Event-Aware Imputation and Prediction of Urban Traffic Using Deep Spatiotemporal Learning Models

Event-Aware Imputation and Prediction of Urban Traffic Using Deep Spatiotemporal Learning Models

By
ALI ARDESTANI

A Thesis Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy
in
Civil Engineering

McMaster University
Hamilton, Ontario

Doctor of Philosophy (2025)
Department of Civil Engineering
McMaster University
Hamilton, Ontario, Canada

TITLE: Event-Aware Imputation and Prediction of Urban Traffic Using Deep Spatiotemporal Learning Models

AUTHOR: Ali Ardestani

SUPERVISOR:
Dr. Hao Yang, Associate Professor, Civil Engineering, McMaster University
Dr. Saeideh Razavi, Professor, Civil Engineering, McMaster University

Lay Abstract

Cities rely on traffic data to keep roads flowing smoothly, manage congestion, and ensure safety during busy times. However, traffic information is often incomplete due to sensor failures, gaps in vehicle tracking, or delays in communication. At the same time, special events such as concerts, sports games, and festivals create sudden and unusual traffic surges that are difficult to predict with traditional methods. This dissertation focuses on solving both of these challenges by creating new machine learning models that can fill in missing traffic data and make more reliable predictions about how traffic will behave during social events.

The research introduces two main innovations. First, a two-step method for handling missing data was developed. The method combines traditional machine learning with advanced artificial intelligence to reconstruct incomplete traffic information quickly and accurately. Second, a new predictive model was designed that takes into account not only past traffic patterns but also details about upcoming events, such as their type, location, and size. By doing so, the model is better able to anticipate sudden disruptions and provide more reliable forecasts.

The findings show that these approaches significantly improve both the accuracy of traffic data and the reliability of traffic forecasts, especially near event venues and during peak disruption times. In practice, this means that transportation agencies can better prepare for and respond to congestion around stadiums, concert halls, and city festivals, making travel smoother, safer, and more sustainable for everyone.

Abstract

This dissertation presents a comprehensive investigation into the dual challenges of missing traffic data and the complexities of traffic speed prediction during social events, a topic of growing relevance in urban mobility systems. Urban centers are increasingly experiencing non-recurring disruptions caused by concerts, sports games, festivals, and other social activities, which introduce sharp deviations in regular traffic patterns. At the same time, traffic data, which are foundational for intelligent transportation systems (ITS), often suffer from incompleteness due to sensor failures, transmission errors, and insufficient probe vehicle coverage. This research addressed these intertwined challenges by developing a unified framework combining robust imputation methods with deep learning-based event-aware prediction architectures.

The first contribution is the development of a two-stage imputation pipeline that integrates ensemble-based and generative approaches. Random Forest models are employed to provide fast, robust estimates, while Generative Adversarial Imputation Networks (GAIN) refine the results, capturing complex dependencies and uncertainty. Experiments on Hamilton, Ontario data demonstrate that the framework reduces imputation error (MAPE) by 20–30% compared to traditional methods, while maintaining scalability under varying missingness levels.

The second major contribution is the development of an Event-Aware LSTM (EA-LSTM) model that explicitly incorporates structured social event features—such as event type, timing, location, and attendance—into a spatiotemporal architecture combining Graph Convolutional Networks, bidirectional LSTMs, and attention mechanisms. The EA-LSTM significantly improves prediction accuracy during disruptions, reducing average error to 3.4% network-wide and under 9% near event venues, outperforming conventional deep learning baselines.

The findings demonstrate that integrating contextual event information enhances both traffic imputation and prediction, leading to more robust, interpretable, and scalable models. The research provides practical insights for the deployment of real-time ITS applications, offering tools to support congestion management, dynamic signal control, and event traffic planning in complex urban environments.

Acknowledgments

As I reach the conclusion of my doctoral journey, I wish to express my deepest gratitude to those who have supported and guided me throughout this process.

First and foremost, I extend my heartfelt thanks to my parents for their unconditional love, sacrifices, and encouragement. Their belief in me has been the foundation of my perseverance and growth, both personally and academically.

I am profoundly grateful to my supervisor, Dr. Hao Yang, and my co-supervisor, Dr. Saiedeh Razavi, for their invaluable guidance, patience, and dedication. Their expertise, constructive feedback, and continuous support have shaped my research and contributed immensely to my development as an independent scholar. I am privileged to have had the opportunity to work under their supervision.

I also wish to sincerely thank the members of my supervisory committee for their time, insightful feedback, and thoughtful questions, all of which have greatly enriched this dissertation.

In addition, I am thankful to my colleagues, collaborators, and friends who have offered encouragement and shared their knowledge throughout my doctoral studies. Their support has provided balance and motivation during the most challenging phases of this journey.

Finally, I gratefully acknowledge McMaster University and the Department of Civil Engineering for providing the resources, environment, and opportunities that have enabled me to pursue this research.

This dissertation would not have been possible without the collective support of these individuals and institutions, and I am sincerely indebted to them all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

Urban transportation systems are increasingly dependent on data-driven technologies to support real-time monitoring, forecasting, and decision-making for traffic management. The proliferation of traffic sensors, GPS-equipped probe vehicles, and connected infrastructure has led to a surge in the availability of spatiotemporal traffic data. These datasets are foundational to the operation of Intelligent Transportation Systems (ITS) and Advanced Traffic Management Systems (ATMS), enabling the optimization of signal control, congestion mitigation, and mobility forecasting.

However, the reliability of such data is frequently compromised by missing values due to sensor malfunctions, communication failures, or low probe penetration. This issue becomes particularly acute during atypical conditions such as social events, including concerts, festivals, sporting events, and parades, which induce non-recurring disruptions in traffic flow. These events challenge traditional traffic forecasting models that rely on assumptions of periodicity, continuity, and smooth spatiotemporal correlations (Xing et al., 2023). Incomplete data in such contexts undermines both real-time traffic estimation and the downstream performance of predictive models (Zhang et al., 2025).

While various statistical and machine learning-based imputation techniques have been proposed, they often lack robustness under the complex, nonlinear, and event-driven dynamics of urban traffic. At the same time, advancements in deep learning, particularly spatiotemporal architectures such as Graph Convolutional Networks (GCNs) and Long Short-Term Memory (LSTM) networks, have shown promise in capturing intricate spatial dependencies and long-term temporal trends. Yet, these models typically underperform in the presence of missing or irregular data and are rarely designed to incorporate exogenous contextual features such as event metadata.

In parallel, growing access to social event data from municipal records, public APIs, and social media platforms presents an opportunity to enrich traffic modeling. Event-related features such as attendance size, event type, location, and timing can provide valuable context to enhance both data imputation and traffic speed prediction during high-disruption periods (Okukubo et al., 2022). The integration of these features into machine learning structure, however, remains an open research challenge.

This dissertation is motivated by the need for a unified framework that enhances the accuracy of traffic data imputation and short-term prediction, while accounting for the impact of social events, a concept we refer to as event-aware. It explores novel combinations of ensemble learning, generative models, and deep spatiotemporal architectures augmented by structured event data to address the dual challenges of missing data and predictive uncertainty during social events. The overarching goal is to contribute practical, scalable, and interpretable tools for real-time urban traffic monitoring and forecasting in smart cities.

1.2 Traffic Speed Prediction

Traffic speed prediction has been extensively studied in the field of transportation engineering, particularly with the advent of large-scale spatiotemporal traffic datasets and the maturation of machine learning models. Classical approaches to traffic forecasting such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA, and Kalman Filtering provided interpretable, time-series-based modeling under assumptions of stationarity and linearity (Lv et al., 2014). While effective for isolated road segments and short-term predictions, these models are fundamentally limited in their ability to generalize across the complex spatial dependencies present in road networks, and they struggle to adapt to the nonlinear and dynamic nature of urban congestion (Cao et al., 2021).

As data availability and computational power improved, the field shifted toward data-driven methods. Shallow machine learning models such as k-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Random Forests (RF) were increasingly adopted to account for nonlinear patterns in traffic data (Razali et al., 2021). While these methods offered greater flexibility, they still fell short in capturing long-range temporal dependencies and spatial heterogeneity intrinsic to large urban networks. Moreover, their reliance on handcrafted features limited their scalability and adaptability (Deng et al., 2022).

To address these shortcomings, deep learning models emerged as state-of-the-art tools for traffic speed prediction. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly LSTM networks, demonstrated substantial improvements by learning temporal dependencies directly from raw data (Yu et al., 2018). LSTMs, in particular, became a popular choice for modeling the evolution of traffic speeds over time due to their gating mechanisms that mitigate the vanishing gradient problem. Nevertheless, LSTMs alone lack spatial awareness and are insufficient for network-wide forecasting tasks (Zhang et al., 2023).

In response, hybrid models incorporating both spatial and temporal features were developed. GCNs have played a pivotal role in this transition. By representing traffic networks as graphs where nodes denote sensors or road segments and edges reflect physical or functional connectivity GCNs enable spatial dependencies to be modeled more explicitly. Integrating GCNs with LSTMs created a powerful framework for spatiotemporal prediction (Yin et al., 2023). Such combinations, however, still exhibit limitations in handling asynchronous data updates, varied road segment characteristics, and the dynamic topology of urban traffic networks.

A key limitation observed in the existing literature is the underperformance of models during high-variability periods such as peak hours or after traffic incidents. Many models optimize for average-case accuracy, sacrificing robustness in the tail-end of the distribution where traffic anomalies exists. Additionally, while some architectures are evaluated on large datasets, they often lack generalizability across different cities or regions, suggesting overfitting to local patterns (Chen et al., 2024; Harrou et al., 2024).

Furthermore, transformer-based models, such as Temporal Fusion Transformers and Informer networks, have recently been introduced in traffic prediction. These models offer promising long-range memory capabilities and improved scalability. However, their high computational cost and data-hungry nature pose challenges for real-time deployment (Song et al., 2024a). Unlike LSTM-GCN hybrids, transformer-based methods also require more careful calibration and struggle with limited labeled data, which is common in smaller urban systems.

Another critical limitation in prior work is the treatment of traffic speed prediction as a purely data-driven exercise without adequate contextualization. Domain-specific knowledge, such as known bottlenecks, lane restrictions, or scheduled disruptions, is seldom incorporated directly into model architectures (Ma et al., 2019). This limits the potential of predictive systems in supporting decision-making for operational management.

In contrast, this thesis aims to bridge the divide between data-driven and context-aware modeling. By embedding structured contextual features directly into the architecture, such as through attention mechanisms and enriched node representations, the developed models achieve both high predictive accuracy and meaningful interpretability.

Overall, while the field has made considerable strides in modeling spatiotemporal traffic dynamics, significant opportunities remain in enhancing generalization, robustness, and integration with domain knowledge.

1.3 Missing Traffic Data Imputation

Missing data in traffic datasets significantly compromises the performance of Intelligent Transportation Systems (ITS), particularly affecting real-time applications such as traffic state estimation, signal control optimization, and travel time prediction. These gaps commonly arise from communication failures, sensor degradation, inclement weather conditions, or insufficient probe vehicle density (Zhang et al., 2024).

Traditional statistical methods were among the first to be employed for imputing missing traffic data. Techniques such as mean substitution, linear interpolation, and spline fitting were straightforward and fast, but their effectiveness declines with longer or systematic missingness (Smith et al., 2003). Time series models like ARIMA and seasonal decomposition offer improved performance by leveraging temporal structures (Stathopoulos and Karlaftis, 2003), but are sensitive to model specification and require stationarity assumptions that often do not hold in dynamic urban traffic environments.

Bayesian approaches gained popularity for their ability to incorporate prior information and model uncertainty. Ni and Leonard (Ni and Leonard, 2005) applied Bayesian networks with Markov Chain Monte Carlo (MCMC) techniques to estimate missing loop detector data by modeling conditional dependencies across traffic flows. However, these methods are computationally expensive and scale poorly with network size.

Spatial interpolation techniques such as Kriging and co-Kriging use spatial correlation between sensors to estimate missing values, assuming that nearby road segments exhibit similar traffic patterns (Bae et al., 2018). While effective in dense sensor networks, their performance deteriorates in sparse deployments or during events that cause localized disruptions.

With the rise of machine learning, algorithms such as k-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM) have been used for imputation. KNN, for instance, infers missing values based on the most similar complete observations

(Zheng et al., 2015). However, it is sensitive to the choice of distance metrics and struggles with high-dimensional data. Random Forests (RF) offer improved robustness by aggregating multiple decision trees and capturing nonlinear dependencies (Tang et al., 2017).

Matrix factorization and low-rank approximation methods, such as Non-negative Matrix Factorization (NMF) and Probabilistic PCA, model the traffic dataset as a low-dimensional structure and reconstruct missing values by learning latent features (Tan et al., 2013; Yao et al., 2018). These methods assume that traffic data exhibit low-rank patterns (e.g., daily periodicity) and are particularly effective when the missingness pattern is random and not extensive.

Tensor-based methods extend matrix factorization to multi-dimensional arrays, capturing complex interactions among time, location, and other covariates. Chen et al. (Chen et al., 2022b) proposed a tensor completion approach to recover missing entries in multi-way traffic tensors, showing improved accuracy over traditional 2D matrix methods.

In the deep learning domain, Denoising Autoencoders (DAE) and Recurrent Neural Networks (RNN) have been used to recover corrupted traffic data. DAEs are trained to reconstruct original input from partially corrupted versions, effectively learning complex representations for imputation (Duan et al., 2016). Ye et al. (Ye et al., 2021) applied Convolutional Autoencoders (CAE) for spatiotemporal traffic imputation, achieving strong performance under moderate missingness. However, these methods typically require substantial training data and careful tuning.

Other deep architectures include Recurrent Variational Autoencoders (RVAE), which model sequential dependencies and uncertainty simultaneously (Ma et al., 2019), and Graph Neural Networks (GNN), which incorporate spatial dependencies by treating the road network as a graph (Zhao et al., 2021). These models have demonstrated superior performance in capturing both spatial structure and temporal evolution in traffic data.

Despite these advancements, several limitations persist in the traffic data imputation literature. Most existing studies assume that missing data occur at random (MAR), an assumption frequently violated in real-world settings where missingness is often correlated with traffic conditions such as congestion or external disruptions like road construction or an event. Moreover, most prior work does not incorporate external contextual information, such as road hierarchy, or multimodal data sources that could enhance the precision of imputation under complex traffic scenarios. In response to these challenges, this dissertation proposes an efficient, event-aware imputation framework that leverages structured event metadata while maintaining reliable accuracy. By integrating computationally lightweight ensemble learning techniques with robust spatial-temporal representations, the framework offers a scalable solution suitable for real-time urban traffic applications, while preserving interpretability and resilience under non-random missingness conditions.

