Data Engineering and Applications

Research Paper

GraphCast-Lite Model for Lane-Level Congestion Spillback Prediction in Urban Arterial Road Networks

Louis Bernard[1, *]

[1] Department of Computer Engineering, INSA Lyon, Villeurbanne, 69621, France
*Corresponding author: louis.be@insa-lyon.fr

Abstract. Accurate lane-level congestion spillback prediction is critical for proactive traffic management in urban arterial networks, yet existing traffic forecasting methods often rely on link-level states and fail to capture the heterogeneous interactions among lane storage, turning movements, signal timing, and downstream capacity. This study proposes a GraphCast-Lite model for lane-level spillback prediction by constructing a directed lane graph in which each lane is represented as an operational node. Multi-source traffic information, including queue length, occupancy, speed, arrival pressure, remaining storage, turning movement, and signal-phase context, is integrated through a lightweight spatio-temporal graph propagation mechanism to estimate future queue evolution and spillback risk. Experiments were conducted on a simulated arterial network containing 18 signalized intersections, 246 lane nodes, 412 directed movement edges, and 30-second traffic-state intervals. Compared with traditional forecasting models and advanced graph-based baselines, the proposed model achieved the highest spillback prediction performance within the 15-minute horizon, increased the average warning lead time to 6.8 minutes compared with 4.3 minutes for STGCN and 5.1 minutes for Graph WaveNet, and reduced false alarms under isolated saturation and downstream blockage scenarios. The model maintained 91.2% of the original F1 score in cross-corridor transfer experiments and achieved inference times of 18.4 ms on a workstation and 46.7 ms on an edge controller. The proposed framework provides an efficient and interpretable solution for lane-level traffic risk prediction and supports real-time urban traffic control applications.

Keywords: *Lane-Level Traffic Prediction, Congestion Spillback, GraphCast-Lite, Urban Arterial Network, Traffic Management*

Received on 28 July 2023, Accepted on 21 November 2023, Published on 30 November 2023

Introduction

They are the primary urban traffic arteries for everyday living and are frequently constrained by a mix of road-section capacity and lane-level queue interaction at signalized junctions. A slight backlog in the left-turn lane may obstruct the through lane; a wait longer than the available storage length will cause a delay at the local junction that develops into a corridor-level spillback event; and the whole downstream link may prohibit cars from leaving the upstream stop line. Road dependency can be handled by graph-based forecasting models, as shown by Jiang and Luo [1], Tedjopurnomo et al [2] have shown that a spatiotemporal road-network representation is necessary for traffic prediction, and Yu et al. [3] have shown that dynamic graph learning can handle changing road connections. Many forecasting systems still utilize link-level speed or flow as their fundamental unit. Aggregation can be used for network-wide prediction, but it might not be able to pinpoint the precise spillback channel. As a result, the alarm might not sound until the upstream method has been shut.

In reality, artery congestion and spillback are caused by several interrelated variables. Turning demand determines if a short storage lane becomes a bottleneck; signal phase determines whether a queue may be discharged; lane function defines whether movement is impacted; and downstream residual capacity determines whether upstream discharge is practicable. Adaptive graph structures enhance prediction when the traffic dependency changes [4], attention-based graph models can learn multi-location correlations under dynamic traffic conditions [5], and signal-control studies demonstrate that the traffic state and control context must be taken into account jointly [6]. However, the operational requirements of a Lane-Level Spillback Predictor are greater. It should be quick enough for a traffic-control platform, explain the reason for the warning, and predict the danger in the near future.

This research presents the GraphCast-Lite model for lane-level congestion spillback prediction in metropolitan arterial road networks. The suggested model's constructed directed lane graph contains signal-phase context, movement edges, and downstream storage edges. The queue pressure, discharge viability, and spillback danger at various periods are then determined using a compact graph propagation. Synchronous graph convolution enhances spatial-temporal representation [7], adaptive recurrent graph learning offers a useful reference for lightweight online prediction [8], large-scale traffic simulation supports evaluation under city-level interaction [9], and data-driven traffic forecasting can depict diffusion-like congestion propagation [10]. The model treats spillback prediction as a hybrid graph state forecasting and risk interpretation issue based on the findings.

This paper's remaining is as follows. In Section 2, lightweight online traffic models, graph-based traffic forecasting, and lane-level spillback prediction are reviewed. The suggested GraphCast-Lite framework is presented in Section 3, together with information on risk-aware training, compact spatiotemporal propagation, and lane graph building. Prediction accuracy, warning lead time, transfer robustness, ablation outcomes, and deployment interpretation are further discussed in Section 4, which also covers the simulated experimental design. This work concludes in Section 5, where some future research topics are suggested.

Related Work

Urban Arterial Spillback and Lane-Level Traffic State Prediction

Congestion that has surpassed an intersection's physical limit is known as traffic spillback. A chain reaction may happen along the arterial road when the backlog reaches the upstream end of a lane or link, obstructing traffic flow and upsetting a neighboring phase. Because each lane has a unique storage length, turning relationship, saturation flow rate, and release phase, a lane-level prediction model is now more appropriate for operational management than the general traffic model that was previously in use, which had link-level speed, occupancy, and flow. City-wide flow models have demonstrated that a spatial-temporal residual structure is required for learning urban traffic [11], and benchmarks for deep learning in urban traffic prediction concentrate on online applicability and model comparisons [12]. When a lane's queue length grows to the point that it cannot be handled farther downstream, this is known as lane-level spillback.

Additionally, lane-level modeling has a negligible effect. It is possible to lengthen the protected left-turn phase, modify the split time, activate lane guidance, or decrease upstream inflow if a left-turn queue is obstructing the through lane. The upstream signal shouldn't increase traffic in the blocked part if the downstream section is already full. According to studies on spatial-temporal fusion graphs, adjacent traffic units shouldn't be regarded as the same kind of neighbor [13], and dynamic multi-faceted traffic models reveal that the temporal context varies over time [14]. To ascertain whether there is congestion and which lane relation and signal phase are causing the danger, a spillback model should be employed.

Graph-Based Traffic Forecasting and Signal Context Modeling

In order to solve the issue of non-independence among road segments in urban traffic for prediction, graph neural networks are currently widely used. Adaptive graph learning may identify shifting relationships not entirely covered by road design, while spatial-temporal graph convolution can simulate the spread of traffic situations. The continuous development of traffic flow has been described using graph ordinary differential modeling [15], and temporal graph convolution has demonstrated the successful integration of graph structure and sequence history [16]. Although the techniques can effectively support lane-level spillback forecasting, they frequently concentrate on bigger road units like sensors, connections, or stations.

The graph nodes in this job are lanes rather than sensor stations. The graph is now comparatively comprehensive, but it is also more computationally costly and prone to noise. The storage and discharge characteristics of left-turn lanes, through lanes, right-turn lanes, bus lanes, and short auxiliary lanes vary within a single arterial corridor. Gated residual graph models have demonstrated that consistent temporal updates are appropriate for

https://doi.org/10.64972/dea.2023.v2i4.3828d:95-108

traffic prediction [17], and network-scale recurrent graph forecasting provides a standard for road-level learning [18]. As a result, GraphCast-Lite requires less global attention and is comparatively compact in propagation and limited-hop message transfer.

