MSGAT-GRU: A Multi-Scale Graph Attention and Recurrent Model for Spatiotemporal Road Accident Prediction

Thrinadh Pinjala[1,2], Aswin Ram Kumar Gannina[1,3], and Debasis Dwibedy[1,4]

[1] Department of Computer Science and Engineering, School of Computing,
Coimbatore, Amrita Vishwa Vidyapeetham, India
[2] cb.en.u4cse22141@cb.students.amrita.edu
[3] cb.en.u4cse22305@cb.students.amrita.edu
[4] debasis.dwibedy@gmail.com

Abstract. Accurate prediction of road accidents remains challenging due to intertwined spatial, temporal, and contextual factors in urban traffic. We propose MSGAT-GRU, a multi-scale graph attention and recurrent model that jointly captures localized and long-range spatial dependencies while modeling sequential dynamics. Heterogeneous inputs, such as traffic flow, road attributes, weather, points of interest, are systematically fused to enhance robustness and interpretability. On the Hybrid Beijing Accidents dataset, MSGAT-GRU achieves RMSE = 0.334 and F1-score = 0.878, consistently outperforming strong baselines. Cross-dataset evaluation on METR-LA under a 1-hour horizon further supports transferability, with RMSE = 6.48 (vs. 7.21 for GMAN model) and comparable MAPE. Ablations indicate that three-hop spatial aggregation and a two-layer GRU offer the best accuracy–stability trade-off. These results position MSGAT-GRU as a scalable and generalizable model for intelligent transportation systems, providing interpretable signals that can inform proactive traffic management and road safety analytics.

Keywords: Accident Prediction, Gated Recurrent Unit (GRU), Graph Neural Networks (GNN), Road Safety Enhancement, Spatio-temporal Data, Traffic Analysis

1 Introduction

Road traffic is an integral part of modern life, supporting economies and connecting communities across the globe. However, road accidents remain a leading cause of injury and death, with the World Health Organization estimating that approximately 1.19 million fatalities occur each year due to road accidents, making it one of the leading causes of mortality worldwide [22]. As urbanization accelerates and traffic density increases, the need for advanced predictive models capable of anticipating accident-prone scenarios has become more urgent. Effective accident prediction not only helps prevent fatalities but also supports

better traffic management, insurance planning, and infrastructure development, thereby contributing to safer and more efficient transportation systems [13], [8].

Research Motivation: Predicting accidents is an inherently complex task due to the interplay of spatial road networks, temporal traffic patterns, weather conditions, and socio-economic factors. Traditional machine learning and neural network models often focus on either spatial relationships or temporal dynamics, limiting their ability to fully capture real-world complexities [11], [1]. Furthermore, existing systems frequently struggle with heterogeneous data integration, class imbalance due to the rarity of accidents, and difficulties in adapting to rapidly changing traffic environments [23], [4]. These limitations reduce the effectiveness of prediction models when applied to diverse urban contexts or deployed in real-time traffic management systems [2], [23]. Motivated by these challenges, our research explores how spatiotemporal graph modeling can enhance accident prediction by integrating diverse data streams and learning patterns across both space and time. By leveraging real-time traffic flow data, meteorological information, road attributes, and calendar-based contextual factors, we aim to address the critical gaps in data heterogeneity, class imbalance, and transferability across cities with different traffic dynamics [10], [19]. Our work is inspired by the need for models that not only predict accident risks with high accuracy but also offer interpretable and actionable insights for transportation planners and safety practitioners.

Our Contribution: We propose MSGAT-GRU, an improved model that integrates multi-scale graph attention mechanisms with gated recurrent units (GRU) to effectively capture complex spatial and temporal dependencies in road accident prediction. We evaluate our approach on the Hybrid Beijing Accidents dataset, which includes approximately 7.7 million records, demonstrating significant improvements over existing methods in key metrics such as error reduction and classification performance. Through extensive experiments, including ablation studies and feature importance analysis, we highlight how the integration of heterogeneous data sources and multi-scale modeling enhances both the robustness and interpretability of accident predictions. The results of this study underscore the potential of the proposed model to contribute meaningfully to intelligent transportation systems and road safety interventions.

Organization: The structure of this article is outlined below. Section 2 provides an overview of prior research in road accident prediction, with particular focus on graph-based frameworks and modeling strategies. Section 3 introduces the problem formulation and describes the dataset employed in this work. Section 4 details the architecture and methodological design of the proposed MSGAT-GRU model. Section 5 reports the experimental evaluation, including comparisons with baseline models and an ablation analysis highlighting the role of different components. Finally, Section 6 summarizes the main contributions and discusses potential avenues for future research.