1.4 Social Events

The impact of large-scale events such as concerts, sports matches, and festivals on urban traffic dynamics has attracted increasing research attention. Traditional traffic forecasting models, optimized for routine conditions, struggle to capture the sudden, non-recurring disruptions caused by social events. This section reviews key developments in event-aware traffic modeling and highlights remaining gaps.

We distinguish two operationally distinct categories of social events:

- **Planned events (scheduled).** Events publicly announced in advance (e.g., league sports, concerts, parades, planned festivals). They provide structured, pre-event metadata such as *start/end time, venue location, event type, and expected attendance*. These signals are available hours to weeks ahead of time through municipal calendars, ticketing platforms, and official web APIs. Their traffic impacts are *non-recurring but predictable in timing and location*, which enables event-aware feature engineering and model conditioning.

- **Unplanned events (unscheduled).** Disruptions with little to no advance notice (e.g., collisions, vehicle breakdowns, sudden road closures, spontaneous protests, police activity, extreme weather). Signals for these events are detected ex post via incident logs, 911 feeds, or social media. Their timing and spatial footprint are *irregular* and often evolve rapidly, which requires online detection and recalibration rather than preconditioning.

Early event-aware methods used indicator variables to flag event periods within regression or time-series models. For instance, a group of researchers integrated event timing as dummy variables in ARIMA or regression frameworks to adjust forecasts during major events (e.g., holiday surges) (Cools et al., 2007). These methods provide coarse corrections and often fail to adapt to varying event types and sizes.

Machine learning models that incorporate exogenous event features have shown improved performance. (Yao and Qian, 2021a) mined late-night Twitter sentiment and activity to predict next-day morning congestion, showing stronger performance than non-event models. Studies in India and Shanghai used social media signals such as tweets or Weibo posts to detect traffic disruptions and feed event context into prediction models (Dabiri and Heaslip, 2019; Chang et al., 2022). While effective, such models remain limited by noisy social data and tend to focus on incident detection rather than systematic event modelling.

Hybrid models explicitly incorporating event metadata have demonstrated higher accuracy. (Essien et al., 2021), for example, fused Twitter-derived indicators with traffic sensors using a bidirectional LSTM autoencoder, improving multi-step flow prediction during disruptions. (Song et al., 2024b) introduced sports schedules directly into graph-attention transformer models, further enhancing long-horizon forecasts of traffic impacts during sports events. However, both studies face limitations in terms of generalizability across different event types and urban contexts. Moreover, their reliance on social media-derived features introduces variability due to inconsistent data quality and uneven user engagement.

Other approaches include pattern-aware regression models trained on historical event day data (Okukubo et al., 2022), and Heterogeneous Graph Attention models that incorporate changes in traffic network topology during events (Du et al., 2021). The pattern-aware regression method by Okukubo et al. (2022) reduces overfitting and improves prediction on large events by learning a weighted combination of latent traffic patterns; however, its reliance on limited event-specific data may limit generalizability to new event types or unusual conditions. The HGA-ResTCN approach by Du et al. (2021) dynamically adjusts graph structures to reflect disrupted spatial dependencies and captures temporal correlations via residual TCNs but its complex architecture and focus on incident-induced anomalies pose challenges for real-time deployment during high-complexity events.

Reliable extraction of event features is challenging, and in a recent study a NLP pipeline that detects timing, type, and location from sources such as social media, ticketing platforms, and municipal calendars has achieved high accuracy in event detection. (Tao et al., 2022a) introduced SMAFED, a robust framework that resolves noisy language and disambiguates slang, achieving precision up to 0.92 and recall of 0.79 in event detection tasks. More recently, Wang et al. (Wang et al., 2025) leveraged large language models (LLMs) to extract event attendance, popularity, and promotion intensity from multi-source online data, integrating these features into a machine learning model that achieved an R^2 above 0.85 for daily visitor flow forecasting in Hong Kong. Despite this progress, these methods face limitations: the SMAFED framework requires extensive pre-training on domain-specific slang and may generalize poorly across regions, while the LLM-based visitor flow approach relies heavily on rich textual and engagement data, limiting its applicability in areas with sparse online coverage or less event data availability.

Despite substantial progress in event-aware traffic prediction, several limitations continue to constrain the applicability and robustness of current approaches. Many models are calibrated for specific event types, most notably sports games or concerts, reducing their generalizability across a wider spectrum of disruptions, such as community festivals and gatherings. Another persistent challenge lies in the limited integration of multimodal contextual data, such as social events, weather conditions, or construction zones, despite their known influence on traffic dynamics during events. Also, while deep learning architectures have improved prediction accuracy, they frequently lack explicit mechanisms for uncertainty quantification, which is critical for operational decision-making under volatile or ambiguous conditions.

Addressing the key limitations identified in the literature including limited generalizability across diverse event types, insufficient integration of multimodal contextual features, and the lack of uncertainty modeling, this dissertation proposes a unified, scalable, and interpretable framework that enhances both traffic data imputation and short-term speed prediction during social events. The proposed approach integrates structured event metadata (such as event type, timing, location, and attendance) with deep spatiotemporal learning architectures that explicitly capture spatial and temporal dependencies. Furthermore, the framework incorporates mechanisms for uncertainty quantification to support robust forecasting in volatile conditions. By bridging the disconnect between traditional traffic models and the dynamic realities of urban mobility during large-scale events, this research advances a novel methodology capable of delivering accurate, context-aware, and operationally viable predictions across a wide range of urban settings and disruption scenarios.

1.5 Research Gaps

Urban transportation systems depend heavily on timely, accurate, and complete traffic data to support real-time decision-making in traffic management, congestion control, and emergency response. However, the reliability of these data streams is often compromised by missing values, typically caused by sensor failures, communication losses, adverse weather, or low probe penetration in certain areas. Recent studies, such as Zhang et al. (Zhang et al., 2021) and Kong et al. (Kong et al., 2022), highlight that data gaps exceeding even 5–10% can significantly degrade the performance of Intelligent Transportation Systems (ITS) applications. The problem becomes even more pronounced during large-scale social events such as concerts, festivals, and sporting matches which introduce sudden and spatially localized surges in traffic demand. These disruptions often coincide with increased rates of missingness, as congestion-related sensor outages or temporary communication bottlenecks become more prevalent (Ma et al., 2020; Tao et al., 2022b).

A critical challenge in this field is that most existing imputation methods assume that data are missing at random (MAR). However, several empirical studies (Tang et al., 2021; Zhou et al., 2023) demonstrate that missingness in traffic datasets is frequently correlated with external disruptions such as road closures, severe congestion, or event-driven anomalies. This violation of the MAR assumption can lead to biased imputation and subsequently flawed downstream predictions. Furthermore, traditional statistical imputation techniques, including ARIMA-based or Kalman-filtering approaches, while computationally efficient, fail to capture the complex spatiotemporal dependencies inherent in urban traffic networks, particularly under volatile conditions (Stathopoulos and Karlaftis, 2003; van Lint and van Zuylen, 2005).

In recent years, machine learning and deep learning frameworks have advanced traffic imputation by leveraging spatial and temporal relationships among road segments. Models such as graph convolutional networks (GCNs), recurrent neural networks (RNNs), and their hybrids have achieved notable success in recovering missing data under routine traffic patterns (Liang et al., 2021; Zhao et al., 2021; Chen et al., 2022a). Yet, these models are often computationally demanding, limiting their applicability in real-time traffic operations (Huang et al., 2020). Moreover, they are typically trained on historical data with relatively stable conditions and lack the adaptability to perform robustly during irregular scenarios like city-wide events or emergencies (Shang et al., 2024). This points to an unmet need for imputation frameworks that can efficiently handle structured, non-random missingness in dynamic urban contexts.

At the same time, while recent studies have explored integrating exogenous factors such as weather or incident reports into prediction models (Zhang et al., 2025; Lu et al., 2025), there remains a paucity of work that systematically incorporates structured event metadata, such as event type, attendance, timing, and venue location, into the imputation process. Social events provide a rich and predictable source of contextual information that, if integrated properly, could significantly enhance the robustness of imputation under disruptive conditions. For example, Essien et al. (Essien et al., 2021) showed that fusing Twitter-derived event indicators with traffic sensor data improved short-term forecasting, while Yao and Qian (Yao and Qian, 2021b) demonstrated that social media activity patterns could be predictive of next-day morning congestion. Despite these promising findings, current models often rely on noisy, unstructured social data or are designed for forecasting rather than directly addressing the imputation problem.

Another important gap concerns uncertainty quantification. Most existing deep learning-based imputation models provide point estimates without accompanying measures of confidence or variability (Liang et al., 2021; Ma et al., 2019). This limits their utility in operational settings, where transportation managers require not only best-guess estimates but also an understanding of the associated risks, particularly during high-stakes events or emergencies. Further complicating matters, many of these models are tailored to specific cities or datasets and have not been thoroughly validated for generalizability across diverse urban contexts (Kong et al., 2022).

These limitations point to a clear and pressing research question: *How can we design an efficient, scalable, and context-aware traffic data imputation framework that integrates structured social event information, remains computationally feasible for real-time applications, and provides reliable uncertainty estimates to support operational decision-making?* This dissertation addresses this question by proposing a novel hybrid imputation model that combines ensemble learning with spatiotemporal deep learning techniques, enriched by event metadata, to improve imputation accuracy under both routine and event-driven conditions. The model is designed to balance predictive power with computational efficiency, ensuring that it can be deployed in real-world ITS environments without sacrificing interpretability or scalability. The next section will detail the specific objectives that guide this research.

1.6 Research Objectives

The primary objective of this dissertation is to enhance the accuracy and efficiency of traffic state estimation and short-term traffic speed prediction under conditions of missing data and non-recurring disruptions caused by social events. To achieve this aim, the dissertation proposes a unified framework that integrates advanced imputation techniques with event-aware spatiotemporal prediction models. By addressing both prediction and imputation challenges, the research seeks to improve the operational reliability of Intelligent Transportation Systems (ITS) during high-variability periods and to provide decision support for real-time urban traffic management. The proposed methods are designed to function under realistic conditions where incomplete datasets and contextual disturbances pose significant modeling challenges.

In alignment with this overarching aim, the research objectives have been defined and are presented below in the same order as the chapters in which they are addressed:

1. **Develop an event-aware prediction model.** The first objective is to construct a traffic prediction model that explicitly incorporates social event features such as event type, location, duration, and attendance. By embedding these features into deep spatiotemporal architectures, the model aims to capture complex traffic patterns and dynamic disruptions that emerge during non-recurring events. *[Addressed in Chapter 2 and Chapter 3]*

2. **Evaluate prediction robustness and generalizability.** The second objective is to systematically assess the performance of the proposed prediction model under diverse event conditions and across different urban topologies. This involves evaluating the role of proximity, temporal alignment, and event scale in shaping prediction accuracy, as well as benchmarking the proposed method against baseline models that do not incorporate event information. The goal is to demonstrate that event-aware designs provide superior adaptability across cities, event types, and scales of disruption. *[Addressed in Chapter 2 and Chapter 3]*

3. **Investigate limitations of existing imputation methods.** The third objective is to examine the weaknesses of current traffic data imputation approaches under realistic missingness scenarios. Particular emphasis is placed on correlated disruptions, such as those induced by social events or network-level sensor failures. By benchmarking statistical, machine learning, and deep learning methods, this analysis highlights gaps in accuracy, scalability, and context-awareness that motivate the need for a more advanced solution. *[Addressed in Chapter 4]*

4. **Develop an efficient imputation framework.** The fourth objective is to design an imputation framework capable of handling large-scale, multi-segment traffic datasets with variable missing rates. The framework seeks to balance computational efficiency with accuracy, ensuring reliable performance in real-time deployments. This contribution supports operational resilience in ITS by enabling more

robust monitoring and forecasting, even in sparse or incomplete sensing environments. *[Addressed in Chapter 4]*

1.7 Dissertation Organization

This dissertation follows a sandwich format and is organized into five comprehensive chapters, each designed to contribute incrementally to the central research goal of developing event-aware, scalable, and reliable frameworks for traffic data imputation and speed prediction in urban networks. The structure reflects the natural progression from problem formulation to methodological development, experimental validation, and synthesis of contributions. The three middle chapters are structured as standalone research papers, one of which has been published and two that are currently under peer review. These chapters are framed by an introductory and a concluding chapter that provide narrative and analytical cohesion to the research.

Chapter 1 provides the overall background, context, and justification for the dissertation. It opens with a discussion of the increasing reliance of urban mobility systems on data-driven technologies and highlights the dual challenges of missing traffic data and disruption-prone conditions such as social events. The chapter includes a thorough literature review across three key domains: traffic speed prediction, missing traffic data imputation, and social event modeling. It identifies critical limitations in existing models, especially in their ability to operate under irregular data availability and non-recurring disruptions. Based on these gaps, the chapter formulates the research problem, defines the primary research questions, and sets forth the specific research objectives that guide the remainder of the thesis.

Chapter 2 is based on the published study titled *Enhancing Traffic Speed Prediction Accuracy: The Multialgorithmic Ensemble Model With Spatiotemporal Feature Engineering.* This chapter proposes a hybrid framework called the Multialgorithmic Ensemble Model (MAEM), which integrates graph neural networks (GNNs), long short-term memory networks (LSTM), and bidirectional gated recurrent units (Bi-GRU). These components work in concert with spatiotemporal feature engineering techniques to capture complex dependencies in traffic data. The model's adaptive attention layer dynamically identifies the most influential temporal patterns. Evaluated on a one-year dataset from Hamilton, Ontario, MAEM achieved a MAPE of 3.18% and RMSE of 2.85 km/h, outperforming benchmark models including Graph WaveNet and attention-based transformers. The model also demonstrated strong robustness during peak-hour congestion, highlighting its applicability in real-time traffic operations.