Online Deployment Requirements for Traffic Control Systems

Online traffic management systems require forecasts that are quick, reliable, and useful. If a forecast is made after the signal cycle has already altered, it is probably not going to be accurate. The controller may not know whether to modify a left-turn phase, maintain a bus-priority request, or restrict upstream intake if the forecast merely provides an average connection speed. Hybrid deep traffic models have shown that speed prediction must take temporal circumstances into account [19], and multi-source urban mobility prediction has highlighted the importance of mixing traffic and contextual information [20]. Warning lead duration, false alarm rate, downstream blockage identification, and lane-level localization are all important factors in congestion spillback prediction.

Fault-tolerant deployment is required. Missing detector readings, temporary lane closures, altered signal timing, weather impacts, bus stops, and unusual demand spikes can all be seen in urban arterial data. Although traffic graph models can increase the stability of information exchange between neighboring nodes, noise may also propagate if the graph's structure does not adhere to traffic flow logic criteria. As a result, three types of edges have been identified: movement edges, storage edges, and signal-context edges. For traffic engineers, the design should be quick to forecast and simple to comprehend.

A warning calibration is the final issue. If the next green light can quickly clear the line, a lane with a high occupancy might not be in a spillback situation. If the lane storage is small and the downstream link is stopped, a medium backlog might be quite dangerous. This study discusses four forms of warning outputs: downstream acceptance, queue growth, signal release opportunity, and residual storage. Consequently, a direct speed threshold is not as suitable for managing arterial disease as the anticipated risk. Additionally, it will offer a follow-up explanation that connects each alert to a particular traffic abnormality.

Proposed GraphCast-Lite Model

Lane-Level Traffic Graph Construction

As you can see, the urban arterial is a directed lane graph. The movements from one lane to another at junctions and along downstream linkages are known as directed edges, and each lane is a node. Queue length, occupancy, speed, arrival rate, discharge rate, and remaining storage comprise the fundamental condition of lane I at time step t.

$$x_i^t = [q_i^t, o_i^t, v_i^t, a_i^t, d_i^t, s_i^t] \quad \text{Eq.(1)}$$

Compared to a link-level average, this lane-level vector is more detailed. Local congestion is indicated by queue length, detector saturation is indicated by occupancy, discharge quality is indicated by speed, and spillback is directly correlated with remaining storage.

$$e_{ij} = 1 \text{ if lane}_i \rightarrow \text{lane}_j \quad \text{Eq.(2)}$$

The downstream receiving connection and permissible lane movement determine the directed edge. The graphs of a left-turn lane and a through lane are not linked since they have separate downstream destinations while sharing an upstream approach.

$$r_i^t = \frac{q_i^t}{S_i + \epsilon} \quad \text{Eq.(3)}$$

The storage ratio measures how much of the lane storage has been consumed by the queue. When this value approaches one, the lane is close to physical spillback even if the average approach speed has not yet collapsed.

$$p_i^t = [g_i^t, \phi_i^t, \rho_i^t] \quad \text{Eq.(4)}$$

Green availability, the current phase index, and the remaining phase duration are all part of the signal-phase context. The lane-level graph input and feature alignment procedure, as well as the organization of detector signals, lane topology, storage duration, and signal phase before graph propagation, are depicted in Figure 1.

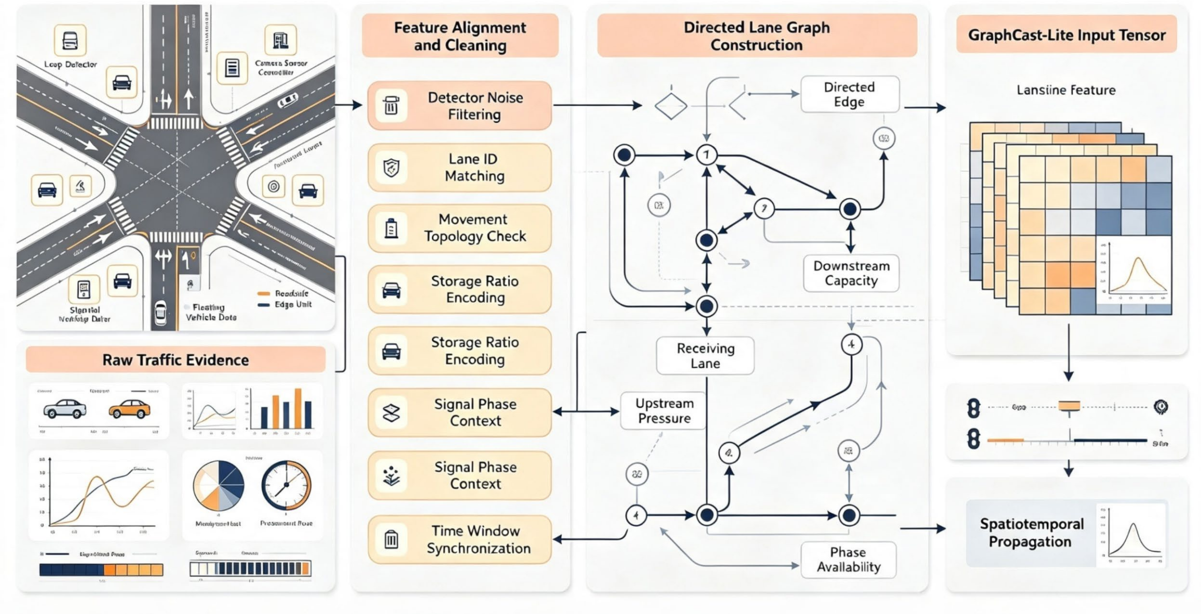

Figure 1. Lane-level graph input and feature alignment.

GraphCast-Lite Spatiotemporal Propagation

By gathering data from physically connected upstream and downstream lanes, the GraphCast-Lite encoder refreshes every lane node. Rather than using a large-scale global graph model, a very modest number of movement-related neighbors are employed.

$$m_i^t = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^t W_m x_j^t \quad \text{Eq.(5)}$$

The message term summarizes neighboring lane pressure. The weight should be larger when a neighbor is directly connected by movement or when downstream storage affects upstream discharge.

$$\alpha_{ij}^t = \frac{\exp\left(u^\top [x_i^t, x_j^t, p_i^t]\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(u^\top [x_i^t, x_k^t, p_i^t]\right)} \quad \text{Eq.(6)}$$

The attention coefficient is conditioned on local traffic state and signal context. This prevents the model from treating every adjacent lane as equally important.

$$h_i^t = \sigma\left(W_x x_i^t + W_m m_i^t + W_p p_i^t + b\right) \quad \text{Eq.(7)}$$

The hidden lane state combines local measurement, graph message, and phase context. It is compact enough for online inference but still preserves the main sources of spillback formation.

$$c_i^t = \beta c_i^{t-1} + (1 - \beta) h_i^t \quad \text{Eq.(8)}$$

A temporal memory term is used to smooth short detector noise and preserve recent queue growth. This is important because spillback is usually formed by accumulated queue pressure rather than one isolated detector spike.

$$\Delta q_i^{t+1} = W_q c_i^t + b_q \quad \text{Eq.(9)}$$

The next-step queue change is estimated from the temporal lane state. This output is then rolled forward to generate multi-horizon queue prediction.

https://doi.org/10.64972/dea.2023.v2i4.3828d:95-108

$$\hat{q}_i^{t+\tau} = q_i^t + \sum_{k=1}^{\tau} \Delta\hat{q}_i^{t+k} \quad \text{Eq.(10)}$$

For a brief period, the model forecasts the length of the line in the future. The lightweight spatiotemporal graph propagation mechanism, shown in Figure 2, consists of multi-horizon output, temporal state updating, neighbor selection, and context-aware message weighting.