2   Foundation and Related Work

We first review the fundamental concepts underlying spatio-temporal graph modeling, which form the basis for advanced traffic analysis. We then present key contributions from existing graph-based approaches for road traffic and accident prediction, highlighting how they address spatial and temporal dependencies. Finally, we discuss the critical research gaps that persist in the existing literature, emphasizing the need for more robust, adaptive, and interpretable models.

2.1   Fundamental Concepts and Definitions

– *Graph Neural Network (GNN):* A class of neural networks designed to operate on graph data structures, consisting of nodes and edges, allowing for information propagation and learning on non-Euclidean data [15].
– *Spatial-Temporal Graph Neural Network (STGNN):* An extension of GNNs that captures both spatial relationships (through graph structure) and temporal dynamics (through sequence modeling) in a unified framework, especially suitable for traffic networks where both dimensions are critical [25].
– *Graph Attention Network (GAT):* A GNN variant that utilizes attention mechanisms to learn dynamically to assign importance weights to the neighboring nodes while passing messages, allowing for more flexible and context-aware information aggregation [18].
– *Gated Recurrent Unit (GRU):* A recurrent neural network (RNN) architecture used for modeling sequential data, employing gating mechanisms (reset and update gates) in order to regulate the flow of information and effectively capture temporal dependencies [6].
– *Attention Mechanism:* A technique that allows neural networks to focus on specific parts of input data that are most relevant to the task, assigning different weights to different input elements [3].

2.2   Related Work in Graph-Based Traffic and Accident Prediction

Graph-based frameworks have emerged as effective tools for modeling the intertwined spatial, temporal, and contextual factors in traffic systems. Prior studies can be grouped into five methodological directions:

Dynamic Graph Representations: Early approaches such as Dynamic-GRCNN [14] and DSTAGNN [9] introduced adaptive graph structures to capture evolving traffic flows. While these frameworks effectively model fluctuations, they neglect external contextual variables, such as weather and incidents, limiting their applicability to accident prediction.

Attention-Based Mechanisms: Models such as TFM-GCAM [5] and STGNN [21] employ attention layers to enhance spatiotemporal feature extraction. Although these methods improve short-term forecasting accuracy, their fixed architectures restrict adaptability to rare and unpredictable accident events.

Multi-Graph Approaches: Frameworks like DMGNN [24] integrate multiple graphs representing traffic patterns, congestion levels, and accident correlations. Despite

improved predictive power, their reliance on handcrafted relationships reduces generalizability across cities with heterogeneous road structures.

Heterogeneous Data Integration: DSTGCN [26] and large-scale accident studies by Nippani et al. [12] highlight the importance of integrating traffic, meteorological, and POI features. However, their coarse temporal resolution and limited strategies for handling class imbalance undermine real-time accident forecasting.

Network-Wide and Multi-View Learning: Recent methods such as MSGNN [16], STZITD-GNN [7], and MG-TAR [17] extend prediction to network-wide scales and introduce uncertainty quantification. Yet, these models still lack interpretability and fail to consistently transfer across diverse urban settings.

In summary, while graph-based learning has demonstrated clear promise in traffic forecasting, existing approaches are constrained by limited contextual integration, insufficient temporal granularity, and challenges in handling imbalanced accident datasets. These shortcomings motivate the need for more adaptive and interpretable frameworks.

2.3   Research Gaps

Building on the above limitations, we identify three critical research gaps that motivate our study in this paper:

(i) Lack of real-time contextual integration: Most existing models cannot incorporate external factors, such as weather, incidents, and points of interest, in real time. Fixed architectures fail to adapt to the evolving spatiotemporal patterns observed in real-world traffic systems.

(ii) Insufficient temporal resolution: Many studies aggregate data at coarse daily or monthly levels, overlooking the rapid fluctuations in traffic dynamics. This limits the ability to provide timely, actionable accident predictions.

(iii) Class imbalance in accident datasets: Since crashes are rare compared to normal traffic flow, severe imbalance skews predictions toward non-accident cases. Current evaluation strategies inadequately address this issue, leading to biased or unreliable outputs.

These gaps underscore the need for a robust and flexible framework capable of fusing heterogeneous, fine-grained data, modeling multi-scale spatial dependencies, and mitigating imbalance. Addressing these challenges forms the basis of our proposed MSGAT-GRU model.

3   Problem Statement

The objective of this study is to predict the probability of a traffic accident occurring on a road segment v_i during the next time interval $T+1$, given observations up to the current time T. We model the road network as an undirected graph $G = (V, E, A)$, where V denotes the set of nodes (road segments), E the set of edges (intersections or spatial connections), and A the adjacency matrix encoding the network topology.

