour intervals for both MB and grid units. Concerning the number of L GNet layers, Figure 5(b) shows the highest mean hit rate can be achieved with two hidden layers for both geographic units and the two coverage levels. Referring to the number of K DNet hidden layers, two layers also provide the best mean hit rate for both geographic units. Concerning the parameter ρ in the loss function, the MB-based model has the best performance with ρ values equal to 0.02,
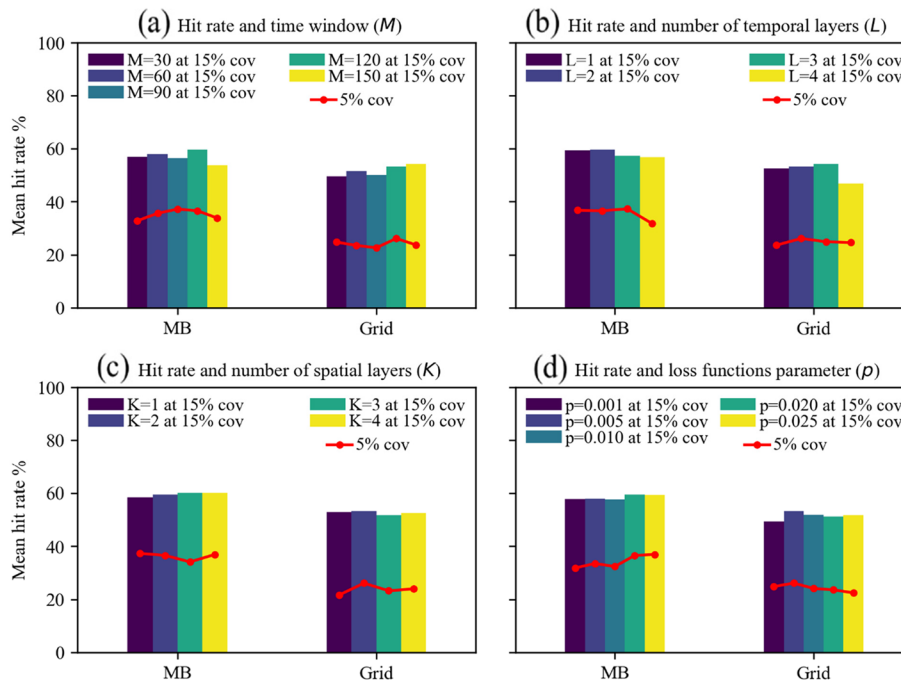

Fig. 5 Sensitivity analysis of the GLDNet based on MB and grid units

while for the grid-based model, the same occurs for ρ values equal to 0.005. It is interesting to find that the optimal value of ρ is lower for the grid case. This difference is potentially due to the variation of the spatial pattern that stems from aggregating data with two distinct geographic units. Concerning computational time, the GLDNet (763.5 minutes for MB-based and 391.8 minutes for grid-based) is at least 15% faster than the STGCN, 20% slower than the ARIMA and 100 times slower than the GBRT, while the HA is the fastest method due to its simplicity.

5 Implications to model applicability

In this section, we demonstrate that the advantages of using statistical/administrative geographic units, such as MB, go beyond the model's improved predictive performance. Hotspot maps based on these units can be easily integrated with other sources of spatial data, such as land use and road network, to help inform decision-makers on the allocation of city resources, including the development of police patrolling and rescuing strategies.

To illustrate the relevance of matching predicted hotspots with contextual information for decision-making, we show an example of how our model results could be used together with land use (Fig. 6 (a) and road network (6 (b)) data to create patrolling and rescue strategies. To do that, we analyse the locations with the highest crash risk probability for the 12h-18h interval of a random day in the test sample.

Figure 6(c) presents the predictive mapping of crash risk at a 10% coverage level. The predicted crash risk is subdivided into six localised hotspots according to their location and are described based on their land use, population and street network. Hotspot number one is located in the city centre and comprises the

CBD as well as some of its first-degree neighbours. This region is largely defined by commercial areas that form the main entertainment and business centre of Victoria that attracts thousands of trips throughout the day conducted by a range of modes of transport. The large number of hotels, services, cafes, pubs, and public and private offices, together with a dense network of motorised and non-motorised modes of transport, creates a unique environment of human mobility and crash patterns that should be accounted for when developing police patrolling and rescue strategies. For instance, the large number of active road users, trip purposes, entertainment venues and visitors should be considered when planning and monitoring pedestrian crossing behaviour and managing crowd behaviour. In contrast, hotspot number two is mostly defined by large parks located between two southeastern suburbs. The parks contain main corridors connecting the south and southeast regions to the city centre. Although this hotspot encompasses large areas, its low network density (Fig. 6b) indicates that most crash occurrences are concentred around specific links. Moreover, the presence of main corridors and a low-density network indicates that crash occurrences in this region are likely to involve motorised vehicles. In fact, over the study period, 53.7% of the crash occurrences involved only motorised vehicles in these areas, in contrast to 39.5% for the hotspot one. Hotspot number three is mostly defined by parks with main corridors connecting the northern regions to the city centre in a similar fashion to hotspot number two. In this sense, strategies would likely be similar for both of these regions, despite their physical distance.