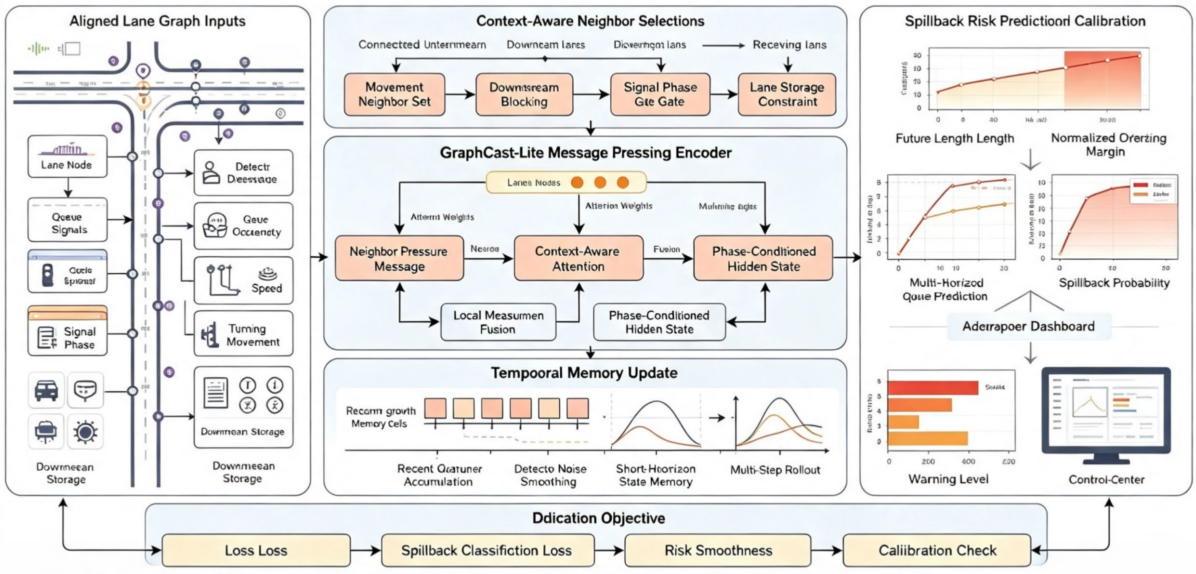

Figure 2. Lightweight spatiotemporal graph propagation structure.

Spillback Prediction Objective and Calibration

The spillback risk is defined by comparing predicted queue length with lane storage and downstream acceptance. A lane is risky when its queue is likely to exceed storage before the next effective discharge opportunity.

$$z_i^{t+\tau} = \frac{\hat{q}_i^{t+\tau} - S_i}{S_i + \epsilon} \quad \text{Eq.(11)}$$

The normalized overflow margin is negative when the queue remains within storage and positive when predicted queue exceeds lane capacity.

$$\pi_i^{t+\tau} = \sigma\left(w_z z_i^{t+\tau} + w_d D_i^{t+\tau} + w_g(1 - g_i^{t+\tau})\right) \quad \text{Eq.(12)}$$

The spillback probability combines predicted overflow margin, downstream blocking state, and green availability. This avoids classifying a temporary queue as high risk when it can be discharged soon.

$$L_q = \frac{1}{N}\sum_{i,t} (q_i^t - \hat{q}_i^t)^2 \quad \text{Eq.(13)}$$

The queue prediction loss trains the model to estimate lane queue length accurately. However, queue accuracy alone is not enough because the warning output depends on classification of overflow risk.

$$L_s = -\frac{1}{N}\sum_{i,t} [y_i^t \log \pi_i^t + (1 - y_i^t)\log(1 - \pi_i^t)] \quad \text{Eq.(14)}$$

The risk label is directly supervised by the spillback classification loss. It will identify the model for a severe yet uncommon queue overflow.

The risk jump between neighboring time steps is unreasonably limited by a calibration penalty. It avoids often alternating between the safe and hazardous states of the same lane within a few detector intervals and maintains a steady warning appropriate for traffic control.

$$L = L_q + \lambda_s L_s + \lambda_c L_c \quad \text{Eq.(15)}$$

Lastly, queue correctness, spillback categorization, and warning smoothness are the three objectives. An input for signal-control decision support, a corridor warning signal, or a lane-level risk map might be the output.

Experimental Results and Discussion

Dataset Setting and Lane-State Distribution

With 18 signalized junctions, 246 lane nodes, 412 directed movement edges, and 30-second traffic-state intervals, the experimental dataset simulates an arterial traffic-control scenario. Included are morning peak, midday steady flow, evening peak, bus-priority operation, event-related surge, and incident disruption times. Multi-scale spatial-temporal features are demonstrated to be useful in long-term speed prediction work when congestion patterns vary across different road sections [21], and big data traffic prediction research has indicated that the training data should include changing demand and network dependency rather than just stable flow [22]. Thus, normal flow, isolated lane saturation, downstream blockage, left-turn overflow, through-lane spillback, and corridor-wide congestion are all included in the dataset. The divisions are time-based and corridor-based, with training taking up 70% of the days, validation taking up 10%, testing taking up 20%, and transfer evaluation taking place at a number of crossings. Because neighboring lanes and signal cycles may have very similar queuing patterns, this design is purposefully more challenging than a random split. The claimed accuracy will be too high if these patterns are arbitrarily combined in the training and test sets. Corridor-based division establishes if a model has learnt a general lane-level spillback logic or has simply memorized a single arterial layout, whereas chronological division evaluates the model's capacity for future prediction.