Each node v_i is associated with a feature vector

$$x_i = \left[x_i^{spatial},\ x_i^{temporal},\ x_i^{external}\right],$$

where $x_i^{spatial}$ includes structural attributes such as road length, number of lanes, and nearby points of interest (POIs), $x_i^{temporal}$ captures recent traffic dynamics such as speed and flow patterns, and $x_i^{external}$ incorporates contextual information such as weather conditions and calendar-based factors.

The prediction task can be formulated as learning a mapping function:

$$\hat{y}_i = f\left(x_i^{spatial},\ x_i^{temporal},\ x_i^{external},\ X_{\text{neighbors}}\right),$$

where $\hat{y}_i \in [0, 1]$ represents the probability of an accident on segment v_i at time $T + 1$, and $X_{\text{neighbors}}$ denotes features from the k-hop neighborhood of v_i.

This formulation must capture four critical aspects: (1) spatial dependencies arising from road connectivity and neighboring segments, (2) temporal evolution of traffic patterns, (3) influence of real-time external factors such as weather and calendar events, and (4) the inherent class imbalance in accident datasets.

Addressing these challenges requires a model that integrates multi-scale spatial reasoning, sequential temporal modeling, and heterogeneous feature fusion. The following section presents our proposed methodology and illustrates the MSGAT-GRU model, designed to meet these requirements.

4 Our Proposed Methodology

This section presents the dataset used in the study, the overall methodology designed to address the problem, and the architecture of the MSGAT-GRU model developed to capture spatial, temporal, and contextual dependencies in accident prediction. We describe how the data sources are integrated and preprocessed to support model training, and how the proposed framework effectively learns from heterogeneous and imbalanced data to improve prediction accuracy.

4.1 Dataset Description

This study utilizes the Hybrid Beijing Accidents dataset, a comprehensive multi-source traffic dataset designed for spatiotemporal prediction tasks [26]. The dataset is publicly available on GitHub[5] and integrates four key components:

- *Accident records:* Detailed logs of accident events, including location, timestamp, severity, and type of incident.
- *Road network data:* Graph representation of the urban road infrastructure, where nodes correspond to road segments and edges capture connectivity through intersections. Attributes such as road type, speed limit, and number of lanes are included.

---

[5] https://github.com/yule-BUAA/DSTGCN

– *Traffic flow data:* Hourly measurements of vehicle count, average speed, and occupancy, providing both real-time and historical traffic dynamics.
– *External contextual data:* Weather conditions (temperature, precipitation, visibility, wind speed, and weather events) and point-of-interest (POI) information (schools, shopping centers, hospitals, etc.), capturing environmental and socio-economic influences on traffic safety.

The dataset spans a three-month period (August–November 2018) and includes millions of traffic and contextual records. To mitigate the inherent class imbalance, we adopt a balanced sampling strategy during training, maintaining an equal ratio of accident to non-accident instances. Extensive preprocessing aligns heterogeneous sources across space and time, ensuring that each road segment is represented as a node enriched with static attributes and dynamic features. This dataset is well-suited for our study as it provides heterogeneous, fine-grained information necessary to capture the spatial, temporal, and contextual dependencies that govern accident risk in large-scale urban environments.

4.2   Various Phases of Our Methodology

The proposed methodology follows a systematic pipeline, illustrated in Figure 1, which integrates heterogeneous data, ensures consistency through preprocessing, and prepares structured inputs for the MSGAT-GRU model.
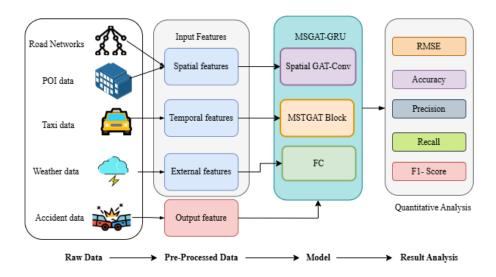

Fig. 1. Workflow of the proposed methodology, highlighting data integration, preprocessing, balanced sampling, and dataset partitioning prior to model training.

Data Integration: The first phase constructs meaningful feature representations

for each road segment by integrating three modalities: (i) *spatial features* combining graph-based road network attributes with point-of-interest (POI) density, (ii) *temporal features* derived from traffic speed sequences resampled at hourly intervals, and (iii) *external features* incorporating weather conditions and calendar-based indicators. This multimodal fusion provides a rich contextual basis for accident prediction.

Preprocessing: To ensure consistency across heterogeneous sources, several preprocessing steps are applied. Road segment attributes are standardized to maintain geometric and topological coherence. Traffic speed data, originally irregular, is aligned through optimized interpolation within ± 7 or ± 14 day windows, with mean imputation as fallback. All feature groups are normalized independently to maintain scale balance and improve convergence stability. For each accident event, a k-hop subgraph is dynamically extracted, enabling localized spatiotemporal context representation.