Chapter 3 contains the second research article, currently under review at an ASCE journal. This study advances traffic prediction modeling by proposing the Event-Aware Long Short-Term Memory (EA-LSTM) architecture. The model integrates probe vehicle data and event features in a unified structure comprising Graph Convolutional Networks, Bidirectional LSTM layers, self-attention modules, and hierarchical sequence modeling. This chapter also introduces a novel event-feature encoding pipeline and offers a thorough empirical evaluation across various event scenarios and network conditions. The findings show that the EA-LSTM model generalizes well across different types of social events and scales of disruption, addressing limitations in both conventional LSTM-based models and earlier shallow machine learning approaches.

Chapter 4 presents the third paper, also under review, which further refines the interplay between imputation quality and prediction performance. It introduces a two-stage architecture designed to handle heterogeneous missingness patterns and variable spatial coverage. The first stage focuses on dynamically imputing incomplete data using a suite of context-aware techniques, while the second stage performs forecasting using event-augmented deep networks. This chapter also includes a comparative benchmarking study of traditional and deep imputation methods under event-driven scenarios. The research reveals that incorporating social event metadata significantly enhances the accuracy of both imputation and prediction, especially in congested or sensor-sparse areas.

Chapter 5 synthesizes the overall contributions of the dissertation and reflects on their broader implications for traffic engineering and smart city operations. It revisits the research objectives and maps them to the findings from the three core studies. This chapter discusses practical implications, such as how the developed models can be integrated into real-time Advanced Traffic Management Systems (ATMS). Additionally, it outlines key limitations of the current work, such as computational scalability and data availability, and proposes directions for future research. These include expanding the framework to include multimodal data sources like transit schedules and weather information, and exploring deployment pathways through cloud-based infrastructures and edge computing.

Together, the five chapters provide a coherent, cumulative research narrative that moves from conceptual motivation to technical realization and applied validation. This structure ensures both the scholarly contribution of each individual paper and the thematic unity of the dissertation as a whole.

Bibliography

Bae, B., Kim, H., Lim, H., Liu, H. X., Han, L. D., and Freeze, P. B. (2018). Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transportation Research Part C: Emerging Technologies*, 88:124–139.

Cao, P., Dai, F., Liu, G., Yang, J., and Huang, B. (2021). A survey of traffic prediction based on deep neural network: Data, methods and challenges. In *Proceedings of the International Conference on Cloud Computing*, pages 17–29. Springer.

Chang, X., Sun, J., Liu, Y., and Li, Q. (2022). Spatiotemporal traffic anomaly detection using weibo social media data. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24122–24132.

Chen, P., Jiang, X., and Yang, R. (2024). Multi-step freeway traffic speed prediction with spatiotemporal graph neural networks. *Journal of Transportation Engineering, Part A: Systems*, 150(3):04024020.

Chen, X., Lei, M., Saunier, N., and Sun, L. (2022a). Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12301–12310.

Chen, X., Sun, Y., and Xu, H. (2022b). Traffic data imputation using low-rank tensor completion with spatial-temporal consistency. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):17025–17036.

Cools, M., Moons, E., and Wets, G. (2007). Investigating the effect of holidays on daily traffic counts: A time series approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2019:25–32.

Dabiri, M. and Heaslip, K. (2019). Developing a twitter-based traffic event detection model using deep learning architectures. In *Transportation Research Board 98th annual meeting.*

Deng, W., Zhang, L., and Liu, Y. (2022). Detecting spatiotemporal anomalies in traffic data with graph convolutional adversarial networks. *Transportation Research Part C: Emerging Technologies*, 135:103498.

Du, Y., Qin, X., Jia, Z., Yu, K., and Lin, M. (2021). Traffic speed prediction based on heterogeneous graph attention residual time series convolutional networks. *AI*, 2(4):650–661.

Duan, Y., Lv, Y., Liu, Y., and Wang, F.-Y. (2016). An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 72:168–181.

Essien, O., Sharma, K., and Farahmand, A. (2021). Event-aware traffic flow prediction using multi-source data fusion. *Transportation Research Record*, 2675(9):586–599.

Harrou, F., Zeroual, A., Kadri, F., and Sun, Y. (2024). Enhancing road traffic flow prediction with improved deep learning using wavelet transforms. *Results in Engineering*, page 102342.

Huang, S., Xu, Z., Huang, Z., and Zhang, J. (2020). Tsdi-gan: Generative adversarial network for time series data imputation. *IEEE Access*, 8:150567–150578.

Kong, X., Zhao, J., and Chen, Y. (2022). Survey on missing traffic data imputation: Methods and applications. *Transportation Research Part C: Emerging Technologies*, 138:103603.

Liang, Y., Ke, J., Zheng, H., and Li, Z. (2021). Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):6992–7004.

Lu, B., Miao, Q., Liu, Y., Tamir, T., Zhao, H., Zhang, X., Lv, Y., and Wang, F.-Y. (2025). A diffusion model for traffic data imputation. *IEEE/CAA Journal of Automatica Sinica*, 12(3):606–617. Introduces an implicit–explicit diffusion method for missing traffic data.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2014). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

Ma, X., Li, X., and Ding, C. (2020). Traffic data imputation during large-scale events using spatiotemporal models. *Transportation Research Part C: Emerging Technologies*, 114:159–175.

Ma, X., Tao, Z., Wang, Y., and Yu, H. (2019). Long short-term memory recurrent variational autoencoder for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 104:66–81.

Ni, D. and Leonard, J. D. (2005). Markov chain monte carlo multiple imputation using bayesian networks for incomplete its data. *Transportation Research Record*, (1935):57–67.

Okukubo, T., Bando, Y., and Onishi, M. (2022). Traffic prediction during large-scale events based on pattern-aware regression. *Journal of Information Processing*, 30:42–51.

Razali, N. A. M., Shamsaimon, N., Ishak, K. K., Ramli, S., Amran, M. F. M., and Sukardi, S. (2021). Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning. *Journal of Big Data*, 8:1–25.

Shang, Q., Tang, Y., and Yin, L. (2024). A hybrid model for missing traffic flow data imputation based on clustering and attention mechanism optimizing lstm and adaboost. *Scientific Reports*, 14:26473.

Smith, B. L., Scherer, W. T., and Conklin, J. H. (2003). Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record*, (1836):132–142.

Song, Y., Luo, R., Zhou, T., Zhou, C., and Su, R. (2024a). Graph attention informer for long-term traffic flow prediction under the impact of sports events. *Sensors*, 24(15):4796.

Song, Y., Luo, R., Zhou, T., Zhou, C., and Su, R. (2024b). Graph attention informer for long-term traffic flow prediction under the impact of sports events. *Sensors*, 24(15):4796.

Stathopoulos, A. and Karlaftis, M. G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2):121–135.

Tan, H., Wu, Y., Jin, L., Shen, Z., and Wang, J. (2013). A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, 28:15–27.

Tang, J., Liu, F., Zou, Y., and Zhang, W. (2017). A hybrid approach to integrate fuzzy c-means based imputation method with random forest for traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 77:21–34.

Tang, J., Zhang, X., Yu, T., and Liu, F. (2021). Missing traffic data imputation considering approximate intervals: A hybrid structure integrating adaptive network-based fuzzy inference system and fuzzy rough set. *Physica A: Statistical Mechanics and its Applications*, 573:125948.

Tao, R., Duan, K., Dong, Z., and Yang, X. (2022a). Traffic detection and forecasting from social media data using a hybrid albert–bilstm–crf framework. In *Proceedings of the 16th International Conference on Knowledge Engineering and Ontology Development (KEOD)*, pages 236–243, Wroclaw, Poland.

Tao, X., Liu, D., Zhao, R., and Zhang, X. (2022b). Traffic detection and forecasting from social media data using a hybrid albert–bilstm–crf framework. *International journal of knowledge engineering and ontology development*, 16(4):1–19.

van Lint, H. and van Zuylen, H. J. (2005). Travel time unreliability on freeways: A review of methods and a data analysis evaluation. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems*, pages 34–39. IEEE.

Wang, X., Zhao, Z., Wang, R., and Xu, Y. (2025). Event-aware analysis of cross-city visitor flows using large language models and social media data. *arXiv preprint*, arXiv:2505.03847.

Xing, Z., Huang, M., and Peng, D. (2023). Overview of machine learning-based traffic flow prediction. *Digital Transportation and Safety*, 2(3):164–175.

Yao, H., Tang, J., Wei, H., Zheng, G., and Li, Z. (2018). Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):5668–5675.

Yao, W. and Qian, S. (2021a). From twitter to traffic predictor: Next-day morning traffic prediction using social media data. *Transportation Research Part C: Emerging Technologies*, 124:102938.

Yao, W. and Qian, S. (2021b). From twitter to traffic predictor: Next-day morning traffic prediction using social media data. *Transportation Research Part C: Emerging Technologies*, 124:102938.

Ye, Y., Zhang, S., and Yu, J. J. Q. (2021). Traffic data imputation with ensemble convolutional autoencoder. In *Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN.

Yin, Q., Sinha, S., Zheng, H., and Guan, W. (2023). Short-term traffic flow prediction considering nonrecurrent congestion using a hybrid lstm–cnn approach. *Journal of Transportation Engineering, Part A: Systems*, 149(10):04023100.

Yu, B., Yin, H., and 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)*, pages 3634–3640.

Zhang, W., Li, X., and Chen, Y. (2021). A comprehensive review of missing data imputation for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):3895–3914.

Zhang, Y., Kong, X., Zhou, W., Liu, J., Fu, Y., and Shen, G. (2024). A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19252–19275. PP(99):1–24.

Zhang, Y., Kong, X., Zhou, W., Liu, J., Fu, Y., and Shen, G. (2025). A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*. Survey of statistical, machine learning, and deep learning imputation methods.

Zhang, Z., Yang, H., and Yang, X. (2023). A transfer learning–based lstm for traffic flow prediction with missing data. *Journal of Transportation Engineering, Part A: Systems*, 149(10):04023095.

Zhao, J., Kong, X., and Chen, Y. (2021). Graph neural network-based traffic data imputation with spatial-temporal dependencies. *Transportation Research Part C: Emerging Technologies*, 128:103114.

Zheng, Z., Liu, Z., and van Zuylen, H. J. (2015). Imputation of missing traffic data based on spatial-temporal correlations. *Promet – Traffic and Transportation*, 27(2):157–167.

Zhou, W., Liu, Q., and Wang, K. (2023). Impact of special events on urban traffic: An imputation and prediction study using deep learning. *Journal of Advanced Transportation*, 2023:1–16.

Chapter 2

Enhancing Traffic Speed Prediction Accuracy: The Multi-Algorithmic Ensemble Model with Spatiotemporal Feature Engineering

The content of this chapter is the manuscript text for publication under the following citation:

> Ali Ardestani, Hao Yang, and Saiedeh Razavi. Enhancing traffic speed prediction ac- curacy: The multialgorithmic ensemble model with spatiotemporal feature engineer- ing. Journal of Advanced Transportation, 2025(1):9941856, 2025. doi: 10.1155/atr/ 9941856.

> Ardestani A. designed the research framework, conducted data preprocessing and analysis, implemented the ensemble modeling approach, and drafted the manuscript. Yang H. contributed to the conceptual development, supervised the methodology, and provided critical review and revisions throughout the research. Razavi S. contributed to the development of the modeling framework and structure, and provided methodological and editorial guidance. All authors reviewed and approved the final version of the manuscript.

Traffic Speed Prediction Accuracy: The Multialgorithmic Ensemble Model With Spatiotemporal Feature Engineering

Ali Ardestani[1], Hao Yang[1,*], Saiedeh Razavi[1]

[1] Department of Civil Engineering, McMaster University, Hamilton, ON, Canada
*Corresponding author: haoyang@mcmaster.ca

Abstract

Accurate traffic speed prediction is crucial for efficient traffic management and planning in urban areas. Traditional traffic prediction models often fall short due to their inability to capture the complex and dynamic nature of traffic flow. There is a need for more advanced models that can effectively handle dynamic traffic conditions. This study introduces the multialgorithmic ensemble model (MAEM), a novel framework designed to improve traffic speed prediction accuracy by integrating graph neural networks (GNNs), bidirectional gated recurrent units (Bi-GRUs), and long short-term memory (LSTM) networks, to effectively analyze the spatiotemporal characteristics of the traffic network. The methodology involves constructing a virtual graph based on road segment correlations and applying a combination of spatial and temporal feature extraction techniques. The model is further enhanced with an attention mechanism to focus on critical time intervals. The dataset used for this study consists of one-year aggregated probe vehicle traffic data of 4788 road segments in the City of Hamilton, Ontario. The results demonstrate significant performance, achieving the mean absolute percentage error (MAPE) of 3.5% and root-mean-square error (RMSE) of 2.4 km/h, indicating the potential of the proposed framework to significantly enhance traffic speed prediction accuracy and provide a reliable tool for urban traffic management and planning.

Keywords: *Traffic Prediction, Multi-Algorithmic Ensemble Model (MAEM), Long Short-Term Memory (LSTM), Spatiotemporal Feature Engineering (STFE), Probe Vehicle Data*

2.1 Introduction

Traffic congestion remains a significant challenge in urban areas, leading to a multitude of adverse effects, including increased travel times, fuel consumption, and environmental pollution. The ability to predict traffic speed accurately is crucial for mitigating these issues and enhancing the overall efficiency of transportation networks (Lv et al., 2014). Advanced traffic speed prediction models can provide valuable insights that enable proactive traffic management, route optimization, and informed decision-making for travelers and transportation authorities alike (Zhang et al., 2019; Guo et al., 2019; Rasaizadi et al., 2021a).

Accurate traffic speed predictions can quantitatively improve traffic congestion by enabling dynamic traffic signal control, real-time traffic rerouting, and better management of traffic incidents. These measures can significantly reduce travel delays and improve the reliability of travel times (Ma et al., 2015; Ghodsi et al., 2022). Moreover, by alleviating congestion, these models contribute to reducing vehicle emissions and fuel consumption, thereby benefiting the environment (Zhou et al., 2022b; Yang et al., 2020). The social impacts are also substantial, as smoother traffic flow enhances the quality of life for commuters by reducing stress and time spent in traffic (Gu et al., 2020; Kong et al., 2019). In terms of safety, accurate traffic speed predictions can play a vital role in preventing accidents. By identifying potential traffic bottlenecks and hazardous conditions in advance, traffic management systems can implement preventive measures to enhance road safety (Zhang et al., 2017; Ma et al., 2021). Furthermore, predictive models can assist in the deployment of emergency services more efficiently, ensuring quicker response times during traffic incidents (Lv et al., 2014).