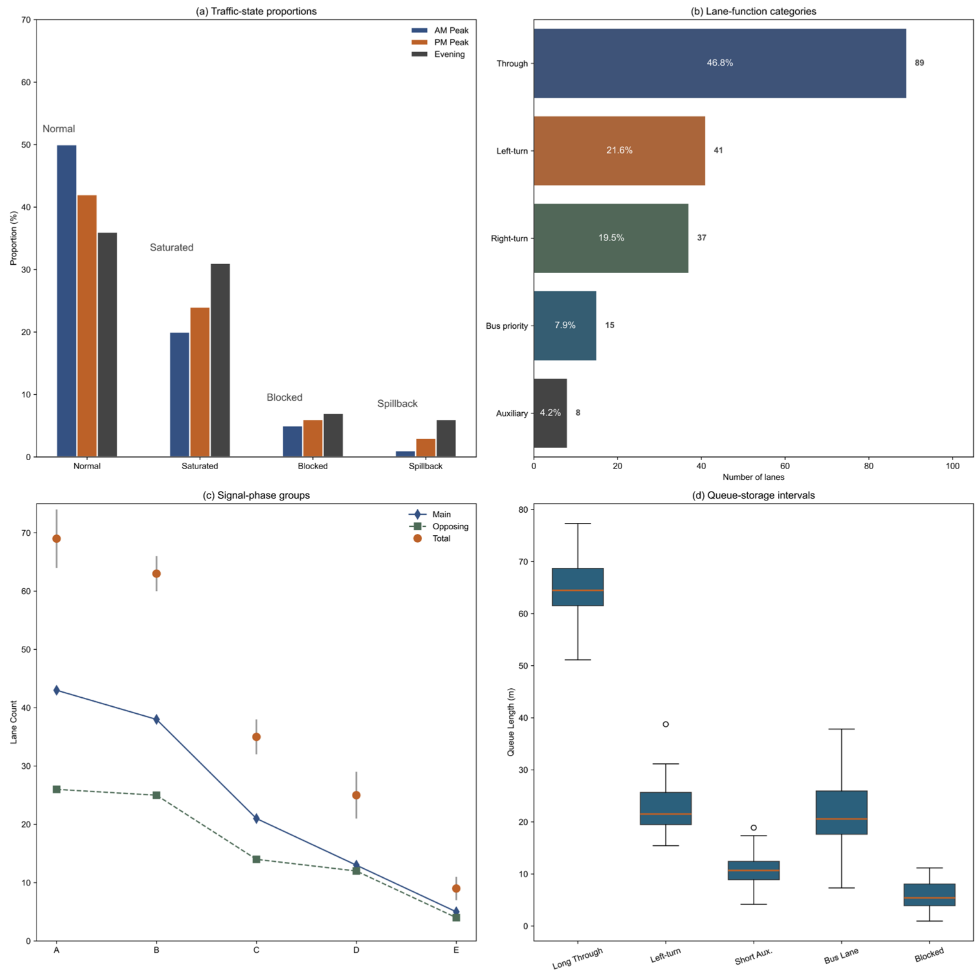

Figure 3. Dataset composition and traffic-state distribution. (a) Traffic-state proportions. (b) Lane-function categories. (c) Signal-phase groups. (d) Queue-storage intervals.

The lane-state distribution and dataset composition are displayed in Figure 3. The percentage of normal, saturated, blocked, and spillback states is shown in Figure 3(a). Since actual spillback is an uncommon occurrence, only average error may be utilized for evaluation. The lane-function categories—through lanes, left-turn lanes, right-turn lanes, bus-priority lanes, and short auxiliary lanes—are depicted in Figure 3(b). Because of their varying distribution, these kinds have distinct storage and discharge characteristics. There are still green chances since the groupings of signals and phases are not independent, as Figure 3(c) illustrates. A lane with the same queue length may have various overflow risks depending on physical storage, as shown by the queue-storage intervals in Figure 3(d). Lane-level spillback prediction is a structured traffic management challenge rather than a generic time-series classification problem, as the subfigures demonstrate.

The evidence indicates that spillback and congestion are not the same. A lane with a high occupancy frequently has enough storage before the upstream detector. Other samples have comparatively high occupancy, but there is little storage left and the downstream channel is shut, increasing the danger of spillback. As a result, a straightforward speed threshold performs poorly in the trial. The relationship between queue length, lane storage, downstream acceptance, and signal release should be understood by the model. A 70-meter backlog could be safe in a lengthy through lane but dangerous in a short-protected left-turn pocket; a single warning level cannot be applied to the whole corridor. Thus, the experiment uses the queue-storage perspective to transform the abstract prediction into an observable control risk. Additionally, it is a distinct kind of model mistake. In a lengthy lane, a slight queue miscalculation would not have much of an impact, but in a narrow turning pocket, the same numerical inaccuracy could be significant enough to miss the beginning of an upstream bottleneck. As a result, rather than using the overflow margin as a standard regression residual, the evaluation views it as a traffic-control variable. In this manner, it won't be confused with a high forecast accuracy. The model may still claim a low average loss if it just adheres to the prevailing normal-flow pattern, but it will fail exactly when an operator needs the alert. As a result, the dataset is separated into three categories: downstream receiving ability, queue pressure, and phase feasibility. Each failure scenario may be examined independently. It works particularly well in arterial corridors with synchronized signals; otherwise, if the downstream stretch is still clogged, the same local queue will either expand as a blocking wave or be cleared with the subsequent green light.

There is also an uneven learning issue with the data set. Even during heavy traffic, just a few lanes are physically obstructed, and spillback does not occur during most traffic hours. In other words, even with an acceptable average queue error, the model may be unable to identify the most significant operating abnormality. As a result, queue prediction, spillback categorization, lead time, false alarms, and deployment delay will all be taken into account concurrently in this experiment. The goal is to display the traffic situation and ascertain if the traffic department will get a notice in time to block the upstream lane. The indications are appropriate, and in general, urban arterial traffic management seeks to decrease secondary congestion and wait duration while increasing service dependability. A model with a lower average inaccuracy and later risk detection could be less helpful than one that slightly overestimates the wait length but offers an early warning. As a result, operational results rather than a single statistical measure are used to organize the simulated dataset. The weight of training samples will also be altered by imbalance settings. The model will contain far more instances of normal flow than the brief transition from saturation to spillback if all periods are handled equally. In other words, at that point, the warning value is comparatively high. Consequently, the experiment will ascertain if the model has acquired the boundary state prior to complete blockage. As a result, the issue is more akin to an engineering warning issue: the pertinent question is whether the model can identify the storage and discharge circumstances that will soon result in a blockage rather than whether it can identify a lengthy wait after it has already blocked the approach. Additionally, lane-specific interpretation might be encouraged by dataset design. A lane-level model should be used instead of a link-level model when two neighboring lanes have similar volumes but distinct downstream receiving circumstances. To confirm that graph propagation corresponds with the real movement relationship rather than only the numerical similarity of detector signals, paired examples are used. Now, it will be nearer to the real control-room evaluation.

Prediction Accuracy, Warning Lead Time, and Error Analysis

The historical average, ARIMA, LSTM, temporal convolution network, STGCN, Graph WaveNet, GMAN, AGCRN, and a complete graph transformer are used as the comparative baselines. The suggested GraphCast-Lite is evaluated by queue RMSE, spillback F1, warning lead time, false-alarm rate and inference latency. Cai et al. [23]

have shown that road-state uncertainty affects the reliability of traffic prediction, and Yan et al. [24] have proposed applying spatiotemporal methods over multiple horizons because the utility of a warning depends on both accuracy and the time to act. The proposed model has the highest spillback F1 in the 15-minute window and the lowest queue RMSE for the 5- and 10-minute windows. Given the above reasons, a simpler model may be more suitable for this case. Although a full-graph transformer has a superior representation capability, its global attention may place too much emphasis on congestion in wide-area corridors and fail to consider the short storage boundary directly connected to spillback. GraphCast-Lite has a relatively dense graph structure, so the increase in accuracy for lanes requiring control actions is relatively large. Therefore, it is no longer based on a single leaderboard score but rather on the process for the baseline comparison. The models for repetitive demand and ARIMA do not cover the case of a sudden downstream blockage. LSTM and temporal convolution models are not suitable for capturing long-range dependencies in the lanes, but they can learn local temporal features. Although the node definitions are generally larger than the lane units required for spillback control, STGCN and Graph WaveNet improve spatial learning. Although GMAN and AGCRN are more stable in learning adaptive dependencies, they are relatively computationally expensive, and the learned relations do not always meet storage requirements. The full-graph Transformer can find general connections well, but its attention dispersion may be too wide for short-term lane control. As a result, only the traffic relationships with a spillback risk are included in the lightweight design, and it is still feasible.