Balanced Sampling: Accident datasets are inherently imbalanced. To mitigate bias, positive samples are derived from real accident records while negative samples are selected from non-accident instances under comparable spatial constraints. Both are validated for geographic consistency to ensure representative learning.

Dataset Partitioning. Finally, the dataset is split into training (70%), validation (10%), and test (20%) subsets. This stratification minimizes selection bias and supports robust evaluation of model generalization.

Overall, this pipeline provides a structured and balanced representation of heterogeneous traffic environments, forming the foundation for the subsequent MSGAT-GRU model design.

4.3 Our Proposed MSGAT-GRU Model

The MSGAT-GRU model is designed to capture complex spatiotemporal dependencies by processing spatial, temporal, and external features through specialized branches before fusing them into a unified representation. Figure 2 illustrates the overall architecture.

Input Features and Branch Separation: Each road segment is characterized by three feature sets: (i) *spatial features*, encoding topological road attributes and points of interest, (ii) *temporal features*, representing historical traffic dynamics, and (iii) *external features*, such as weather and calendar data. These are processed independently to preserve modality-specific characteristics.

Spatial Feature Processing: Spatial features are passed through stacked Graph Attention (GAT) layers, enabling node-level embeddings that account for heterogeneous neighbor influence. For node i, the representation is updated as:

$$h'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j \right),$$
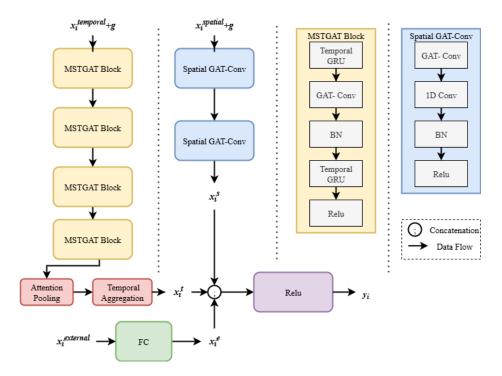
Fig. 2. Architecture of the proposed MSGAT-GRU model. Spatial, temporal, and external features are processed through dedicated branches with graph attention and recurrent units before feature fusion for accident risk prediction.

where α_{ij} is the attention coefficient learned via normalized importance scores. This dynamic weighting allows the model to prioritize influential neighbors over uniform aggregation.

Temporal Feature Processing: Temporal sequences are modeled using Multi-Scale Spatiotemporal Graph Attention (MSTGAT) blocks. Each block combines bidirectional GRUs, which capture sequential dependencies, with multi-scale GAT modules aggregating information from 1-hop, 2-hop, and 3-hop neighbors:

$$h_i^{\text{multi-scale}} = \big\|_{s=1}^{S} h_i^{(s)},$$

where $\|$ denotes concatenation across scales. This design preserves both local and global traffic dependencies without over-smoothing.

External Feature Processing: External factors are embedded through a lightweight fully connected (FC) network with normalization and ReLU activation:

$$z_{\text{external}} = \text{FC}(\text{BN}(\text{ReLU}(\text{Linear}_1(x_{\text{external}})))).$$

This ensures compact yet informative representations of weather and calendar influences.

Attention Pooling and Temporal Aggregation: To derive graph-level embeddings, attention pooling assigns learnable weights β_i to nodes, emphasizing their relative importance:

$$h_{\text{temp-agg}} = \sum_{i=1}^{N} \beta_i h_i.$$

The resulting sequence is passed through a bidirectional GRU, enabling the model to jointly capture short- and long-term temporal patterns.

Feature Fusion and Prediction: Finally, outputs from the three branches are concatenated:

$$h_{\text{final}} = [h_{\text{temp}} \,\|\, h_{\text{spatial}} \,\|\, z_{\text{external}}],$$

and passed through an FC layer with dropout and normalization to predict accident probability $\hat{y}$. This fusion strategy ensures comprehensive integration of spatial, temporal, and contextual dependencies.

The modular design of MSGAT-GRU provides three advantages: (i) fine-grained modeling of heterogeneous data sources, (ii) preservation of multi-scale dependencies across space and time, and (iii) interpretable representations that highlight influential road segments and contextual factors. The key notations and mathematical definitions used in this architecture are summarized in Table 1.

5 Performance Evaluation

This section presents the experimental evaluation of the proposed MSGAT-GRU model on the Hybrid Beijing dataset using widely accepted regression and classification metrics, including RMSE, Accuracy, Precision, Recall, and F1-score. The obtained results are compared against established baseline models to demonstrate the effectiveness of our approach. Additionally, we assess the model's performance on the METR-LA dataset by employing standard error metrics such as MAE, RMSE, and MAPE, and benchmark it against the state-of-the-art GMAN model. Furthermore, an ablation study is conducted to investigate the influence of key architectural parameters, specifically, the GAT hop size and GRU depth, on prediction performance. Finally, the computational cost associated with the model's design is analyzed to provide insights into its practical applicability.