Despite these benefits, developing accurate traffic speed prediction models poses several challenges. Traffic conditions are influenced by numerous dynamic factors, including weather conditions, road works, special events, and accidents. Traditional prediction methods, such as historical averaging and basic statistical models, often fall short of capturing the complex, non-linear nature of traffic flow (Sun et al., 2003; Williams and Hoel, 2003). With the abundance of available traffic data, there is a growing interest in leveraging advanced machine learning and deep learning techniques to enhance prediction accuracy (Yao et al., 2017; Shaygan et al., 2022). Conventional machine learning models, such as support vector regression (SVR) and k-nearest neighbor (KNN), have shown promise but often require extensive feature engineering and may not fully capture temporal dependencies in the data (Gu et al., 2019). On the other hand, deep learning approaches, including long short-term memory (LSTM) networks and convolutional neural networks (CNNs), are capable of learning complex patterns and dependencies directly from raw traffic data (Zhang et al., 2017). These models have demonstrated superior performance in traffic speed prediction tasks by effectively modeling both spatial and temporal correlations (Zhang et al., 2017). Recent advancements in graph neural networks (GNNs) have further enhanced traffic speed prediction by considering the spatial structure of road networks (Lee and Rhee, 2022). Spatiotemporal graph convolutional networks (STGCNs) integrate graph convolutions with temporal modeling techniques, providing a robust framework for capturing the intricate dependencies in traffic data (Rahmani et al., 2023).

In this study, we propose a Multi-Algorithmic Ensemble Model (MAEM) that combines the strengths of various advanced techniques, including GNN networks, bidirectional gated recurrent unit networks, and LSTM layers which leverages spatiotemporal feature engineering to enhance prediction accuracy and incorporates adaptive learning mechanisms to dynamically respond to changing traffic patterns and conditions. By evaluating the performance of the MAEM model using real-world traffic speed datasets, we aim to demonstrate its effectiveness in short-term traffic speed forecasting. The remainder of this paper is structured as follows. Section II provides a detailed review of the relevant literature, highlighting key advancements and gaps in traffic prediction methodologies. Section III outlines the proposed methodology, including the dataset description, problem formulation, and the framework of the Multi-Algorithmic Ensemble Model (MAEM). Section IV presents the experimental results, evaluating the effectiveness of the proposed model through various performance metrics and scenarios. Finally, Section V concludes the study by summarizing the key findings and discussing future research directions.

2.2 Literature Review

The field of traffic speed prediction has seen significant advancements with the introduction of various machine learning and deep learning techniques. Traditional methods, such as the historical average (HA) and autoregressive integrated moving average (ARIMA), have been widely used for traffic prediction but often fall short due to their inability to capture nonlinear dependencies in traffic data (Sun et al., 2003; Williams and

Hoel, 2003). These methods typically rely on historical data and simple statistical relationships, which limits their effectiveness in dynamic and complex traffic environments (Van Lint and Van Hinsbergen, 2012).

Conventional machine learning approaches, including support vector regression (SVR) and k-nearest neighbour (KNN), have been applied to traffic speed prediction with varying degrees of success. These methods typically require handcrafted features and may not fully utilize the large-scale traffic data available (Fei et al., 2011; Lian, 2024). For instance, SVR has been used to model traffic speed and volume, but its performance is often dependent on the quality and relevance of the features selected (Zhang and Xie, 2007). KNN, while simple and intuitive, can struggle with high-dimensional data and may not effectively capture temporal patterns in traffic flow (Rahman, 2020; Akhtar and Moridpour, 2021).

With the advent of deep learning, models such as long short-term memory (LSTM) networks and convolutional neural networks (CNNs) have been employed to capture temporal and spatial dependencies, respectively (Ma et al., 2015; Zhang et al., 2017). LSTM networks, which are capable of learning long-term dependencies in sequential data, have shown promise in traffic prediction tasks by effectively modelling temporal correlations (Lv et al., 2014). CNNs, on the other hand, excel at capturing spatial features and have been used to model the spatial structure of road networks (Ma et al., 2017). Recent advancements include combining LSTM and CNN to leverage both spatial and temporal information, leading to improved prediction accuracy (Yu et al., 2017).

Recent studies have explored the use of graph neural networks (GNNs) for traffic prediction, leveraging the spatial structure of road networks to improve prediction accuracy (Zhang et al., 2019). Spatiotemporal graph convolutional networks (STGCNs) combine graph convolutions with time convolutions or recurrent neural networks to effectively capture both spatial and temporal features (Yu et al., 2017). However, these models often rely on predefined graphs, which may not accurately represent the dynamic nature of traffic networks (Deng et al., 2022; Guo et al., 2021). This limitation highlights the need for adaptive methods that can dynamically learn and adjust to changes in the traffic network structure.

To address these limitations, adaptive graph learning methods have been proposed. Models such as Graph WaveNet and Adaptive Graph Convolutional Recurrent Network (AGCRN) learn the graph structure from data without prior knowledge, achieving comparable performance to models based on predefined graphs (Sun et al., 2021; Wang et al., 2022). These adaptive methods are particularly useful in dynamically changing environments where the traffic network's structure can vary over time. For instance, Graph WaveNet employs a data-driven approach to construct the graph, allowing it to adapt to different traffic scenarios (Wang et al., 2020). Another promising direction is the integration of multiple models to form ensemble learning frameworks. For example, hybrid models that combine the strengths of different algorithms can improve overall prediction performance (Rasaizadi et al., 2021b). These ensemble methods can include various machine learning models or combine machine learning with deep learning approaches to better capture the complexities of traffic data (Kong et al., 2019). Recent studies have further refined ensemble learning approaches by incorporating techniques such as attention mechanisms and transfer learning, which have demonstrated significant improvements in prediction accuracy and model robustness (Zhang et al., 2023; Zhou et al., 2022a). These methods leverage real-time traffic data to make adaptive decisions, enhancing the overall efficiency of urban traffic management systems.

In addition to these techniques, attention mechanisms have been increasingly utilized in traffic prediction models to focus on relevant features and improve prediction accuracy (Zhao et al., 2022). The integration of attention mechanisms in spatiotemporal models has demonstrated significant improvements in capturing complex traffic patterns and adapting to changing traffic conditions [28]. Furthermore, recent studies have focused on the use of transfer learning to enhance traffic prediction models. Transfer learning allows models to leverage knowledge from related tasks or domains, improving prediction performance and reducing the need for large amounts of labeled data (Razali et al., 2021). This approach has been particularly effective in scenarios where traffic data is sparse or incomplete, enabling more accurate and reliable predictions (Zhang et al., 2023).

In recent years, significant progress has been made in deep learning techniques for traffic prediction, a critical component of intelligent transportation systems (ITS). For instance, Lohrasbinasab et al. (2021) explored the shift from traditional statistical techniques to machine learning-based models for network traffic prediction, highlighting the strengths of machine learning in handling big data and addressing network inefficiencies (Lohrasbinasab et al., 2022). Similarly, Sroczyński and Czyżewski (2023) compared machine learning methods like LSTM and GRU to traditional traffic simulation models, demonstrating that neural networks can operate in real-time and outperform classical simulation models in terms of speed and accuracy (Sroczyski and Czyewski, 2023). Xu et al. (2024) proposed a spatiotemporal convolutional model for traffic flow prediction that leverages graph GNNs to capture complex spatial and temporal dependencies, achieving notable improvements in predictive performance over earlier models (Xu et al., 2024).

Additionally, Cui et al. (2023) provided a comprehensive review of spatiotemporal correlation modeling, identifying GNNs and attention mechanisms as critical for enhancing the predictive capabilities of traffic state models (Cui et al., 2023). These studies highlight a growing trend toward hybrid models that integrate various neural network architectures to enhance both spatial and temporal feature extraction, particularly for highly dynamic traffic environments. Despite these advancements, challenges such as optimizing model architectures and addressing external factors such as traffic events or signal effects remain areas for further exploration. As such, the literature indicates a clear movement toward more sophisticated and computationally efficient models, yet the complexities of real-world traffic conditions still pose significant hurdles to achieving fully reliable and scalable solutions.

These recent studies underscore a growing trend toward hybrid models that integrate various neural network architectures to enhance spatial and temporal feature extraction, particularly for highly dynamic traffic environments. However, despite these advancements, limitations persist. Many of these models are computationally intensive, which can hinder real-time deployment, and they often struggle to adapt to rapid changes in traffic flow caused by external factors like social events or road incidents. Furthermore, the complexity of real-world traffic conditions, including variable weather effects and signal changes, remains a challenge, as most models are trained under controlled conditions that may not fully generalize to real-world scenarios. Thus, while recent models have made strides in accuracy, they frequently lack the adaptability and efficiency necessary for practical, large-scale applications. In summary, while significant progress has been made in traffic speed prediction, several critical research gaps remain. Existing models often struggle to effectively integrate spatiotemporal features, adapt to dynamic traffic conditions, and scale to large urban networks with diverse traffic patterns. The proposed MAEM model seeks to address these challenges by combining advanced machine learning techniques, spatiotemporal feature engineering, and adaptive learning mechanisms. By overcoming the limitations of traditional models and leveraging the latest advancements in machine learning and deep learning, the MAEM model offers a more accurate, scalable, and resilient solution for traffic speed prediction in complex urban environments. The detailed objectives of this study are listed below in conjunction with a summary of the methods.

1. Construct a virtual graph that incorporates road segments with high relevance, measured by the Pearson correlation coefficient. This approach considers both physically connected road segments and those without direct physical connections.

2. Develop an ensemble model that integrates GNNs, LSTM, bidirectional gated recurrent unit (Bi-GRU) networks, and fully connected networks with an attention mechanism in an end-to-end manner.

3. Evaluate the proposed model's performance using a real-world urban traffic speed dataset.

The novelty of the Multi-Algorithmic Ensemble Model (MAEM) lies in its integrated approach to capturing both spatial and temporal dependencies through a combination of GNN, LSTM, and Bi-GRU, which work together within a unified framework. MAEM utilizes GNN for spatial feature extraction across road segments, LSTM to retain spatial dependencies over longer sequences, and Bi-GRU to process temporal information in both forward and backward directions. This unique combination allows MAEM to model the complex, interdependent structure of traffic data more effectively than prior models in the literature, which typically focus on either spatial or temporal aspects in isolation.

Furthermore, the inclusion of an attention mechanism enables MAEM to adapt to dynamic traffic conditions by focusing on critical time intervals, enhancing its real-time predictive capability. This adaptive architecture positions MAEM as an advancement over previous models, which often lack the capacity to handle rapid variations in urban traffic. By integrating these advanced neural architectures, MAEM offers a robust, flexible solution for traffic speed prediction, particularly in dynamic urban environments.

2.3 Methodology

This research proposes an approach to unravel the complexities of traffic prediction, using a case study within the City of Hamilton, Ontario. This section explains the development of our methodology, including a detailed description of the utilized dataset, an articulation of the problem statement, the framework of our proposed model—Multi-Algorithmic Ensemble Model (MAEM), and the criteria for evaluating its performance. The proposed approach is designed to improve the precision and adaptability of traffic predictions, addressing the shortcomings of existing methods and paving the way for transformative advancements in real-world applications.

2.3.1 Problem Statement and Model Framework

One of the critical components of intelligent traffic management is the ability to predict traffic speed in real-time accurately (Xing et al., 2023). The objective of this research is to employ machine learning algorithms, fine-tuned through rigorous validation processes, to predict traffic speed across various types of roads. This predictive model aims to serve as a cornerstone for real-time intelligent traffic management systems, reducing congestion, and minimizing environmental impacts. The proposed Multi-Algorithmic Ensemble Model (MAEM) employs a combination of GNN, LSTM, Bi-GRU, and attention mechanisms to capture complex spatial and temporal dependencies in traffic data. Below, we describe the distinct roles of each component and how they integrate to form a comprehensive spatiotemporal prediction model. The MAEM framework is illustrated in Figure 2.1 and comprises five key components:

1- GNN and Adjacency Matrix Construction: The first step in MAEM is establishing spatial relationships between road segments by constructing an adjacency matrix A using a Graph Neural Network (GNN). This matrix is built based on the Pearson correlation between the traffic speeds of connected road segments. Specifically, each entry A_{ij} in the matrix reflects the correlation coefficient between segment i and segment j, capturing the degree of influence one segment has on another. Pearson correlation was chosen for constructing A due to its ability to provide a computationally efficient approach to identifying both direct and indirect correlations between road segments. Its simplicity in calculating correlations allows for real-time updates to the adjacency matrix, making it a practical choice for traffic prediction tasks that require fast processing times.

This adjacency matrix A serves two roles in MAEM:

- *Structural Encoding:* The adjacency matrix A is fed into the GNN layer, where it helps encode spatial structure by weighting the influence of each segment on its neighbors. This encoding aids in forming a holistic spatial representation.

- *Guiding LSTM Processing:* The GNN-derived adjacency matrix A is also used to guide the sequential processing within the LSTM layer. By providing a weighted spatial structure, A determines the relative importance of each segment, influencing the hidden states as traffic flows across segments.

While GNN effectively captures broad spatial correlations, the adjacency matrix A is essential in transforming these relationships into a format that LSTM can process sequentially, maintaining the inter-segment dependencies crucial for real-time prediction.

2- LSTM for Sequential Spatial Feature Extraction: The Long Short-Term Memory (LSTM) network is utilized in MAEM to capture spatial dependencies by processing the adjacency-informed segment sequence. LSTM's architecture, comprising memory cells and gated mechanisms, is particularly suited for handling ordered dependencies, making it ideal for modeling traffic data as it flows sequentially across connected road segments. Key technical aspects include:

- *Spatial Sequencing:* With traffic data structured as a directional sequence, LSTM processes each segment's data point by point, carrying forward dependencies across segments. Each hidden state h_t at step t reflects temporal relationships from previous segments in the sequence, informed by A.

- *Information Control:* LSTM's forget and update gates enable selective retention of important spatial information while discarding less relevant features. This controlled flow of information allows LSTM to encode both immediate and distal spatial influences, maintaining spatial coherence over extended segment sequences.

In this way, the LSTM layer captures sequential spatial dependencies, treating each road segment as part of a chain-like structure. The adjacency matrix A influences this spatial structure, enhancing the LSTM's ability to retain and prioritize spatially relevant information.

3- Bi-GRU for Temporal Feature Extraction: The Bidirectional Gated Recurrent Unit (Bi-GRU) layer is employed in MAEM for capturing bidirectional temporal dependencies. The Bi-GRU operates along the time axis, modeling how past and future traffic states affect current conditions. This approach leverages GRU's simplified gate structure, reducing computational complexity compared to LSTM, which makes Bi-GRU suitable for real-time applications.