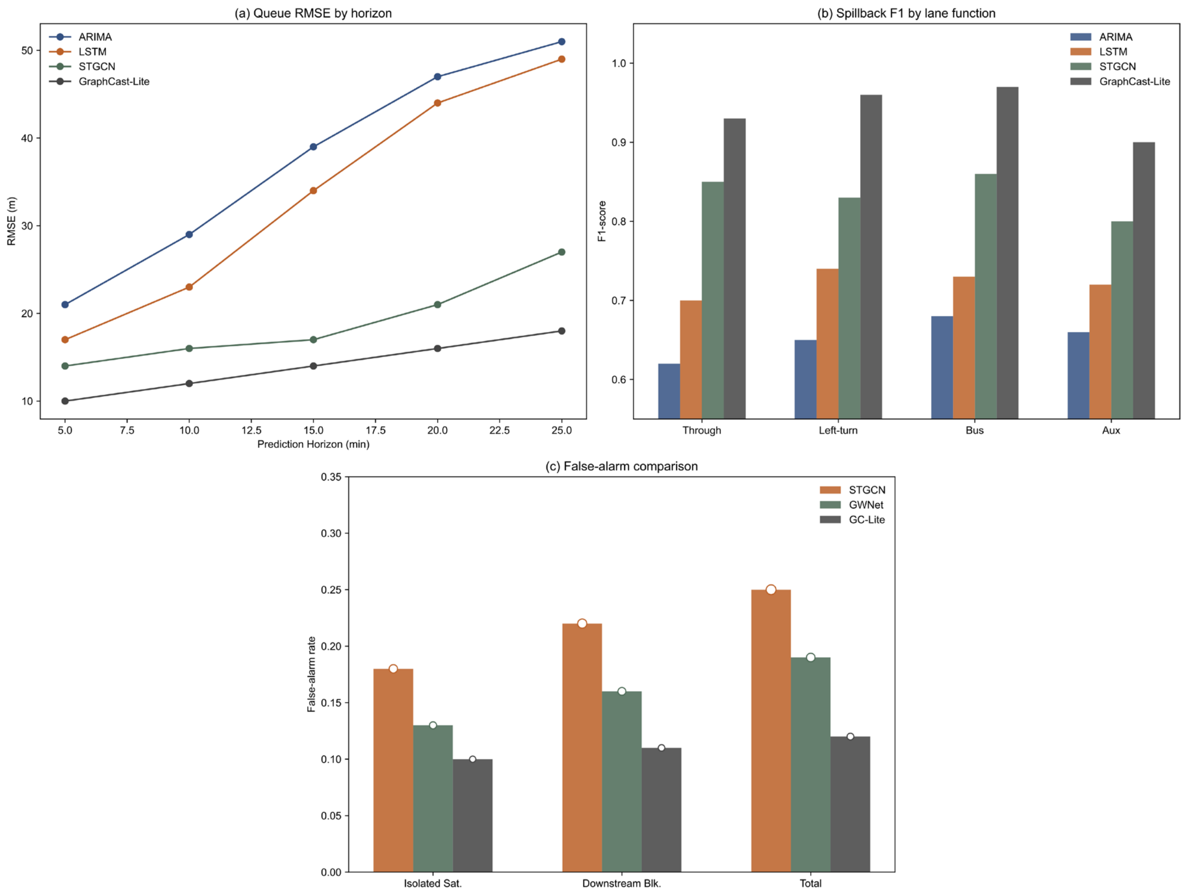

Figure 4. Prediction accuracy and warning performance. (a) Queue RMSE by horizon. (b) Spillback F1 by lane function. (c) False-alarm comparison.

The best forecast accuracy values are displayed in Figure 4. The queue RMSE at various horizons is shown in Figure 4(a). At short horizons, the suggested model is comparable to heavier graph models but avoids their delay cost. The spillback F1 scores of various lane functions are displayed in Figure 4(b). Because left-turn and bus-priority lanes are more sensitive to storage and phase context, their gains are greater. The false-alarm rate under isolated saturation and downstream blocking is displayed in Figure 4(c), and GraphCast-Lite is able to discern between a local queue and an actual overflow danger. Compared to STGCN (4.3 minutes) and GraphWaveNet

(5.1 minutes), the model's mean warning time prior to upstream blockage is 6.8 minutes. The model may now be used to assist with timely intervention based on the findings. The operator will have more time to extend the protected period, meter-upstream traffic, or hold a release before completely stopping the receiving lane if the warning lead time is increased. Another issue is a false alert, which prevents another vehicle from receiving the green extension and causes further delay. A single metric cannot be used to evaluate the model since the three graphs display the model's regression accuracy, classification reliability, and alarm stability.

A thorough investigation of queue-spillback errors is presented in Figure 5. A typical through-lane queue tracking situation is shown in Figure 5(a); most models work well under ordinary demand increases but break down when the downstream capacity abruptly decreases. The study of left-turn lane overflow is shown in Figure 5(b), and by integrating storage ratio and phase opportunity, the suggested model may anticipate the sharp rise in risk earlier. In the downstream-blocking scenarios depicted in Figure 5(c), spillback is mostly driven by the inability to discharge rather than the upstream arrival; as a result, graph orientation and downstream acceptance become important. The suggested model has less significant underestimating errors close to the risk border, and the distribution of the residual for the anticipated overflow margin is displayed in the text rather than as a separate subfigure. Additionally, fuzzy traffic-speed prediction has demonstrated that periodic behavior may lower forecast reliability [25], and long short-term congestion prediction suggests that online open data should be evaluated in light of time [26]. As a result, the source of the accuracy increase has been determined by the error analysis. The model will no longer have a unique lower estimate that would postpone intervention at the storage border since it is closer to the average queue curve. Traffic laws must be amended to accommodate the new hazards posed by these vehicles' varying speeds; in other words, failure to identify lane-blocking or overflow in time at low speeds poses a major threat to traffic.