Hotspots four and five are mostly commercial or industrial areas defined by large MB with main corridors connecting the west and north-western regions to
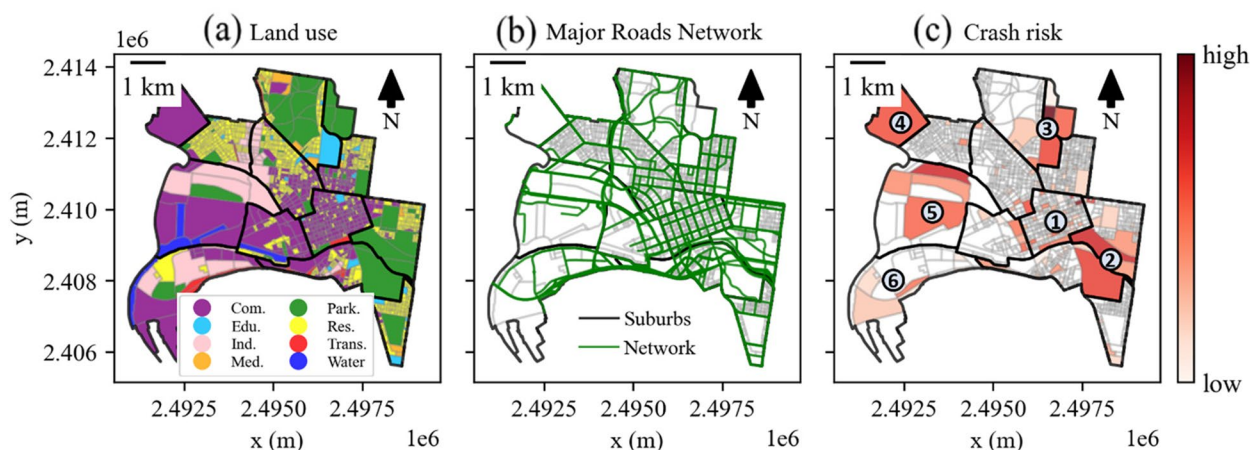

Fig. 6 GLDNet predictive hotspot mapping of crash risk for a random day at a 10% coverage level for MB and contextual information, including land use and major street network

the city centre, including the principal routes from the city centre to Victoria's main airports. Although similar, hotspot number four is mainly defined by a few arterial roads, while hotspot five consists of a main freeway and several arterial roads. Furthermore, hotspot four has a higher population density and a higher number of leisure activity centres, including cafes and pubs. As a result, different patrolling strategies would be necessary for mitigating risky driver behaviours. Lastly, hotspot number six includes access to local beaches and the port of Melbourne and is defined by large and small MB with mixed land used. In terms of network, a major freeway is also present in this region, however, with a smaller length compared to hotspot number four. Therefore, this region would likely benefit from a mixed patrolling strategy targeting the main roads and the interaction between motorised and non-motorised users in local and arterial streets, particularly on the access to the local beaches and ship terminals.

Overall, the described hotspots vary in different dimensions. Including this information with the probability of crash risk throughout an integrated framework can greatly benefit the decision-making of government agencies on the allocation of city resources. As discussed earlier, this is not a straightforward process when considering a gridded representation of space, as it requires further translation of data.

6 Conclusions

This study implemented a GNN model, the GLDNet, to predict citywide short-term crash risk. In contrast to previous GNN applications for crash risk prediction, the implemented model considers the historical occurrence of crash events as the only data input and is not constrained by a gridded representation of space, which greatly increases its applicability by practitioners in traffic management and police enforcement agencies. Overall, the results of this study have important implications and recommendations to researchers and practitioners seeking to extend the application of GNN frameworks to predict crash risk in short time horizons and at a macroscopic level:

1. Even if the only data source available to practitioners is historic crash occurrences, adopting a GNN modelling framework that considers the sparsity of crash events, such as the one used in this study, can bring forecasting accuracy benefits compared to traditional methods. Furthermore, the use of administrative spatial units is likely to improve such benefits, but caution should be given when using units defined based on population and dwelling counts in dense areas. In

this case, the modeller may want to create a criterion for aggregation that caps the minimum size of a unit.