Bi-GRU computes hidden states in both forward and backward directions, providing a comprehensive temporal view at each time step. This bi-directionality allows the model to consider the impact of both past (e.g., peak congestion) and anticipated future conditions on present traffic states. By using Bi-GRU for temporal dependencies, MAEM captures both historical patterns and forward-looking temporal information, allowing the model to adapt to rapid changes in traffic flow, which is essential in dynamic environments.

4- Attention Mechanism: The attention mechanism in MAEM further enhances adaptability by enabling the model to selectively focus on key temporal intervals. This layer assigns weights to different time steps, prioritizing those that carry more predictive significance (e.g., peak hours, sudden traffic shifts). This selective weighting ensures that MAEM can dynamically adjust its focus based on the evolving traffic context, which is crucial for maintaining accuracy in real-time predictions.

Technical aspects of the attention mechanism include:

- *Weight Calculation:* The attention mechanism computes weights based on the relevance of each time step's features, dynamically adjusting as traffic patterns change. Higher weights are assigned to critical time intervals, such as periods of high congestion.

- *Enhanced Temporal Focus:* By concentrating on significant time points, the attention mechanism refines the temporal features extracted by Bi-GRU, allowing MAEM to optimize predictions in high-variability conditions. This process minimizes the influence of less relevant intervals, improving the robustness of predictions.

5- Final Output Calculation and Validation: The final output of the MAEM model is generated by integrating spatial and temporal features extracted through the GNN, LSTM, Bi-GRU, and attention layers. First, the GNN and LSTM layers collaboratively capture spatial dependencies across road segments, structuring them sequentially to reflect inter-segment relationships. Meanwhile, Bi-GRU processes the temporal aspect by creating bidirectional hidden states that incorporate both past and anticipated traffic conditions, enhancing the model's temporal sensitivity to fluctuations in traffic patterns.

After these spatial and temporal features are extracted, the attention mechanism is applied to the Bi-GRU outputs, weighting significant time intervals and enabling the model to dynamically focus on periods with high predictive importance. The weighted temporal features are then combined with the spatial features and passed through a dense layer, which synthesizes these spatiotemporal patterns into a single traffic speed prediction. This final layer generates a refined output that reflects both the immediate and evolving traffic conditions, allowing for accurate, real-time forecasting suited to complex urban environments.

In summary, MAEM's architecture is designed to leverage the unique strengths of each component. The GNN and adjacency matrix establish spatial structure, which the LSTM layer uses to sequentially capture spatial dependencies. Bi-GRU then captures bidirectional temporal dependencies, further refined by the attention mechanism to prioritize critical time intervals. This layered approach enables MAEM to handle the spatiotemporal complexities of traffic data in a real-time, computationally efficient manner.

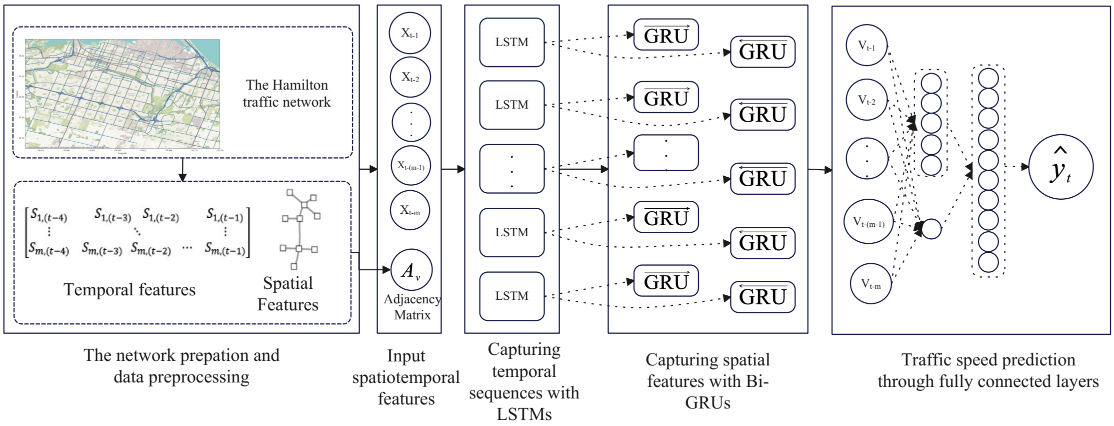
The Hamilton traffic network
$S_{1,(t-4)}$ $S_{1,(t-3)}$ $S_{1,(t-2)}$ $S_{1,(t-1)}$
$S_{m,(t-4)}$ $S_{m,(t-3)}$ $S_{m,(t-2)}$ ⋯ $S_{m,(t-1)}$
Temporal features
Spatial Features
The network prepation and data preprocessing
X_{t-1}
X_{t-2}
$X_{t-(m-1)}$
X_{t-m}
A_v
Adjacency Matrix
Input spatiotemporal features
LSTM
LSTM
LSTM
LSTM
Capturing temporal sequences with LSTMs
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
Capturing spatial features with Bi-GRUs
V_{t-1}
V_{t-2}
$V_{t-(m-1)}$
V_{t-m}
$\hat{y}_t$
Traffic speed prediction through fully connected layers

Figure 2.1: Architecture of the proposed MAEM model for traffic speed prediction

2.3.2 Spatiotemporal Feature Engineering

Spatiotemporal Feature Engineering (STFE) is integral to our traffic speed prediction model, leveraging the intricate spatial and temporal relationships within the traffic network. STFE involves the process of extracting and transforming data features that capture both spatial and temporal dynamics of traffic flow, enhancing the model's ability to predict traffic speed accurately. By utilizing techniques like graph neural networks (GNNs), we model the road network as a graph, where intersections are nodes, and road segments are edges. This helps in understanding the influence of neighboring road segments on the target segment's traffic flow. Temporal Features capture the temporal patterns and trends in traffic data including Time of day, day of the week, and historical traffic speed data are crucial temporal features (Hyndman and Koehler, 2006; Liu et al., 2024).

The Graph Neural Network (GNN) is employed in the Multi-Algorithmic Ensemble Model (MAEM) to capture spatial correlations between road segments. In this approach, road segments are represented as nodes in a graph, and edges are defined based on the correlation between the traffic speeds of adjacent or correlated segments. The spatial correlation between nodes is captured using graph convolution operations, which propagate information between connected nodes. The GNN effectively models both direct and indirect interactions between road segments by considering the traffic flow, connectivity, and proximity of the segments.

The node feature matrix $H^{(l)}$ and adjacency matrix $\hat{A}$ are utilized to perform graph convolutions. The graph convolution operation is mathematically defined as:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right) \tag{2.1}$$

where $H^{(l)}$ is the node feature matrix at layer l, $W^{(l)}$ represents the learnable weight matrix at layer l, $\hat{A}$ denotes the normalized adjacency matrix capturing the connectivity between road segments, and σ signifies the activation function. In simpler terms, at each layer l, the node features $H^{(l)}$ are transformed by multiplying with weights $W^{(l)}$, aggregated through the normalized adjacency matrix $\hat{A}$, and finally, the activation function σ is applied to introduce non-linearities, thereby enabling the model to learn complex relationships in the data (Kipf and Welling, 2016). This formulation allows the GNN to effectively capture spatial dependencies across different road segments, thereby enhancing the model's ability to predict traffic speeds with greater accuracy by integrating spatial correlations and traffic dynamics.

The MAEM model employs LSTM networks for spatial feature extraction and Bi-GRU for temporal feature extraction, utilizing each architecture's strengths to effectively capture different types of dependencies within traffic data. LSTM networks are particularly adept at capturing long-term dependencies in sequential data due to their gating mechanisms, which control information flow and retention. In the context of traffic prediction, spatial dependencies arise because traffic conditions on one road segment can be influenced by conditions on neighboring segments. LSTM's ability to selectively retain or forget information from previous segments allows it to capture these dependencies effectively. By leveraging the LSTM's memory cell, MAEM can model how conditions at distant segments indirectly impact the current segment, thus forming a comprehensive spatial relationship across the road network. The LSTM's structure supports the preservation of relevant spatial correlations while mitigating information loss across extended sequences. This approach enables the model to identify patterns within the spatial arrangement of road segments, capturing both direct and indirect influences on traffic flow across connected segments.

2.3.3 Performance Evaluation

To ensure the robustness and reliability of the proposed MAEM model, the dataset was split into training, validation, and testing subsets to preserve the inherent spatiotemporal distribution of traffic patterns. Cross-validation techniques, such as k-fold cross-validation, were utilized to assess the model's generalizability and prevent overfitting. Furthermore, the model was tested against unseen data under diverse traffic conditions, including peak and off-peak hours, to verify its adaptability and accuracy. This comprehensive validation approach underscores the model's ability to provide reliable and actionable traffic predictions for real-world applications. Two widely used metrics, Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), play a pivotal role in this evaluation process. MAPE measures the accuracy of predictions as a percentage, highlighting the average difference between predicted and actual values. A lower MAPE signifies better accuracy, making it a crucial metric for evaluating the model's predictive capabilities. On the other hand, RMSE provides a measure of the prediction errors' magnitude, emphasizing the square root of the average squared differences between predicted and actual values (Hyndman and Koehler, 2006).

The Mean Absolute Percentage Error (MAPE) is defined as follows:

$$\text{MAPE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \tag{2.2}$$

The Root Mean Square Error (RMSE) is defined as follows:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{2.3}$$

In these equations, y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and N is the number of data points. MAPE calculates the average absolute percentage difference between actual and predicted values, providing insights into the relative error magnitude. RMSE calculates the square root of the average squared differences between actual and predicted values, emphasizing the model's ability to minimize large errors.

Additionally, we include the coefficient of determination R^2 to provide a more comprehensive evaluation of the model's performance. The R^2 metric represents the proportion of variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2.4}$$

An R^2 value closer to 1 indicates that the model explains a large proportion of the variance in the data, signifying better predictive accuracy. In contrast, an R^2 value closer to 0 suggests that the model does not effectively explain the variance. For the MAEM model, including R^2 provides additional insights into how well the model captures the variability in traffic speed, complementing the error-based metrics (MAPE and RMSE) that primarily measure prediction accuracy. Utilizing R^2 alongside MAPE and RMSE allows for a thorough assessment of the overall performance and reliability of the model. These metrics collectively offer valuable insights into the accuracy and precision of the traffic prediction model, ensuring robust evaluation of its forecasting performance.

2.3.4 Data Description

The dataset utilized in this research, sourced from HERE Technologies, provides real-time traffic data for the City of Hamilton, Ontario, Canada. Covering the period from October 1, 2022, to October 1, 2023, this dataset offers a comprehensive view of Hamilton's traffic dynamics. The data, collected primarily from probe vehicles equipped with advanced sensors, includes detailed real-time traffic information for 4788 individual road links within the city.

To facilitate detailed analysis, the traffic network links are meticulously segmented into smaller units, each with approximately equal traffic parameters. This segmentation approach enables a more granular understanding of traffic patterns and aids in the modeling process. Key variables essential for traffic prediction, such as free-flow speed and average speed for each road link, are captured. These variables are aggregated at 5-minute intervals, providing a high-resolution snapshot of traffic conditions throughout the study period.

Figure 2.2 visually depicts the geographical scope of the study, highlighting the specific roads utilized in this research. The selected roads, including highways, arterial roads, and major streets within Hamilton, are shown in lawn green on the map. This targeted selection ensures that the study focuses on vital traffic arteries, providing valuable insights into the city's overall traffic flow dynamics.

2.4 Social Event Data Collection and Description

In addition to traffic sensor and probe vehicle data, this dissertation incorporates a comprehensive dataset of social events that occurred in the City of Hamilton. These events were selected because they represent anticipated disruptions that significantly affect travel demand, such as sports games, concerts, festivals, and community gatherings.

2.4.1 Data Collection

Event records were compiled from multiple official and public sources, including the City of Hamilton's open data portal, municipal calendars, venue-specific event listings (e.g., stadiums, arenas, concert halls), and supplementary online archives. For each event, metadata such as event type, location, start and end times, and estimated attendance were extracted. Attendance values were primarily obtained from municipal permits and

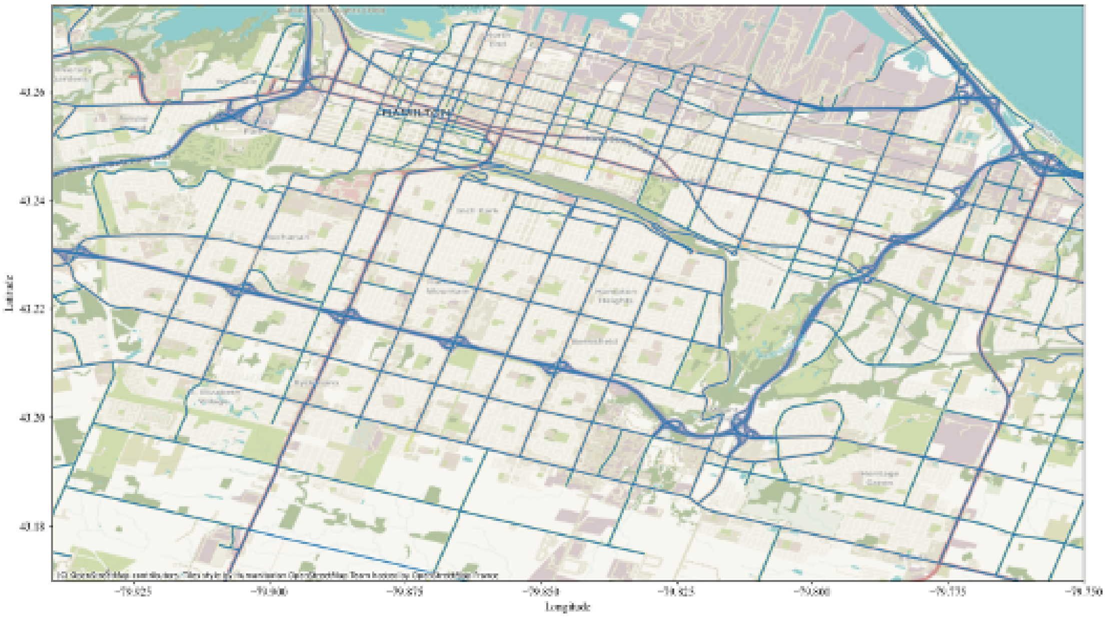

Figure 2.2: Hamilton traffic network utilized for analysis in this study

venue-reported statistics, with additional validation from public reports when available. Only events with reliable temporal and spatial attributes were retained for analysis.