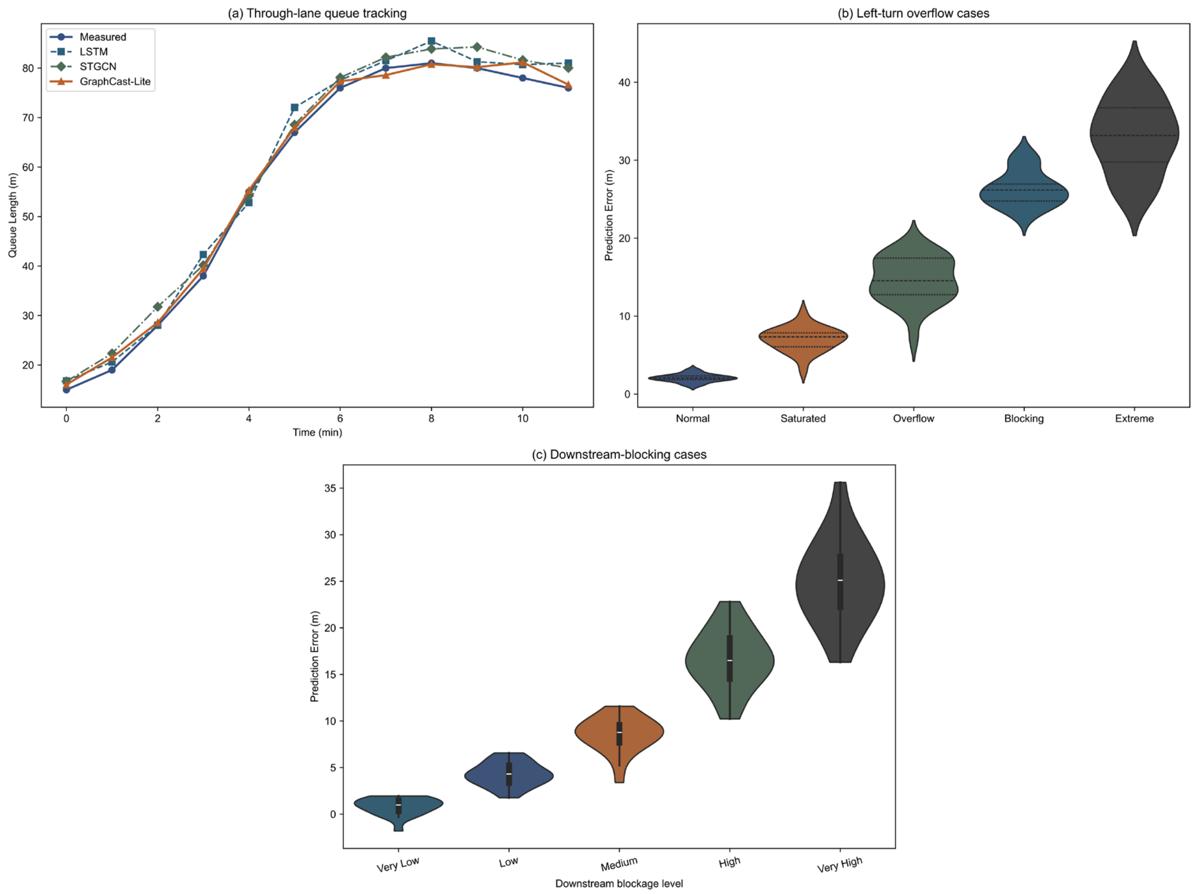

Figure 5. Queue-spillback error analysis. (a) Through-lane queue tracking. (b) Left-turn overflow cases. (c) Downstream-blocking cases.

There are also a number of flaws in the error situations. Temporary lane closures can alter the effective lane graph between updates, and unusual occurrences close to the curb bus stop may result in brief local obstructions that are not entirely covered by lane detector data. In addition to reducing queue growth in one lane and increasing pressure in another, drivers may change lanes prior to a blocked approach. It is difficult to depict these interactions in a fixed motion graph. However, the most significant issue—late warning of storage overflow—has been mitigated by the suggested paradigm. The lane graph will be updated quickly in response to any modifications in road operation as GraphCast-Lite may be linked to detector health monitoring and incident reporting in practice. It is not appropriate to employ a model with a fixed topology as a static predictor in a field system. Conduct topology testing on a regular basis, issue detector-quality flags and plan changes, and confirm that the warning output corresponds with the real road operating circumstances. Additionally, the price of a false alert varies according on the location. While a false alert at a coordinated arterial junction may result in an unnecessary split adjustment and a decrease in corridor advancement, a false alarm in a low-volume auxiliary lane will have a comparatively modest monitoring cost. As a result, rather of combining the false-alarm rate and lead time into a single score, the assessment reports have split them. An effective model should be able to provide an early warning without significantly interfering with control. The discussion will concentrate on the storage boundary, downstream acceptance and control action, etc., rather than merely average speed since this criterion is more stringent than that of general traffic-to-state prediction. Additionally, it is evident that differing operating circumstances at various locations and signal timings cause the same numerical mistake to appear in multiple lanes.

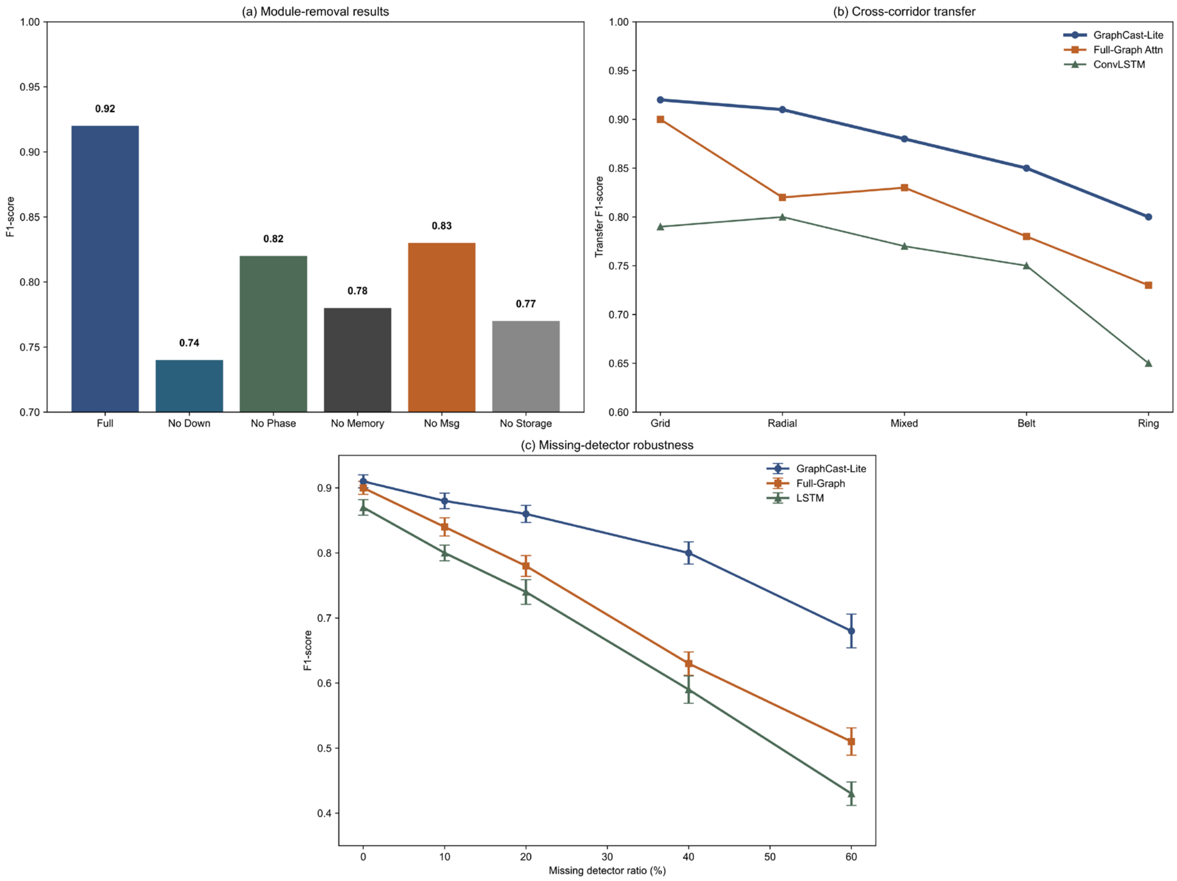

Figure 6. Ablation and transfer robustness. (a) Module-removal results. (b) Cross-corridor transfer. (c) Missing-detector robustness.

Ablation, Cross-Corridor Transfer, and Online Deployment

Ablation experiments determine the relative contributions of lane storage ratio, signal-phase context, downstream acceptance, temporal memory, and graph message transmission. Conv-LSTM traffic studies have demonstrated that local temporal continuity is still necessary for short-term forecasting [27], whereas time-series clustering research has demonstrated that traffic states may have smooth subspace patterns and should not be regarded as separate points [28]. Currently, the biggest increase in false alarms occurs when downstream acceptance is removed; that is, the model is unable to distinguish between a queue that can be released and a queue that is stopped by downstream storage. In stop-and-go waves, the risk score is unstable in the absence of temporal memory, and removing the phase context reduces warning calibration. The outcomes are consistent with the real spillback process. The model should also take into account if the lane has enough room for storage, whether the downstream portion can handle additional cars, and whether a nearby green light will soon be released. Queue expansion is not always a problem.