5.1 Result Comparison and Discussions

We conducted extensive experiments on the Hybrid Beijing Accidents dataset using a 60-minute prediction horizon across diverse traffic settings, including urban, suburban, and rural road segments. The MSGAT-GRU model was benchmarked against prominent baselines, including DSTGCN, MSGCN, STGAT-GRU, and

Table 1. Notations and Definitions used in the MSGAT-GRU Model

Notation	Definition
$G_t = (V, E, X_t)$	raph at time t, with nodes V, edges E, and node feature matrix X_t.
V	Set of nodes representing road segments in the road network.
E	Set of edges representing connections between adjacent road segments.
$X_t \in \mathbb{R}^{|V| \times F}$	Node feature matrix at time t, with F features per node.
$x_t^i \in \mathbb{R}^F$	Feature vector for node v_i at time t, including spatial, temporal, and external features.
$p_{t+\Delta t}^i$	Predicted probability of a traffic accident at node v_i within future time window Δt.
T	Number of historical time steps used for prediction.
f	Function mapping historical graph sequence to accident probabilities.
x_i^{spatial}	Spatial features for node v_i, combining graph network data and POI data.
x_i^{temporal}	Temporal features for node v_i, derived from traffic speed data.
x_i^{external}	External features for node v_i, including weather and calendar information.
x_i^{b}	Point of Interest (POI) data for node v_i.
x_i^{g}	Graph network data for node v_i (e.g., road type, number of lanes).
x_i^{w}	Weather attributes for node v_i (e.g., temperature, precipitation).
x_i^{c}	Calendar features for node v_i (e.g., day of week, hour).
y	Binary label indicating accident occurrence at a specific time and location.
A	Adjacency matrix defining the road network topology.
h_i'	Updated feature representation for node v_i after GAT convolution.
α_{ij}	Normalized attention score between nodes v_i and v_j in GAT.
h_t	Hidden state of the GRU at time t for temporal feature processing.
$h_i^{\text{multi-scale}}$	Concatenated multi-scale spatial representation for node v_i.
z^{external}	Processed external feature embedding after fully connected layers.
β_i	Attention score for node v_i in temporal feature pooling.
$h^{\text{temp-agg}}$	Aggregated graph-level temporal representation.
h^{spatial}	Pooled spatial feature representation across nodes.
h^{final}	Final concatenated feature vector for accident probability prediction.
$\hat{y}$	Predicted accident probability output by the model.

Table 2. Performance comparison of MSGAT-GRU with baseline models on the Beijing dataset for 60-minute accident prediction

Model	RMSE	Accuracy	Precision	Recall	F1-Score
STGAT	0.378 ± 0.015	0.814 ± 0.012	0.831 ± 0.013	0.762 ± 0.014	0.795 ± 0.010
STGCN_GraphSAGE	0.370 ± 0.011	0.843 ± 0.009	0.777 ± 0.012	0.785 ± 0.013	0.781 ± 0.008
STGCN_LSTM	0.361 ± 0.013	0.801 ± 0.011	0.789 ± 0.010	0.798 ± 0.012	0.794 ± 0.009
STGAT_GRU	0.344 ± 0.012	0.812 ± 0.008	0.824 ± 0.011	0.893 ± 0.009	0.857 ± 0.007
MSGCN	0.357 ± 0.009	0.878 ± 0.006	0.835 ± 0.008	0.817 ± 0.010	0.826 ± 0.005
MSGCN_GRU	0.352 ± 0.008	0.883 ± 0.005	0.841 ± 0.007	0.828 ± 0.009	0.834 ± 0.006
DSTGCN	0.343 ± 0.011	0.867 ± 0.007	0.821 ± 0.0012	$\mathbf{0.896 \pm 0.016}$	0.8508 ± 0.011
MSGAT-GRU (Ours)	$\mathbf{0.334 \pm 0.016}$	$\mathbf{0.886 \pm 0.004}$	$\mathbf{0.896 \pm 0.005}$	0.861 ± 0.006	$\mathbf{0.878 \pm 0.004}$

STGCN variants. Table 2 summarizes the comparative performance. MSGAT-GRU consistently outperforms baselines across both regression and classification metrics. It reduces RMSE to 0.334, improving upon DSTGCN by approximately 1%, and achieves an F1-score of 0.878, a 2.1% gain over the strongest competitor. While these numerical margins may appear modest, they are significant in highly imbalanced accident datasets where even small improvements translate into better identification of rare events and fewer false alarms. Figure 3 illustrates performance across prediction horizons. As expected, accuracy decreases with longer horizons due to higher uncertainty. Nevertheless, MSGAT-GRU maintains relatively stable performance even at 60 minutes, outperforming all baselines in both RMSE and F1-score.