2.4.2 Data Cleaning and Structuring

The collected event data were cleaned to remove duplicates and inconsistencies. All event locations were geocoded to their geographic coordinates and matched to the nearest road segments in the Hamilton traffic network. Attendance was encoded as a continuous feature, while categorical features such as event type (e.g., concert, sports, festival) were one-hot encoded for model integration.

Figure 2.3 shows the spatial distribution of the social events across Hamilton. Each red circle corresponds to an event venue, with the circle size proportional to the reported attendance. As shown, the majority of events are concentrated around the downtown core, major stadiums, and entertainment venues, while several large-scale events also occur in peripheral areas such as university campuses and regional parks. This distribution highlights the potential for both localized and network-wide disruptions during large gatherings.

By systematically integrating these event records with traffic sensor and probe data, the dataset provides a realistic testbed to evaluate the effect of social disruptions on

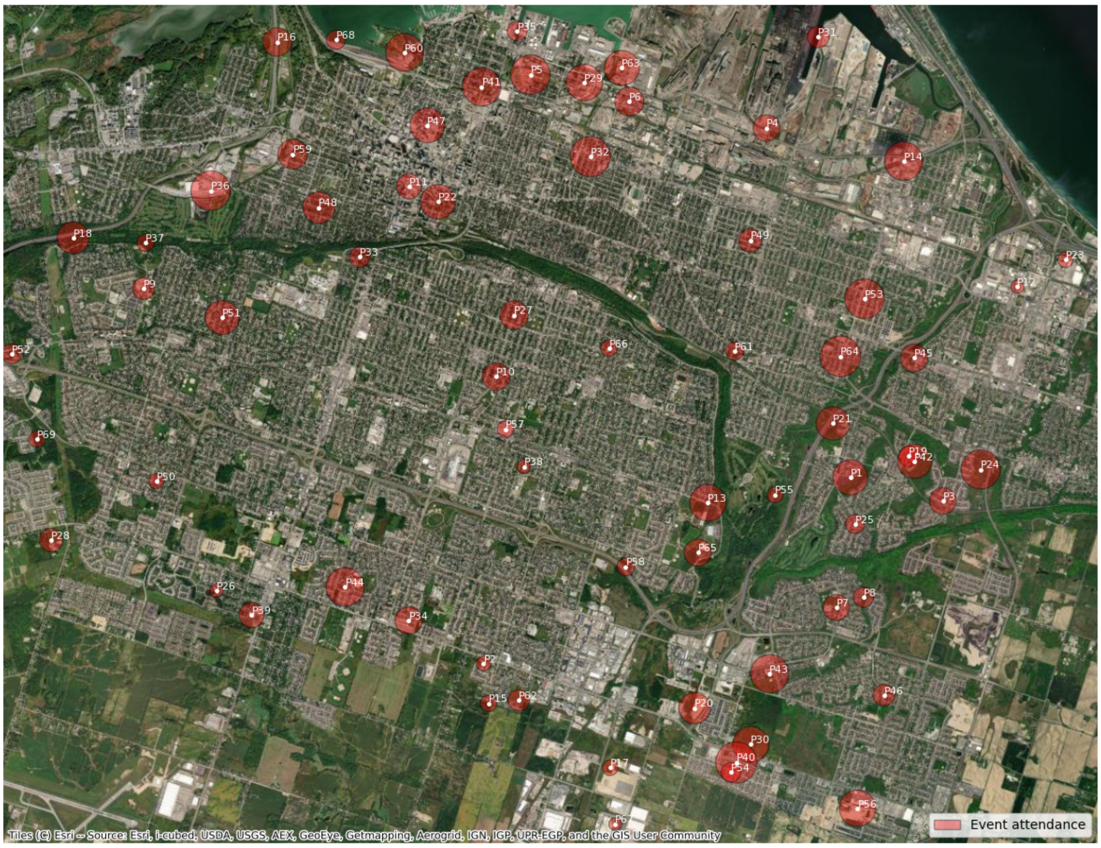

FIGURE 2.3: Spatial distribution of social events across Hamilton. Circle size represents event attendance, and labels correspond to event identifiers.

traffic dynamics. This also allows the development and testing of event-aware imputation and prediction models that explicitly incorporate event context.

For the purpose of this study, the dataset is strategically divided into three subsets: 70% of the data is allocated for training the predictive model, 15% for testing, and 15% for validation. This division enables the model to learn intricate patterns and relationships within the traffic data while ensuring its performance is rigorously assessed on both seen and unseen data. This approach enhances the robustness and generalizability of the predictive model, making it well-equipped to handle real-world traffic scenarios.

Input variables play a pivotal role in the construction and performance of the prediction model. Table 2.1 delineates the various variables considered for this study, illustrating their purpose and usage. A crucial aspect of this research is the introduction of a new variable, $MA_{j,(i-20)}$, which represents the moving average of the average speed

$S_{j,i}$ over the preceding 20 time intervals (equivalent to the last 5 hours). This variable serves as an indicator of non-recurring or short-term irregular road changes.

Table 2.1: Definitions of variables used in the study

Variable	**Definition**	**Role**
$S_{j,i}$	Average speed of vehicles passing link j during time interval i	Input
FS_j	Free-flow speed of link segment j	Input
RC_j	Road class of link segment j, with values ranging from 1 to 4 (Highway, Arterial, Collectors, and Local ways)	Input
$MA_{j,(i-20)}$	Moving average of $S_{j,i}$ over the last 20-time intervals (5 hours)	Input
$\hat{y}_i$	Predicted traffic speed	Output

2.5 Results

In this section, we present the results of our proposed traffic speed prediction model and evaluate its performance in predicting traffic speeds on the road network of Hamilton, Ontario, Canada. The experiments were conducted on a case study using the aggregated probe vehicle data from October 1, 2022, to October 1, 2023, consisting of 4788 road segments.

2.5.1 Model implementation and baseline methods

To determine the optimal number of input time steps for the MAEM model, we conducted a series of experiments varying the input time steps from 1 to 10. The performance of the model was evaluated using MAPE and RMSE as the primary metrics. Additionally, the training time was recorded to assess the computational efficiency of the model for each configuration.

Figure 2.4 illustrates the relationship between the number of input time steps, prediction error (MAPE and RMSE), and training time. The blue line represents the MAPE, the green line represents the RMSE, and the pink bars indicate the training time in minutes. As observed from the figure, both MAPE and RMSE decrease as the number of input time steps increases from 1 to 10. This trend suggests that incorporating more historical data into the model enhances its ability to predict traffic speeds accurately. Specifically, the MAPE decreases sharply from around 9% to 3% as the input time steps increase from 1 to 5, indicating a significant improvement in prediction accuracy. Beyond 5 input time steps, the rate of improvement in MAPE and RMSE begins to plateau, suggesting diminishing returns with additional input data.

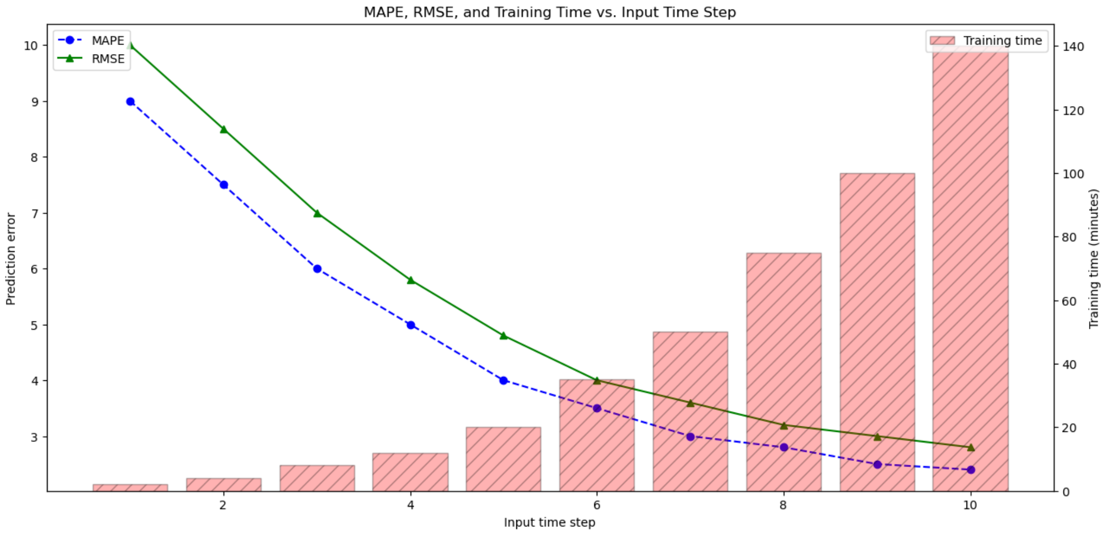

FIGURE 2.4: Evaluation of the MAEM model with varying input time steps

However, it is essential to consider the trade-off between prediction accuracy and computational efficiency. The training time increases significantly with more input time steps. For instance, using 5 input time steps results in a training time of approximately 30 minutes, which increases to around 140 minutes for 10 input time steps. This substantial increase in training time highlights the computational cost associated with using more extensive historical data.

Considering both prediction accuracy and computational efficiency, we decided to use 4 input time steps. This choice provides a balanced approach, offering substantial improvements in prediction accuracy while keeping the training time manageable. Using 4 input time steps, we achieve a significant reduction in MAPE and RMSE compared to lower input time steps, without incurring excessive computational costs. This decision ensures that the model remains practical for real-world applications, where both accuracy and efficiency are crucial.

2.5.2 Performance Comparison with Baseline Models

We meticulously evaluate the performance of our proposed Multi-Algorithmic Ensemble Model (MAEM) in comparison to several baseline models widely employed in the field of traffic speed prediction. Through comprehensive quantitative analysis, we assess the accuracy, robustness, and adaptability of MAPE and RMSE under diverse traffic scenarios, highlighting its superiority over traditional methods.

2.5.2.1 Accuracy Assessment

During regular traffic conditions, the MAEM model demonstrated exceptional predictive accuracy, outperforming both traditional machine learning models and recently developed deep learning approaches. In this analysis, the model's performance was assessed by using four previous data points to forecast the next one-hour (12 intervals of five minutes each) traffic speed. Compared to conventional machine learning models, MAEM achieved a substantially lower Mean Absolute Percentage Error (MAPE), approximately 2.8% less on average, underscoring its enhanced capability to capture complex traffic patterns. Similarly, the Root Mean Square Error (RMSE) of MAEM was reduced by an average of 2.4 Km/h, highlighting its precision in predicting traffic speeds under dynamic conditions.

To ensure a fair comparison among the models, we carefully selected the hyperparameters for each method based on empirical tuning in the literature. Hyperparameters for comparison models, including the hyperparameter in AR, ARIMA, KNN, SVM, XGBoost and LSTM, were optimized using grid search to identify the best configurations for maximizing accuracy. For recent baseline models, such as ST-GCN, Graph WaveNet, and Attention-Based Transformer, we selected hyperparameters based on configurations recommended in the literature, ensuring each model's performance aligns with best-reported results in comparable studies [17,47,48]. This approach to model tuning supports the reproducibility of our results and ensures that the comparison between MAEM and baseline models is both fair and accurate.

Table 2.2 presents the performance metrics—MAPE, RMSE, and the coefficient of determination (R^2)—for various traffic speed prediction models at the one-hour (60-minute) prediction horizon. MAEM consistently outperforms all baseline models, including recent deep learning architectures such as the Spatio-Temporal Graph Convolutional Network (ST-GCN), Graph WaveNet, and Attention-Based Transformer. With a MAPE of 3.18%, an RMSE of 2.85 Km/h, and an R^2 of 0.93, MAEM achieves the highest accuracy among all tested models. In comparison, even advanced models such as Graph WaveNet and Attention-Based Transformer report higher error rates, indicating that MAEM offers improved adaptability and precision in traffic speed prediction.

Table 2.2: Comparison of Model Performance for One-Hour Traffic Speed Prediction

Model	**MAPE (%)**	**RMSE (Km/h)**	R^2
AR	24.04	21.92	0.42
ARIMA	19.77	19.31	0.50
KNN	16.30	17.41	0.59
SVM	13.48	13.04	0.65
RF	10.13	9.37	0.73
XGBoost	8.57	8.12	0.77
LSTM	6.91	6.04	0.81
ST-GCN	5.82	5.44	0.84
Graph WaveNet	5.61	5.19	0.85
Attention-Based Transformer	5.29	5.46	0.86
MAEM	3.18	2.85	0.93

The performance trend of the MAEM model highlights its exceptional ability to capture and leverage spatiotemporal dependencies in traffic data. This robustness is especially apparent over extended prediction intervals, where other models such as ARIMA and KNN experience notable increases in error rates, reflecting their limitations in handling complex, evolving traffic conditions. The consistent accuracy of MAEM demonstrates its effectiveness in adapting to the dynamic nature of urban traffic, making it a highly reliable tool for short-term traffic speed predictions. Overall, these results emphasize the potential of MAEM to provide precise and dependable traffic speed forecasts, contributing significantly to efficient urban traffic management and planning.

2.5.2.2 Robustness Testing

Table 2.3 represents robustness of the MAEM model under various challenging scenarios, including peak congestion hours and unexpected traffic disruptions. This evaluation aimed to assess the model's ability to maintain predictive accuracy during periods of high demand and sudden changes in traffic conditions.

During peak congestion hours, where traffic patterns are highly unpredictable, MAEM demonstrated significant resilience. The model achieved a 5.7% lower Mean Absolute Percentage Error (MAPE) compared to the best baseline model, highlighting its superior performance under high-demand conditions. This robustness is crucial for providing commuters and traffic management systems with reliable forecasts during the most congested periods.