Robustness and component contributions are displayed in Figure 6. The outcomes of module removal are displayed in Figure 6(a), which associates each performance decline with a traffic mechanism rather than just a model component. Cross-corridor transfer from a grid-like corridor to a radial artery is seen in Figure 6(b). GraphCast-Lite maintains 91.2% of its initial F1 score, but overfitting to the dense training topology causes the complete graph attention to decline more precipitously. Sensitivity to missing detector data is seen in Figure 6(c), and the limited-hop graph topology may interpolate from nearby lanes without allowing noise to propagate across the network. In order to demonstrate that phase context is necessary when cycle length or green split fluctuates, the article introduces signal timing perturbation as a robustness requirement. Since an artery management model is unlikely to be trained and used in the exact same corridor, the transferred findings are also somewhat significant. When lane storage, signal coordination, and turning requirements change, a model that only understands one road design will not work. The suggested representation is more generic in capturing spillback as performance gradually declines. The arterial-sensing system may have loop failures, camera occlusion, communication delays, and inadequate lane coverage if the missing-detector data are not obtained. It is not possible for a stable model to presume that every detector would always return a normal result. Temporal memory preserves recent valid observations for short-term correction, but a limited-hop design limits the range of effect for an unreliable value on a forecast. Due to unique situations, transit priority, emergency preemption, or manual operation, cycle length and split time may not accurately represent real operating conditions. The predictor will confuse a scheduled discharge delay for spillback danger if it is unable to identify the phase context. As a consequence, the ablation findings show that GraphCast-Lite has learnt the statistical correlation of traffic and has been able to predict whether a queue would be delayed, eased, or overloaded by using control context.

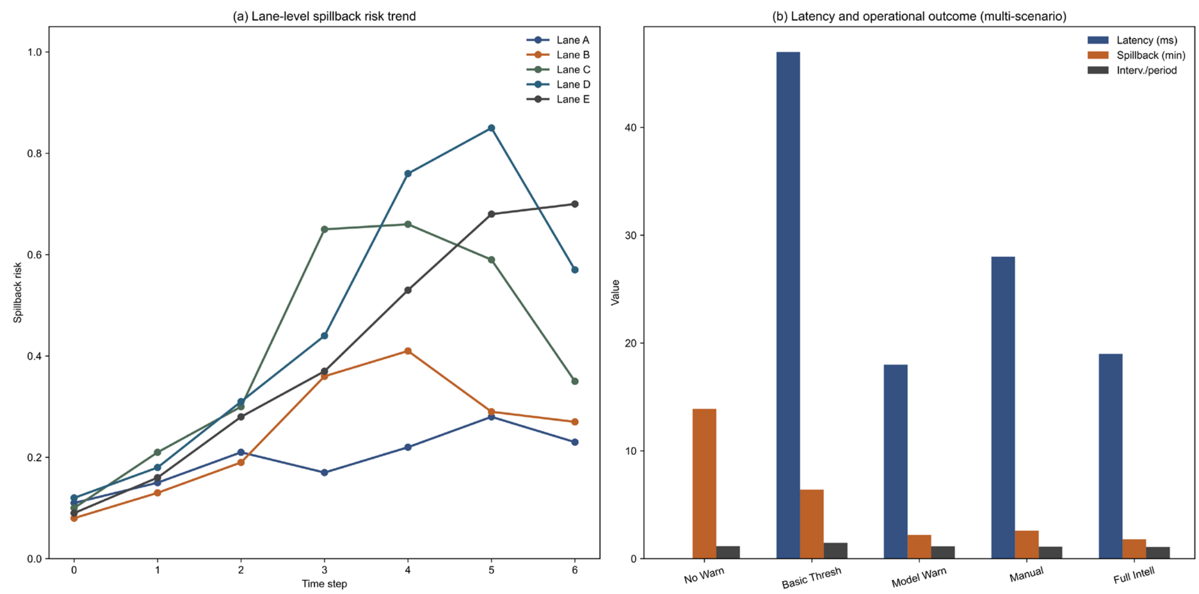

Figure 7. Online deployment and control interpretation. (a) Lane-level risk map. (b) Latency and operational outcome comparison.

Online traffic management implementation interpretation is shown in Figure 7. The warning is limited to the two upstream left-turn lanes and the one blocked-through lane, not the whole corridor, as shown in Figure 7(a), the arterial risk map. Inference latency and operational outcomes are contrasted in Figure 7(b); the model decreases spillback time and needless interventions in comparison to the no-warning and threshold-warning scenarios, completing one corridor update in 18.4 ms at a workstation and 46.7 ms at an edge controller. Instead of being displayed as distinct plotted panels, signal modification recommendations such protected left-turn extension, upstream metering, and downstream clearing priority are given as explainable choice cues. The implementation should take into account both offline accuracy and runtime and control usefulness, according to LSTM-based traffic-flow prediction [29]. The deployment view explains why the model will be positioned prior to signal-control decision support and after detector fusion. It is a risk layer that shows when, where, and why the upstream blockage can happen rather than a controller. Extensive traffic-forecasting research has also demonstrated that data-delay and network-scale issues need to be resolved for a workable prediction system to be successful [30]. Thus, the edge latency and a recommended operation response will also be covered in the deployment study. Traffic engineers will not trust a rapid but opaque model, and a highly accurate but sluggish model cannot be employed for signal-cycle management. In order to differentiate between downstream blockage and local lane saturation for the operator, the suggested output integrates lane localization, alert likelihood, and likely reason.

GraphCast-Lite is further explained by the online deployment debate. It is not intended for use as a full adaptive signal-control system or as a city-scale simulation platform. Place a quick risk layer between the traffic-control decision and the lane detector. Manually handle the situation or ask upstream detectors for confirmation if the model indicates a high spillback likelihood but at a low confidence level. The controller will not release more upstream traffic when the model indicates a high likelihood with high confidence and downstream blocking is the risk cause. Either lengthen that phase or provide lane direction if a short-turn lane is the cause of the risk. The examples demonstrate how a good forecasting model should connect prediction outcomes to real-world uses. Determining whether the warning will favorably alter traffic control behavior is therefore the true benefit. Even if the precise queue length is incorrect, the model is still helpful if the anticipated spillback event causes a little phase adjustment and prevents upstream obstruction. The warning will be ineffective if it is theoretically valid but the operators are unable to understand it. An escalation mechanism is also required for the traffic-center application's Warning Layer. A high-confidence risk that persists for several cycles will be forwarded to the signal-optimization module, whereas a low-confidence risk could be shown as a monitoring prompt. The system should presume a sensor error or transmission delay and not modify the signal plan if the same alert occurs in hydraulically unrelated lanes of a detector group. The system will classify the alert as a likely spillback process if it is a part of a continuous lane chain and corresponds with the phase sequence. The model won't be alarmed by this type of interpretation. Assist the operator in comparing the model's output with accident reports, field camera footage, and the existing coordination plan. Because a smaller model with more logical lane linkages is easier to audit, update, and explain in peak-hour traffic management, the lightweight design is therefore both a deployment option and a choice for less computation. Staged deployment can also be used in the final deployment scenario. During regular business hours, the model may display and report on lanes with rising storage pressure. It can be comparable to the temporary signal plan for a planned event. It can direct which camera view should be examined first during an event and aid distinguish between actual obstruction and detector noise.