Discussion: The improved performance stems from three design choices. First, the multi-scale GAT captures hierarchical spatial correlations (1-hop to 3-hop
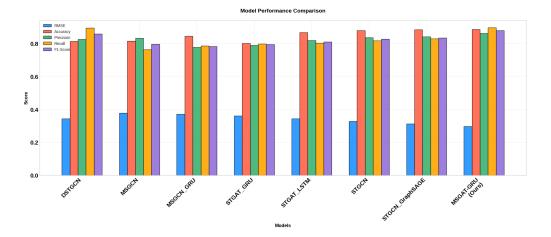
Fig. 3. Comparison of RMSE, Accuracy, Precision, Recall, and F1-score across different prediction horizons. MSGAT-GRU consistently outperforms baseline models, particularly at longer horizons where uncertainty is higher.

neighborhoods) more effectively than single-scale baselines. Second, bi-directional GRUs model sequential dynamics across time, improving recall without sacrificing precision. Third, integrating external features ensures that contextual factors such as weather and calendar events are explicitly considered. Together, these components yield a balanced improvement across regression and classification metrics, underscoring the robustness of the framework for accident prediction.

5.2   Ablation Study: Impact of GAT Hop and GRU Depth

To better understand the contributions of key architectural choices, we conducted an ablation study varying two hyperparameters: (i) the number of Graph Attention (GAT) hops and (ii) the depth of Gated Recurrent Units (GRU). The objective was to evaluate how spatial aggregation scope and temporal modeling depth influence accident prediction performance. The results are summarized in Table 3.

Effect of GAT Hop Sizes: Extending the receptive field from single-hop to multi-hop neighborhoods improves spatial representation learning. Incorporating 1–2–3 hops enables the model to capture both local interactions (e.g., immediate neighbors) and broader network-wide correlations (e.g., downstream congestion effects). This hierarchical aggregation consistently reduces RMSE and enhances F1-scores, confirming the benefit of multi-scale spatial reasoning.

Effect of GRU Depth: Increasing GRU layers improves temporal modeling up to a point. Two-layer GRUs provide the best trade-off between expressiveness and stability, yielding the strongest overall performance. Deeper configurations (three layers) show diminishing returns and occasional instability, likely due to overfitting and increased computational cost.

Table 3. Ablation study results under different GAT hop configurations and GRU depths on the Beijing dataset. Best results in bold.

GAT Hops	GRU Depth	RMSE	Precision	Recall	F1-Score
1	1	0.351	0.837	0.801	0.819
1-2	1	0.342	0.846	0.814	0.830
1-2-3	1	0.341	0.850	0.828	0.839
1	2	0.344	0.845	0.816	0.830
1-2	2	0.338	0.854	0.832	0.843
1-2-3	**2**	**0.334**	**0.896**	**0.861**	**0.878**
1	3	0.349	0.842	0.807	0.824
1-2	3	0.347	0.848	0.823	0.835
1-2-3	3	0.339	0.858	0.887	0.872

Key Insight: The best configuration, three-hop GAT aggregation with two-layer GRUs, achieves an RMSE of 0.334 and F1-score of 0.878. This balance between spatial context and temporal depth is further visualized in Figure 4, which highlights performance trends across different model variants. The figure illustrates that multi-hop aggregation provides robust spatial learning, while two-layer GRUs offer the optimal trade-off between accuracy and stability.
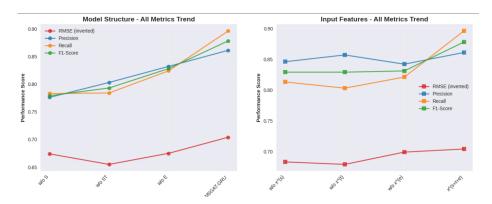

Fig. 4. Performance trends under varying GAT hop sizes and GRU depths. Multi-hop aggregation improves spatial learning, while two-layer GRUs achieve the best balance between predictive accuracy and stability.

5.3   Comparison with State-of-the-Art Model on METR-LA Dataset

To further assess the generalizability of MSGAT-GRU beyond the Hybrid Beijing dataset, we benchmarked it on the widely used METR-LA traffic forecasting dataset under a 1-hour prediction horizon, following standard spatio-temporal evaluation protocols. The state-of-the-art Graph Multi-Attention Network (GMAN) [27] was selected as the comparison baseline, given its strong

performance in modeling spatial dependencies through attention mechanisms. As shown in Table 4, MSGAT-GRU attains a lower RMSE (6.48 vs. 7.21) and

Table 4. Performance comparison between MSGAT-GRU and GMAN on the METR-LA dataset for 1-hour prediction

Model	MAE	RMSE	MAPE (%)
GMAN	3.40	7.21	9.72
MSGAT-GRU	3.43	**6.48**	**9.69**

a marginally better MAPE (9.69% vs. 9.72%), while exhibiting a slightly higher MAE (3.43 vs. 3.40). These results suggest that although GMAN better minimizes small deviations, MSGAT-GRU is more effective in controlling large errors, critical for accident prediction, where rare but severe deviations can disproportionately affect safety outcomes. The ability of MSGAT-GRU to aggregate spatial cues across multiple hops and simultaneously encode temporal dynamics allows it to adapt more robustly to abrupt fluctuations and noise inherent in real-world urban traffic.