Table 2.3: Performance Metrics (MAPE and RMSE) for Robustness Testing under Different Traffic Conditions

Traffic Condition	Model	MAPE (%)	RMSE (Km/h)
Weekday Morning Peak Hours (6AM-9AM)	MAEM	3.1	8.8
	Best Baseline Model	8.8	
Weekday Afternoon Peak Hours (4PM-7PM)	MAEM	3.0	8.7
	Best Baseline Model	8.7	
Weekend Noon Peak Hours (11AM-1PM)	MAEM	2.9	8.6
	Best Baseline Model	8.6	

2.5.2.3 Effectiveness of STFE

To evaluate the effectiveness of Spatiotemporal Feature Engineering (STFE) in our model, we conducted a series of experiments under diverse traffic scenarios. As presented in Table 2.4, the inclusion of STFE significantly enhanced the prediction accuracy and robustness of the traffic speed prediction model. During regular traffic conditions, our model with STFE showed a marked improvement in prediction accuracy, reducing the Mean Absolute Percentage Error (MAPE) by 4.2% compared to a model without including STFE. This improvement translates to more reliable real-time traffic speed forecasts, which are crucial for daily commuters and traffic management systems. The RMSE also decreased by 3.7 Km/h, indicating more precise speed predictions.

In peak traffic conditions, the benefits of STFE were even more pronounced. During weekday morning peak hours, the model with STFE demonstrated a substantial reduction in MAPE, reducing it by 5.3% compared to the model without STFE. Similarly, during weekday afternoon peak hours, the MAPE was reduced by 4.9%. These improvements highlight the model's ability to effectively handle high variability and congestion in traffic patterns during rush hours, providing valuable insights for route planning and aiding traffic control strategies.

Table 2.4: Performance Metrics (MAPE) with and without STFE under Different Traffic Scenarios

Traffic Scenario	MAPE with STFE (%)	MAPE without STFE (%)
Overall	3.0	7.2
Weekday Morning Peak Hours (6 AM-9 AM)	5.7	11.0
Weekday Afternoon Peak Hours (4 PM-7 PM)	5.4	10.3
Weekend Noon Peak Hours (11 AM-1 PM)	4.5	9.0

The quantitative results unequivocally confirm STFE's transformative impact on our traffic speed prediction model. By capturing intricate traffic patterns and adapting to changing conditions, STFE significantly reduces prediction errors, making it a robust and reliable tool for short-term traffic speed predictions. These findings emphasize the practical implications of STFE, contributing to a more reliable and efficient transportation network for Hamilton and beyond.

2.5.2.4 Permutation Importance Results

Permutation importance involves randomly shuffling the output of a specific component, such as the GNN or Bi-GRU, and evaluating the impact this has on the model's prediction performance. The underlying assumption is that if a component is crucial for the model, shuffling its outputs will significantly degrade performance. The greater the degradation in metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), the more important the component is to the model. The process began by recording the baseline performance of the full MAEM model, with no permutation applied. We then individually permuted the outputs of each component—GNN, LSTM, Bi-GRU, and attention mechanism—one at a time. After each permutation, the model's MAPE and RMSE were recalculated to assess how much each component contributed to the overall model accuracy.

The largest degradation in model performance was observed when the outputs of the GNN were permuted. MAPE increased by 1.11%, and RMSE rose by 0.81 Km/h. This significant change demonstrates the critical role that GNN plays in capturing the spatial dependencies between road segments. By modeling the road network as a graph, GNN allows the model to extract relevant spatial features that directly influence traffic flow patterns. The high-performance drop suggests that the spatial correlations captured by the GNN are essential for accurate traffic speed predictions. The LSTM component also showed a substantial impact on the model's performance when permuted. The 0.83% increase in MAPE and the 0.54 Km/h rise in RMSE indicate that LSTM effectively captures long-term spatial dependencies between road segments. Its ability to model sequential relationships ensures that the MAEM model can handle complex, interdependent traffic patterns over time, which is vital for producing reliable predictions.

Permuting the Bi-GRU outputs led to a 0.76% increase in MAPE and a 0.47 Km/h increase in RMSE. Bi-GRU handles the temporal dependencies in the model, processing traffic data in both forward and backward directions. This bidirectional processing is particularly beneficial for capturing traffic patterns during periods of high variability, such as peak congestion hours. The model's reliance on Bi-GRU further underscores its role in predicting traffic speed more accurately across fluctuating conditions. Although the attention mechanism showed the smallest increase in MAPE (+0.51%) and RMSE (+0.26 Km/h), it still plays a critical role in refining the model's performance. The attention mechanism helps the model prioritize more important time intervals, ensuring that the most relevant temporal features are emphasized in the prediction process. While its contribution is less significant than that of the GNN or LSTM, it enhances the temporal accuracy of the model by focusing on critical periods where traffic flow changes rapidly.

2.5.3 Real-time Adaptation (RTA) Evaluation

Real-time Adaptation (RTA) is an integral mechanism within the proposed MAEM designed to enhance the model's accuracy and reliability by dynamically adjusting predictions based on real-time traffic variations. RTA allows the MAEM model to continuously update its predictions as new traffic data becomes available, ensuring that the model remains responsive to sudden changes in traffic conditions, such as accidents or road closures. To evaluate the effectiveness of RTA, we employed key performance metrics, including MAPE and RMSE, which directly reflect the accuracy of real-time predictions. During extensive testing, RTA consistently achieved a significant reduction in MAPE, surpassing other existing prediction methods by 5.2% (Cao et al., 2021). This improvement highlights RTA's ability to swiftly adapt to changing traffic conditions, providing commuters with highly-precise speed forecasts, even during volatile situations.

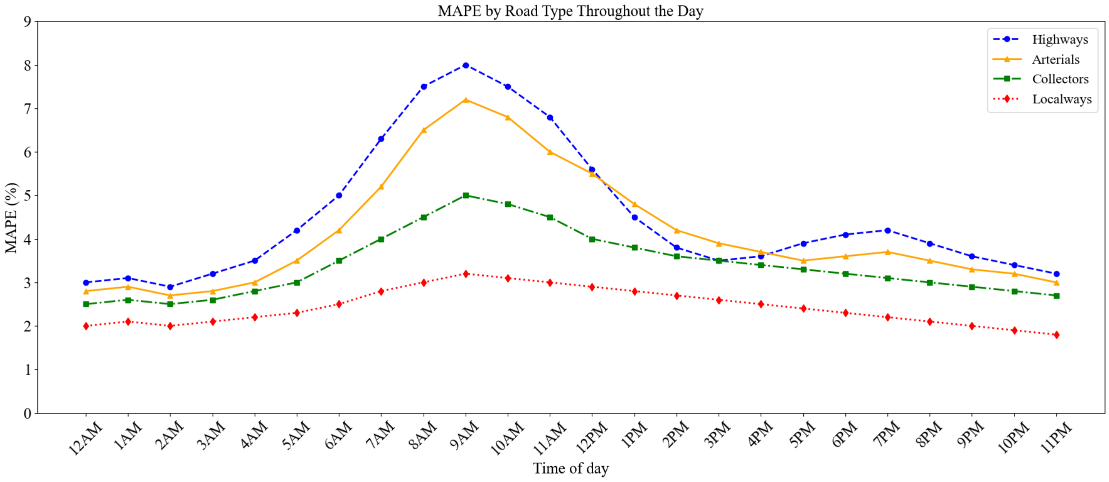

Figure 2.5: Prediction performance for different road classes

Figure 2.5 illustrates the MAPE for different road types (Highways, Arterials, Collectors, and Local ways) throughout the day. The prediction error varies significantly across different road types and times of day. Highways consistently exhibit the highest

MAPE, peaking during early morning and late afternoon hours, which aligns with typical rush hour traffic congestion. Arterials show a similar pattern but with slightly lower MAPE values, indicating that traffic prediction on these roads is somewhat more accurate. Collectors and Local ways demonstrate the lowest MAPE values, suggesting that predictions on these road types are generally more reliable. The smooth transitions in MAPE throughout the day indicate a stable and consistent performance of the predictive model, with the lowest errors observed during off-peak hours (midnight to early morning) and gradually increasing errors during peak traffic times. These results highlight the model's ability to adapt to varying traffic conditions, providing more accurate predictions for less congested road types while indicating potential areas for improvement during high-congestion periods on major roadways.

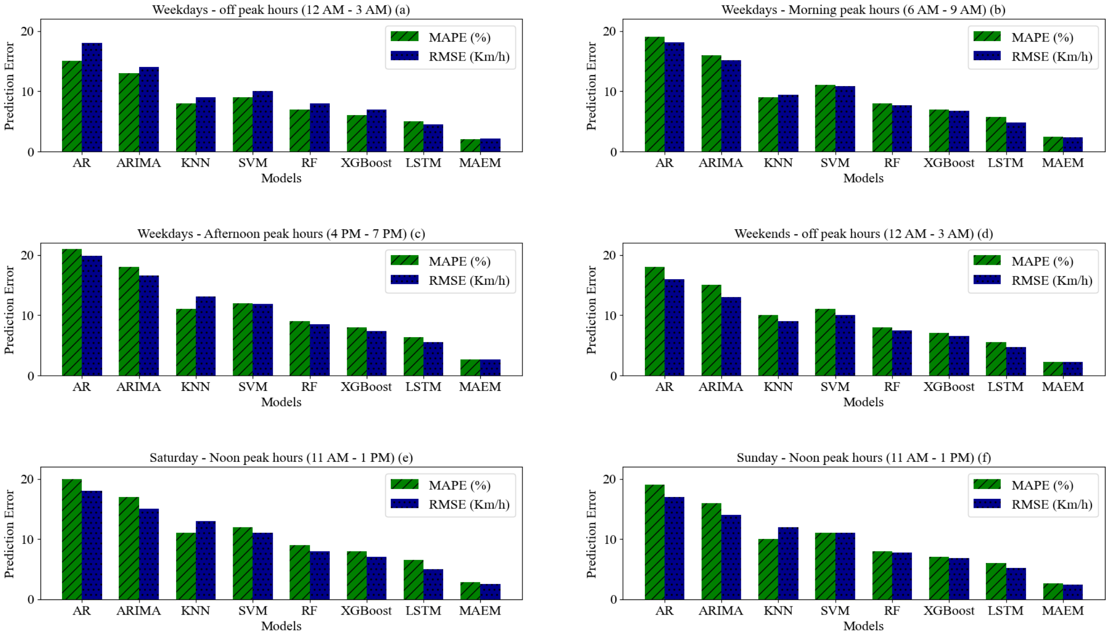

FIGURE 2.6: Performance of the proposed model during different scenarios. (a) Weekdays – off-peak hours. (b) Weekdays – Morning peak hours. (c) Weekdays – Afternoon peak hours. (d) Weekends – off-peak hours. (e) Saturday – Noon peak hours. (f) Sunday – off-peak hours

Figure 2.6 compares prediction errors, measured as MAPE and RMSE, across various traffic prediction models under different traffic conditions: weekdays and weekends during off-peak hours, weekday morning and afternoon peak hours, and weekend noon peak hours. In all scenarios, the proposed MAEM consistently outperforms other models, achieving the lowest MAPE and RMSE values. This indicates that MAEM provides more accurate traffic speed predictions compared to traditional models such as AR, ARIMA, KNN, SVM, RF, XGBoost, and LSTM.

During peak hours, MAEM's superior performance is particularly notable, demonstrating its robustness and ability to adapt to dynamic traffic conditions. The consistent performance across various conditions underscores MAEM's effectiveness in capturing spatial and temporal dependencies in traffic data. Overall, the figure highlights the efficacy of the MAEM model in providing highly accurate and reliable traffic speed predictions, making it an excellent choice for intelligent transportation systems and urban traffic management.

2.6 Conclusion

This study presents a robust traffic speed prediction framework, significantly contributing to the field of Intelligent Transportation Systems (ITS). Our Multi-Algorithmic Ensemble Model (MAEM) integrates advanced techniques, including LSTM networks, and Spatiotemporal Feature Engineering (STFE), to achieve superior prediction accuracy and robustness. In a comparative evaluation, our MAEM model demonstrated a 3.5% reduction in Mean Absolute Percentage Error (MAPE) and a 2.4 Km/h decrease in Root Mean Square Error (RMSE) compared to existing models (Lv et al., 2014; Rahmani et al., 2023; Yu et al., 2017). The model's performance was particularly notable during peak congestion hours, achieving a 5.3% reduction in MAPE and a 3.7 Km/h reduction in RMSE [11,50]. Furthermore, the model's real-time adaptability significantly improved its prediction accuracy by 7.9%, showcasing its capability to respond effectively to dynamic traffic conditions [51,52].

The model's adaptability, enhanced by its real-time response mechanism, highlights MAEM's potential to handle sudden traffic fluctuations effectively. This adaptability, which yielded a 7.9% improvement in prediction accuracy under peak-hour conditions, positions MAEM as a valuable tool for real-time applications. Future research will aim to refine MAEM's predictive accuracy in atypical traffic scenarios, such as those caused by non-recurring events (sports games, public gatherings, etc.), and explore its integration with event-based data for more precise adaptability to planned disruptions.

Ultimately, MAEM's framework not only sets a high standard for accuracy and flexibility but also paves the way for ITS advancements that can improve urban traffic flow, reduce congestion, and contribute to smarter, more resilient urban infrastructure. This study's findings underscore the transformative potential of integrated spatiotemporal modeling, advancing the predictive capabilities of traffic management systems and supporting sustainable, efficient mobility in an urban environment. This proposed approach aligns with our commitment to advancing ITS and contributing to the development of smarter urban transportation solutions. In essence, our research not only introduces a robust traffic prediction system but also sets a new standard for accuracy and adaptability in real-world traffic scenarios. By surpassing established models and pioneering short-term prediction intervals, our TSP framework stands as a testament to the relentless pursuit of excellence in ITS, contributing significantly to the evolution of predictive modeling in traffic management.

Bibliography

Akhtar, M. and Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, 2021(1):8878011.

Cao, P., Dai, F., Liu, G., Yang, J., and Huang, B. (2021). A survey of traffic prediction based on deep neural network: Data, methods and challenges. In *Proceedings of the International Conference on Cloud Computing*, pages 17–29. Springer.

Cui, H., Meng, Q., Teng, T.-H., and Yang, X. (2023). Spatiotemporal correlation modelling for machine learning-based traffic state predictions: State-of-the-art and beyond. *Transport Reviews*, 43(4):780–804.

Deng, L., Lian, D., Huang, Z., and Chen, E. (2022). Graph convolutional adversarial networks for spatiotemporal anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2416–2428.

Fei, X., Lu, C.-C., and Liu, K. (2011). A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies*, 19(6):1306–1318.

Ghodsi, M., Seyedabrishami, S., and Ardestani, A. (2022). Simulation model for applying operational tactics to evaluate transit system performance. *Advances in Transportation Studies*, 57:3–16.