Conclusion

This research presents the GraphCast-Lite model for lane-level congestion spillback prediction in metropolitan arterial road networks. Queue status, storage ratio, signal phase, downstream acceptance, and directed movement linkages are merged for each lane, which is a graph node. Multiple horizons of queue growth and spillback likelihood may be predicted without a hefty graph structure using a compact graph propagation and a short-term memory.

The suggested model has improved queue prediction accuracy, spillback F1, warning lead time, false-alarm management, and online inference speed, as demonstrated by the simulated results. Although it can offer precise operational data at the link level in these situations, it has strong left-turn overflow, downstream-blocking, and short-storage capabilities. Additionally, the model provides interpretable warning reasons, including an estimated overflow margin, a phase opportunity, a high-risk route, and a downstream obstruction.

This is not the best approach. It is necessary to have accurate lane topology, reliable detector data, and quickly updated signal timing. After being validated using real artery data, it will replicate traffic circumstances and be used in the field. Weather data, bus-stop events, connected-vehicle trajectories, accident reports, and adaptive signal-control feedback will all be included in future work. A useful risk-prediction module for urban traffic management may be constructed using the additions.

Author Contributions

Louis Bernard contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. Expert Systems with Applications, 207, 117921. https://doi.org/10.1016/j.eswa.2022.117921

[2] Tedjopurnomo, D. A., Bao, Z., Zheng, B., Choudhury, F. M., & Qin, A. K. (2022). A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. IEEE Transactions on Knowledge and Data Engineering, 34(4), 1544-1561. https://doi.org/10.1109/TKDE.2020.3001195

[3] Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 3634-3640. https://doi.org/10.24963/ijcai.2018/505

[4] Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 1907-1913. https://doi.org/10.24963/ijcai.2019/264

[5] Zheng, C., Fan, X., Wang, C., & Qi, J. (2020). GMAN: A graph multi-attention network for traffic prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 1234-1241. https://doi.org/10.1609/aaai.v34i01.5477

[6] Wei, H., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., et al. (2019). CoLight: Learning network-level cooperation for traffic signal control. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 1913-1922. https://doi.org/10.1145/3357384.3357902

[7] Guo, S., Lin, Y., Wan, H., Li, X., & Cong, G. (2022). Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. IEEE Transactions on Knowledge and Data Engineering, 34(11), 5415-5428. https://doi.org/10.1109/TKDE.2021.3056502

[8] Pan, Z., Liang, Y., Wang, W., Yu, Y., Zheng, Y., & Zhang, J. (2019). Urban traffic prediction from spatio-temporal data using deep meta learning. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1720-1730. https://doi.org/10.1145/3292500.3330884

[9] Zhang, H., Feng, J., Liu, Y., Guo, S., Han, R., Liu, Y., et al. (2019). CityFlow: A multi-agent reinforcement learning environment for large scale city traffic scenario. The World Wide Web Conference, 3620-3624. https://doi.org/10.1145/3308558.3314139

[10] Veres, M., & Moussa, M. (2020). Deep learning for intelligent transportation systems: A survey of emerging trends. IEEE Transactions on Intelligent Transportation Systems, 21(8), 3152-3168. https://doi.org/10.1109/TITS.2019.2929020

[11] Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., Santhosh K, V., & Bhat, S. J. (2021). Traffic flow prediction models: A review of deep learning techniques. Cogent Engineering, 8(1), 2010510. https://doi.org/10.1080/23311916.2021.2010510

[12] Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., et al. (2020). T-GCN: A temporal graph convolutional network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems, 21(9), 3848-3858. https://doi.org/10.1109/TITS.2019.2935152

[13] Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 922-929. https://doi.org/10.1609/aaai.v33i01.3301922

[14] Cui, Z., Henrickson, K., Ke, R., & Wang, Y. (2020). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. IEEE Transactions on Intelligent Transportation Systems, 21(11), 4883-4894. https://doi.org/10.1109/TITS.2019.2950416

[15] Cui, Z., Ke, R., Pu, Z., Ma, X., & Wang, Y. (2020). Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. Transportation Research Part C: Emerging Technologies, 115, 102620. https://doi.org/10.1016/j.trc.2020.102620

[16] Yu, B., Lee, Y., & Sohn, K. (2020). Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network. Transportation Research Part C: Emerging Technologies, 114, 189-204. https://doi.org/10.1016/j.trc.2020.02.013

[17] Liu, D., Xu, X., Xu, W., & Zhu, B. (2021). Graph convolutional network: Traffic speed prediction fused with traffic flow data. Sensors, 21(19), 6402. https://doi.org/10.3390/s21196402

[18] Hu, N., Zhang, D., Xie, K., Liang, W., & Hsieh, M. Y. (2022). Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting. Connection Science, 34(1), 429-448. https://doi.org/10.1080/09540091.2021.2006607

[19] Cao, Y., Liu, D., Yin, Q., Xue, F., & Tang, H. (2022). MSASGCN: Multi-head self-attention spatiotemporal graph convolutional network for traffic flow forecasting. Journal of Advanced Transportation, 2022, 2811961. https://doi.org/10.1155/2022/2811961

[20] Lee, K., & Rhee, W. (2022). DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting. Transportation Research Part C: Emerging Technologies, 134, 103466. https://doi.org/10.1016/j.trc.2021.103466

[21] Song, X., Guo, Y., Li, N., & Zhang, L. (2021). Online traffic flow prediction for edge computing-enhanced autonomous and connected vehicles. IEEE Transactions on Vehicular Technology, 70(3), 2101-2111. https://doi.org/10.1109/TVT.2021.3057109

[22] Jin, K., Wi, J., Lee, E., Kang, S., Kim, S., & Kim, Y. (2021). TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting. Expert Systems with Applications, 186, 115738. https://doi.org/10.1016/j.eswa.2021.115738

[23] Cai, L., Janowicz, K., Mai, G., Yan, B., & Zhu, R. (2020). Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. Transactions in GIS, 24(3), 736-755. https://doi.org/10.1111/tgis.12644

[24] Yan, H., Ma, X., & Pu, Z. (2022). Learning dynamic and hierarchical traffic spatiotemporal features with transformer. IEEE Transactions on Intelligent Transportation Systems, 23(11), 22386-22399. https://doi.org/10.1109/TITS.2021.3102983

[25] Reza, S., Ferreira, M. C., Machado, J. J. M., & Tavares, J. M. R. S. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Systems with Applications, 202, 117275. https://doi.org/10.1016/j.eswa.2022.117275

[26] Feng, H., & Jiang, X. (2022). Multi-step ahead traffic speed prediction based on gated temporal graph convolution network. Physica A: Statistical Mechanics and its Applications, 606, 128075. https://doi.org/10.1016/j.physa.2022.128075

[27] Mallick, T., Balaprakash, P., Rask, E., & Macfarlane, J. (2020). Graph-partitioning-based diffusion convolutional recurrent neural network for large-scale traffic forecasting. Transportation Research Record, 2674(9), 473-488. https://doi.org/10.1177/0361198120930010

[28] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[29] Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., & Li, T. (2018). Predicting citywide crowd flows using deep spatio-temporal residual networks. Artificial Intelligence, 259, 147-166. https://doi.org/10.1016/j.artint.2018.03.002

[30] Doğan, E. (2020). Analysis of the relationship between LSTM network traffic flow prediction performance and statistical characteristics of standard and nonstandard data. Journal of Forecasting, 39(8), 1213-1228. https://doi.org/10.1002/for.2683