2. Researchers should consider that, in parallel to the development of more complex and robust GNN crash models, moving from arbitrary grids towards a more meaningful representation of space will likely leverage the predictive performance of macroscopic short-term crash risk models (as well as GNN applications to similar problems in other fields of study).

3. If new geographic units are developed with the specific purpose of crash risk prediction (or prediction of analogous sparse events) using GNN models, not only spatial information of the study area (i.e., land use) should be considered, but also the properties of the network itself (i.e., network density and degree of centrality).

4. In addition to prediction performance, the use of a non-gridded representation of space increases the GDNet applicability by decision-makers for facilitating its integration with other sources of spatial data needed for developing targeted strategies to mitigate crash occurrences and their resulting injuries.

This study also presents some limitations that can be addressed in future research. Our analysis does not differentiate weekdays from weekends, which may influence the prediction since both day types have distinct traffic conditions. In addition, other geographic units and temporal resolutions should be investigated to understand which spatial and temporal segmentation can provide more meaningful predictions. In terms of methodological advances, adopting a pruning process is a straightforward path to reduce the GLDNet computational costs and increase its applicability to larger networks. Finally, including other data sources in the GLDNet framework by adding dedicated structures (GCN and a fusion network layer) or by modelling the spatial interaction between neighbours is a direct approach to improve the GLDNet performance, although at the cost of reducing its applicability due to data availability and increased computational costs.

Declarations

Ethics approval and consent to participate
Ethics approval was not required for this study.

Consent for publication
The authors certify that the submission is original work and is not under review at any other publication. All co-authors agree with consent for publication if this paper is accepted.

Competing interests
The authors have no competing interests to declare that are relevant to the content of this article.

References

Adepeju, M., Rosser, G., & Cheng, T. (2016). Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study. *International Journal of Geographical Information Science*, *30*, 2133–2154. https://doi.org/10.1080/13658816.2016.1159684.

Anselin, L., & Rey, S. J. (2014). *Modern spatial econometrics in practice: A guide to GeoDa*. GeoDaSpace and PySAL: GeoDa Press LLC.

Arvin, R., Khattak, A. J., & Qi, H. (2021). Safety critical event prediction through unified analysis of driver and vehicle volatilities: Application of deep learning methods. *Accident Analysis & Prevention*, *151*, 105949. https://doi.org/10.1016/j.aap.2020.105949.

Bao, J., Liu, P., & Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention*, *122*, 239–254. https://doi.org/10.1016/j.aap.2018.10.015.

Basso, F., Pezoa, R., Varas, M., & Villalobos, M. (2021). A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. *Accident Analysis & Prevention*, *162*, 106409. https://doi.org/10.1016/j.aap.2021.106409.

Bowers, K. J. (2004). Prospective Hot-Spotting: The Future of Crime Mapping? *British Journal of Criminology*, *44*, 641–658. https://doi.org/10.1093/bjc/azh036.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, *65*, 1509–1526. https://doi.org/10.1080/01621459.1970.10481180.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric Deep Learning Going beyond Euclidean data. *IEEE Signal Processing Magazine*, *34*, 18–42. https://doi.org/10.1109/Msp.2017.2693418.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., & Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. *Transportation Research Part A: Policy and Practice*, *127*, 71–85. https://doi.org/10.1016/j.tra.2019.07.010.

Chen, Q., Song, X., Yamada, H., & Shibasaki, R. (2016). *Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference*. Phoenix, Arizona: Thirtieth AAAI conference on artificial intelligence.

Conover, M. D., Gonçalves, B., Flammini, A., & Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, *1*, 6. https://doi.org/10.1140/epjds6.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). *Language modeling with gated convolutional networks*. In *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Asutralia: PMLR.

Hou, Y., & Edara, P. (2018). Network Scale Travel Time Prediction using Deep Learning. *Transportation Research Record*, *2672*, 115–123. https://doi.org/10.1177/0361198118776139.

Hu, J., Huang, M.-C., & Yu, X. (2020). Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. *Accident Analysis & Prevention*, *144*, 105665. https://doi.org/10.1016/j.aap.2020.105665.

Huang, Y., Zhang, F., & Hu, J. (2022). *Deep Spatial–Temporal Graph Modeling of Urban Traffic Accident Prediction*. In *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*. Singapore: Springer.

Ite, W., & Pande, A. (2016). *Traffic Engineering Handbook*, (7th ed.,). John Wiley & Sons.