Overall, the transfer learning results on METR-LA reinforce the robustness of MSGAT-GRU across diverse datasets, complementing the improvements observed in Section 5.2. This demonstrates the model's capability to generalize beyond city-specific conditions and underscores its potential for deployment in heterogeneous traffic environments.

5.4   Inferences

The experimental results collectively highlight how MSGAT-GRU addresses key challenges in spatio-temporal accident prediction. Multi-scale graph attention captures both localized and long-range dependencies, while recurrent modeling preserves sequential patterns, leading to consistent improvements in recall and F1-score on imbalanced accident data. The ablation study further confirmed that a three-hop receptive field combined with a two-layer GRU yields the optimal balance between accuracy and stability. The cross-dataset validation on METR-LA provided additional evidence of generalizability, showing that the model remains robust under traffic conditions markedly different from Beijing. This indicates that the fusion of heterogeneous features, including traffic flow, weather, and points of interest, not only enhances prediction within a single domain but also facilitates transferability across diverse urban settings.

Taken together, these findings establish MSGAT-GRU as a scalable and interpretable framework for accident prediction. By effectively unifying spatial reasoning, temporal sequence modeling, and contextual integration, the model sets a strong foundation for advancing intelligent transportation systems and for supporting real-time traffic safety interventions.

5.5   Computation Cost Analysis

All experiments were conducted on a machine with an Intel Core i7-12700H, NVIDIA GeForce RTX 3050 Ti, and 16GB RAM. With the default training configuration, each epoch took approximately 10 minutes. This overhead primarily stems from the *multi-scale* graph attention used in the architecture (see Figure 2), which aggregates information across multiple hop distances in parallel rather than relying on single-hop propagation.
[6] To contextualize cost, let N be the number of nodes, E the number of edges, T the look-back length, S the number of spatial scales (hops), H the number of attention heads, and d the hidden size. A single MSGAT layer incurs

$$\mathcal{O}\big(SH \cdot (Ed + Nd^2)\big)$$

per time step due to edge-wise message passing and node-wise transformations; the temporal GRU contributes

$$\mathcal{O}\big(T\,Nd^2\big).$$

Hence, inference scales *linearly* with S and H, *near-linearly* with E for sparse graphs, and *linearly* with T. In practice, the transition from single-hop to three-hop aggregation increases spatial cost roughly proportionally to S, which is consistent with the accuracy gains observed in the ablation (Table 3).
From a deployment standpoint, three techniques were most beneficial: (i) sparse adjacency storage to reduce memory bandwidth, (ii) moderate hidden sizes to control Nd^2 terms, and (iii) caching static embeddings for frequently queried segments. While multi-scale attention introduces overhead relative to single-hop baselines, the *accuracy improvements* reported earlier (Tables 2 and 3) provide a favorable accuracy–cost trade-off for safety-critical applications.

6   Conclusion

We proposed MSGAT-GRU, a multi-scale graph attention and recurrent model for spatio-temporal accident prediction that fuses traffic flow, road-network attributes, weather, and points of interest. On the Hybrid Beijing Accidents dataset, the model achieves RMSE = 0.334 and F1-score = 0.878, outperforming strong graph-based baselines across both regression and classification metrics. Cross-dataset evaluation on METR-LA (1-hour horizon) further indicates robust transferability with a lower RMSE = 6.48 than GMAN (7.21), while maintaining comparable MAPE. Ablations confirm that three-hop spatial aggregation coupled with a two-layer GRU provides the best accuracy–stability trade-off.
Limitations: Balanced sampling improves learning on imbalanced data but may shift operating points in deployment where accidents are rare. Multi-scale attention introduces additional computational overhead, which can be constraining

---

[6] The MSTGAT blocks are depicted in the model architecture (Figure 2); they are not part of the data workflow in Figure 1.

under tight latency or edge-resource budgets.

Outlook: Future work will explore (i) *lightweight* variants (e.g., sparse or neighborhood-sampled attention, parameter sharing, and quantization/pruning), (ii) *explainability and calibration* for risk-aware decision support, and (iii) integration of real-time connected-vehicle and driver-behavior streams for finer-grained, short-horizon risk estimation. These directions aim to retain the demonstrated accuracy while improving deployability and trust in intelligent transportation systems.

References

1. Abbas MS (2023) Accident prediction using machine learning: analyzing weather conditions, and model performance. University of Oulu, Faculty of Information Technology and Electrical Engineering, Master's thesis, May 2023.
2. Azadani MN and Boukerche A (2021) Driving behavior analysis guidelines for intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems, 23(7): 6027–6045.
3. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations (ICLR), 2015.
4. Berhanu Y, Alemayehu E, and Schröder D (2023) Examining car accident prediction techniques and road traffic congestion: A comparative analysis of road safety and prevention of world challenges in low-income and high-income countries. Journal of Advanced Transportation. https://doi.org/10.1155/2023/6643412.
5. Chen J, Zheng L, Hu Y, Wang W, Zhang H, Hu X (2024) Traffic flow matrix-based graph neural network with attention mechanism for traffic flow prediction. Information Fusion, 104(C). https://doi.org/10.1016/j.inffus.2023.102146.
6. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): 1724–1734.
7. Gao X, Jiang X, Haworth J, Zhuang D, Wang S, Chen H, Law S (2024) Uncertainty-aware probabilistic graph neural networks for road-level traffic crash prediction. Accident Analysis & Prevention, 208: 107801.
8. Khan AF and Ivan P (2023) Integrating machine learning and deep learning in smart cities for enhanced traffic congestion management: an empirical review. J. Urban Dev. Manag, 2(4): 211–221.
9. Lan S, Ma Y, Huang W, Wang W, Yang H, Li P (2022) Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In International conference on machine learning: 11906-11917, PMLR.
10. Marcillo P, Valdivieso AL, Caraguay, and Hernández-Álvarez M (2022) A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources. Applied Sciences, 12(9): 4529. https://doi.org/10.3390/app12094529.
11. Miglani A and Kumar N (2019) Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. Vehicular Communications, 20:100184.

12. Nippani A, Li D, Ju H, Koutsopoulos H, Zhang H (2023) Graph Neural Networks for Road Safety Modeling: Datasets and Evaluations for Accident Analysis. In Advances in Neural Information Processing Systems 36:52009–52032.
13. Partheeban P, Arunbabu E, and Hemamalini RR (2008) Road accident cost prediction model using systems dynamics approach. Transport, 23(1): 59–66.
14. Peng H, Wang H, Du B, Bhuiyan MZ, Ma H, Liu J, Wang L, Yang Z, Du L, Wang S, Yu PS (2020) Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. Information Sciences, 521: 277–290.
15. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. IEEE Transactions on Neural Network, 20(1): 61–80.
16. Tran T, He D, Kim J, Hickman M (2023) MSGNN: A Multi-structured Graph Neural Network model for real-time incident prediction in large traffic networks. Transportation research part C: emerging technologies, 156: 104354.
17. Trirat P, Yoon S, Lee JG (2023) MG-TAR: Multi-view graph convolutional networks for traffic accident risk prediction. IEEE Transactions on Intelligent Transportation Systems, 24(4): 3779–3794.
18. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In Proceedings of the International Conference on Learning Representations (ICLR).
19. Wang W, Yang S, and Zhang W (2021) Risk prediction on traffic accidents using a compact neural model for multimodal information fusion over urban big data. arXiv:2103.05107.
20. Wang L, Geng X, Ma X, Liu F, Yang, Q (2019) Cross-City Transfer Learning for Deep Spatio-Temporal Prediction. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), pp. 1893–1899.
21. Wang X, Ma Y, Wang Y, Jin W, Wang X, Tang J, Jia C, Yu J (2020) Traffic flow prediction via spatial temporal graph neural network. In Proceedings of the web conference, 2020: 1082–1092.
22. World Health Organization (2023) Road Traffic Injuries. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
23. Wu P, Chen T, Wong YD, Meng X, Wang X, and Liu W (2023) Exploring key spatio-temporal features of crash risk hot spots on urban road network: A machine learning approach. Transportation Research Part A: Policy and Practice, 173:103717.
24. Ye Y, Xiao Y, Zhou Y, Li S, Zang Y, Zhang Y (2023) Dynamic multi-graph neural network for traffic flow prediction incorporating traffic accidents. Expert Systems with Applications, 234: 121101.
25. Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI): 3634–3640.
26. Yu L, Du B, Hu X, Sun L, Han L, Lv W (2021) Deep spatio-temporal graph convolutional network for traffic accident prediction. Neurocomputing, 423:135-47.
27. Zheng, Chuanpan, et al. (2020) Gman: A graph multi-attention network for traffic prediction. Proceedings of the AAAI conference on artificial intelligence, 34(1). https://doi.org/10.1609/aaai.v34i01.5477.