Jahan, I., Ahmed, M. F., Ali, M. O., & Jang, Y. M. (2022). Self-gated rectified linear unit for performance improvement of deep neural networks. *ICT Express*, *9*, 320–325. https://doi.org/10.1016/j.icte.2021.12.012.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, *135*, 105371. https://doi.org/10.1016/j.aap.2019.105371.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. https://arxiv.org/abs/1707.01926.

Li, Z., Kolmanovsky, I., Atkins, E., Lu, J., Filev, D. P., & Michelini, J. (2016). Road Risk Modeling and Cloud-Aided Safety-Based Route Planning. *IEEE Transactions on Cybernetics*, *46*, 2473–2483. https://doi.org/10.1109/TCYB.2015.2478698.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, *54*, 187–197. https://doi.org/10.1016/j.trc.2015.03.014.

Ma, X. L., Dai, Z., He, Z. B., Ma, J. H., Wang, Y., & Wang, Y. P. (2017). Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors*, *17*, 818. https://doi.org/10.3390/s17040818.

Mannering, F. (2018). Temporal instability and the analysis of highway accident data. *Analytic Methods in Accident Research*, *17*, 1–13. https://doi.org/10.1016/j.amar.2017.10.002.

Mukhopadhyay, A., Pettet, G., Vazirizade, S. M., Lu, D., Jaimes, A., Said, S. E., … Dubey, A. (2022). A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management. *Accident Analysis & Prevention*, *165*, 106501. https://doi.org/10.1016/j.aap.2021.106501.

Newman, M. (2018). *Networks*, (2th ed.,). Oxford university press.

Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, *10*, 1372. https://doi.org/10.1038/s41598-020-58166-5.

Rozemberczki, B., Scherer, P., He, Y., Panagopoulos, G., Riedel, A., Astefanoaei, M., … Sarkar, R. (2021). *PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models*. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Queensland, Australia. https://doi.org/10.1145/3459637.3482014.

Shi, Q., & Abdel-Aty, M. (2015). Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, *58*, 380–394. https://doi.org/10.1016/j.trc.2015.02.022.

Sieveneck, S., & Sutter, C. (2021). Predictive policing in the context of road traffic safety: A systematic review and theoretical considerations. *Transportation Research Interdisciplinary Perspectives*, *11*, 100429. https://doi.org/10.1016/j.trip.2021.100429.

Teng, S.-H. (2016). Scalable Algorithms for Data and Network Analysis. *Foundations and Trends®. Theoretical Computer Science*, *12*, 1–274. https://doi.org/10.1561/0400000051.

Wang, B. B., Lin, Y. F., Guo, S. N., & Wan, H. Y. (2021a). GSNet: Learning Spatial-Temporal Correlations from Geographical and Semantic Aspects for Traffic Accident Risk Forecasting. *AAAI Conference on Artificial Intelligence*, *35*. https://doi.org/10.1609/aaai.v35i5.16566.

Wang, S., Cao, J., & Yu, P. (2020). Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 3681–3700. https://doi.org/10.1109/tkde.2020.3025580.

Wang, S., Zhang, J., Li, J., Miao, H., & Cao, J. (2021b). Traffic Accident Risk Prediction via Multi-View Multi-Task Spatio-Temporal Networks. *IEEE Transactions on Knowledge and Data Engineering*, *1-1*. https://doi.org/10.1109/tkde.2021.3135621.

Wu, L., Cui, P., Pei, J., & Zhao, L. (2022). *Graph Neural Networks: Foundations, Frontiers, and Applications*, (1th ed.,). Springer.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*, 4–21. https://doi.org/10.1109/TNNLS.2020.2978386.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting https://arxiv.org/abs/1709.04875

Zhang, Y. (2020). *Graph Deep Learning Models for Network based Spatio-Temporal Data Forecasting: From Dense to Sparse*. University College London.

Zhang, Y., & Cheng, T. (2020). Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. *Computers, Environment and Urban Systems*, *79*, 101403. https://doi.org/10.1016/j.compenvurbsys.2019.101403.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019.

Zhou, Z., Wang, Y., Xie, X., Chen, L., & Liu, H. (2020a). RiskOracle: a minute-level citywide traffic accident forecasting framework. *Proceedings of the AAAI conference on artificial intelligence*, *34*, 1258–1265. https://doi.org/10.1609/aaai.v34i01.5480.

Zhou, Z., Wang, Y., Xie, X., Chen, L., & Zhu, C. (2020b). Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 3786–3799. https://doi.org/10.1109/TKDE.2020.3034312.

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention*, *135*, 105323. https://doi.org/10.1016/j.aap.2019.105323.

Publisher's Note