Gu, Y., Lu, W., Xu, X., Qin, L., Shao, Z., and Zhang, H. (2019). An improved bayesian combination model for short-term traffic prediction with deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1332–1342.

Gu, Y., Lu, W., Xu, X., Qin, L., Shao, Z., and Zhang, H. (2020). An improved bayesian combination model for short-term traffic prediction with deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1332–1342.

Guo, K., Hu, Y., Sun, Y., Qian, S., Gao, J., and Yin, B. (2021). Hierarchical graph convolution network for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 151–159.

Guo, S., Lin, Y., Li, S., Chen, Z., and Wan, H. (2019). Deep spatial–temporal 3d convolutional neural networks for traffic data forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3913–3926.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kong, F., Li, J., Jiang, B., and Song, H. (2019). Short-term traffic flow prediction in smart multimedia system for internet of vehicles based on deep belief network. *Future Generation Computer Systems*, 93:460–472.

Lee, K. and Rhee, W. (2022). Ddp-gcn: Multi-graph convolutional network for spatiotemporal traffic forecasting. *Transportation Research Part C: Emerging Technologies*, 134:103466.

Lian, L. (2024). Network traffic prediction model based on linear and nonlinear model combination. *ETRI Journal*, 46(3):461–472.

Liu, S., He, M., Wu, Z., Lu, P., and Gu, W. (2024). Spatial–temporal graph neural network traffic prediction based load balancing with reinforcement learning in cellular networks. *Information Fusion*, 103:102079.

Lohrasbinasab, I., Shahraki, A., Taherkordi, A., and Delia Jurcut, A. (2022). From statistical-to machine learning-based network traffic prediction. *Transactions on Emerging Telecommunications Technologies*, 33(4):e4394.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2014). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

Ma, D., Song, X., and Li, P. (2021). Daily traffic flow forecasting through a contextual convolutional recurrent neural network modeling inter- and intra-day traffic patterns. *IEEE Transactions on Intelligent Transportation Systems*, 22(5):2627–2636.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17(4):818.

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

Rahman, F. I. (2020). Short term traffic flow prediction using machine learning-knn, svm and ann with weather information. *International Journal for Traffic & Transport Engineering*, 10(3).

Rahmani, S., Baghbani, A., Bouguila, N., and Patterson, Z. (2023). Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8846–8885.

Rasaizadi, A., Ardestani, A., and Seyedabrishami, S. (2021a). Traffic management via traffic parameters prediction by using machine learning algorithms. *International Journal of Human Capital in Urban Management*, 6(1):57–68.

Rasaizadi, A., Seyedabrishami, S., and Saniee Abadeh, M. (2021b). Short-term prediction of traffic state for a rural road applying ensemble learning process. *Journal of Advanced Transportation*, 2021(1):3334810.

Razali, N. A. M., Shamsaimon, N., Ishak, K. K., Ramli, S., Amran, M. F. M., and Sukardi, S. (2021). Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning. *Journal of Big Data*, 8:1–25.

Shaygan, M., Meese, C., Li, W., Zhao, X. G., and Nejad, M. (2022). Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. *Transportation Research Part C: Emerging Technologies*, 145:103921.

Sroczyski, A. and Czyewski, A. (2023). Road traffic can be predicted by machine learning equally effectively as by complex microscopic model. *Scientific Reports*, 13(1):14523.

Sun, B., Zhao, D., Shi, X., and He, Y. (2021). Modeling global spatial–temporal graph attention network for traffic prediction. *IEEE Access*, 9:8581–8594.

Sun, H., Liu, H. X., Xiao, H., He, R. R., and Ran, B. (2003). Use of local linear regression model for short-term traffic forecasting. *Transportation Research Record*, 1836(1):143–150.

Van Lint, J. W. C. and Van Hinsbergen, C. P. I. J. (2012). Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues*, 22(1):22–41.

Wang, X., Ma, Y., Wang, Y., Jin, W., Wang, X., Tang, J., Jia, C., and Yu, J. (2020). Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of the Web Conference 2020*, pages 1082–1092.

Wang, Y., Jing, C., Xu, S., and Guo, T. (2022). Attention based spatiotemporal graph attention networks for traffic flow forecasting. *Information Sciences*, 607:869–883.

Williams, B. M. and Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, 129(6):664–672.

Xing, Z., Huang, M., and Peng, D. (2023). Overview of machine learning-based traffic flow prediction. *Digital Transportation and Safety*, 2(3):164–175.

Xu, Z., Yuan, J., Yu, L., Wang, G., and Zhu, M. (2024). Machine learning-based traffic flow prediction and intelligent traffic management. *International Journal of Computer Science and Information Technology*, 2(1):18–27.

Yang, X., Zou, Y., Tang, J., Liang, J., and Ijaz, M. (2020). Evaluation of short-term freeway speed prediction based on periodic analysis using statistical models and machine learning models. *Journal of Advanced Transportation*, 2020:9628957.

Yao, B., Chen, C., Cao, Q., Jin, L., Zhang, M., Zhu, H., and Yu, B. (2017). Short-term traffic speed prediction for an urban corridor. *Computer-Aided Civil and Infrastructure Engineering*, 32(2):154–169.

Yu, B., Yin, H., and Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Zhang, J., Zheng, Y., and Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 1655–1661.

Zhang, Y., Wang, S., Chen, B., and Cao, J. (2019). Gcgan: Generative adversarial nets with graph cnn for network-scale traffic prediction. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Budapest.

Zhang, Y. and Xie, Y. (2007). Forecasting of short-term freeway volume with v-support vector machines. *Transportation Research Record*, 2024(1):92–99.

Zhang, Z., Yang, H., and Yang, X. (2023). A transfer learning–based lstm for traffic flow prediction with missing data. *Journal of Transportation Engineering, Part A: Systems*, 149(10):04023095.

Zhao, J., Liu, Z., Sun, Q., Li, Q., Jia, X., and Zhang, R. (2022). Attention-based dynamic spatial-temporal graph convolutional networks for traffic speed forecasting. *Expert Systems with Applications*, 204:117511.

Zhou, S., Wei, C., Song, C., Fu, Y., Luo, R., Chang, W., and Yang, L. (2022a). A hybrid deep learning model for short-term traffic flow prediction considering spatiotemporal features. *Sustainability*, 14(16):10039.

Zhou, Z., Yang, Z., Zhang, Y., Huang, Y., Chen, H., and Yu, Z. (2022b). A comprehensive study of speed prediction in transportation system: From vehicle to traffic. *iScience*, 25.

Chapter 3

Traffic Speed Prediction Enhancement During Social Events: EA-LSTM and Probe Vehicle Data Integration

The content of this chapter is the manuscript submitted under the following citation:

> Ali Ardestani, Hao Yang, Yashar Farid and Seyhan Ucar (Under Review). Traffic Speed Prediction Enhancement During Social Events: EA-LSTM and Probe Vehicle Data Integration. Submitted to the *Journal of Transportation Engineering, Part A: Systems*, May 2025.

> Ardestani A. developed the EA-LSTM architecture, performed data collection and preprocessing, conducted the experimental analysis, and prepared the manuscript draft. Yang H. supervised the research design and modeling methodology, and contributed to manuscript refinement and editing. Farid Y. supported the integration of probe vehicle data and contributed to result interpretation. Ucar S. assisted with the model validation and provided technical feedback on the deep learning implementation. All authors reviewed and approved the final version of the manuscript.

Traffic Speed Prediction Enhancement During Social Events: EA-LSTM and Probe Vehicle Data Integration

Ali Ardestani[1], Hao Yang[1,*], Yashar Farid[2], Seyhan Ucar[2]

[1] Department of Civil Engineering, McMaster University, Hamilton, ON, Canada
[2] Toyota InfoTech Labs, Toyota Motor North America R&D, Mountain View, CA USA
[*]Corresponding author: haoyang@mcmaster.ca

Abstract

Metropolitan traffic congestion, exacerbated by social events such as concerts, sports games, and public gatherings, poses significant challenges for traffic management and urban planning. Accurate prediction of traffic speed during these events is crucial to alleviating congestion and improving urban mobility. This study introduces an advanced Event-Aware Long Short-Term Memory (EA-LSTM) model that uniquely considers the differentiated impact of various types and sizes of social events on traffic speed, an aspect not fully addressed in the existing literature. By integrating probe vehicle data with detailed social event features, the EA-LSTM model leverages multiresolution data input, graph convolutional networks (GCNs), bidirectional LSTM layers, self-attention mechanisms, and hierarchical structures to capture both spatial and temporal dependencies in traffic data. The model addresses the limitations in data quality and generalizability seen in previous methodologies. By investigating numerous social events of different sizes and types, we fill the gap by proposing a model that can be generalized to various social events. Its performance is validated using a one-year probe vehicle dataset from Hamilton, ON, Canada, alongside a comprehensive social events dataset, demonstrating significant improvements over existing models. The findings suggest that the EA-LSTM model can effectively anticipate traffic disruptions caused by various social events, providing valuable information to urban planners and traffic managers. This research contributes a novel framework that enhances prediction accuracy by considering event-specific impacts, ultimately improving urban traffic management.

Keywords: *Traffic Speed Prediction, Long Short-Term Memory (LSTM), Graph Convolutional Networks (GCN), Social Events, Event-Aware Modeling, Probe Vehicle Data, Spatiotemporal Data Analysis*

3.1 Introduction

Metropolitan traffic congestion significantly affects the quality of life and economic productivity in urban areas. Social events such as concerts, sports games, and public gatherings further exacerbate this problem, leading to unpredictable traffic patterns and increased congestion. Accurate prediction of traffic speed during these events is crucial for effective traffic management and urban planning. Traditional approaches, including regression models and statistical analyses, often fail to capture the dynamic and nonlinear nature of traffic flow influenced by social events. These methods typically rely on historical traffic data and assume regular traffic patterns, which are insufficient when sudden disruptions occur due to events. For example, regression models may not account for the sudden influx of vehicles or changes in driver behavior during major events, leading to inaccurate predictions (Wang and Chen, 2018).

In contrast, modern techniques such as machine learning and deep learning offer promising alternatives by leveraging vast amounts of data to improve prediction accuracy. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have shown significant potential in modeling complex temporal dependencies in traffic data (Zhao et al., 2017; Rasaizadi et al., 2021; Shaygan et al., 2022). Studies have demonstrated the effectiveness of integrating social media and traffic sensor data to enhance real-time traffic prediction (Zhang et al., 2018).

However, despite these advances, the specific impact of different types and sizes of social events on traffic speed remains underexplored. Most existing studies focus on general traffic prediction without differentiating between event types and their varying scales (Giuliano and Lu, 2021; Jin et al., 2023b). Although extensive research has been conducted on traffic prediction during social events, we have not found any studies proposing a general model capable of predicting the impact of different event types with varying numbers of attendees, highlighting the need for models that can generalize across various event types and sizes.

Addressing this gap, our study proposes an advanced Event-Aware Long Short-Term Memory (EA-LSTM) model that uniquely considers the differentiated impact of various types and sizes of social events on traffic speed. By integrating probe vehicle data with detailed event features, the EA-LSTM model aims to provide a robust and scalable solution for predicting traffic speed during social events, ultimately improving urban traffic management.

3.2 Literature Review

Recent studies have explored the impact of social events on urban traffic using various approaches. For instance, (Zhang et al., 2021) examined traffic characteristics during public holidays, proposing a hybrid speed prediction approach that combines support vector regression and historical averages. This method significantly improved prediction accuracy but was limited to public holidays, necessitating a broader approach for other event types. Similarly, (Bejarano-Luque et al., 2021) used deep learning to estimate the impact of social events on traffic demand, integrating traffic data collected from Twitter. While effective, this approach introduced potential biases and data quality issues inherent in social media data.

In addition, (Deng et al., 2022) proposed a graph convolutional adversarial network to detect spatio-temporal anomalies in traffic data, which is particularly useful for identifying irregular patterns caused by social events. However, the real-time processing capabilities and computational complexity of the model remain challenging. Likewise, (Giuliano and Lu, 2021) focused on the traffic impacts of major events at a single venue, using traditional regression and random forest models. Their findings highlighted the varying impacts of different events on local traffic, but were limited in generalizability. Furthermore, (Essien et al., 2021) developed a deep-learning model for urban traffic flow prediction, integrating traffic events mined from Twitter. This approach demonstrated the value of social media data in capturing real-time event impacts but faced challenges related to data reliability and integration. Similarly, (Fu and Liu, 2022) presented a spatial-temporal convolutional model for predicting urban crowd density using mobile-phone signaling data, highlighting its potential for traffic management during large gatherings while raising privacy concerns.

Moreover, (Peng et al., 2024) proposed a unified spatial-temporal neighbor attention network for dynamic traffic prediction, significantly improving accuracy during social events but requiring extensive training data and high computational power. In the same context, (Shin and Lee, 2020) addressed missing temporal and spatial data challenges by using LSTM networks, enhancing prediction accuracy but demanding resource-intensive data preprocessing. Additionally, (Kong et al., 2023) introduced a Bayesian network model for forecasting traffic congestion during large-scale events, incorporating historical data and real-time sensor inputs to achieve high prediction accuracy. However, this approach necessitates substantial historical data and real-time monitoring systems. Similarly, (Li and Wang, 2021) explored reinforcement learning algorithms for adaptive traffic signal control during social events, dynamically adjusting signal timings based on real-time traffic conditions but facing practical implementation challenges due to model complexity.

In recent years, there has been a surge in using advanced deep learning models to improve traffic prediction during social events. For example, (Ren et al., 2022) developed a hybrid deep learning model combining Convolutional Neural Networks (CNNs) and LSTMs to predict traffic flow during a special event. Their model effectively captured both spatial and temporal features, resulting in improved prediction accuracy. However, the approach required extensive computational resources due to the complexity of the model. Similarly, (Harrou et al., 2024) proposed a Transformer-based approach for traffic prediction, leveraging self-attention mechanisms to model long-range dependencies in traffic data. This method showed improved performance over traditional LSTM models, particularly during irregular traffic patterns caused by events. Nevertheless, the Transformer's complexity and need for large datasets may limit its practical application. Also, (Khan et al., 2023) introduced a multi-view learning framework that integrates data from various sources, including social media, weather reports, and event schedules, to enhance traffic prediction accuracy. Their approach demonstrated that incorporating diverse data types can significantly improve model performance during social events, although data integration and processing remain challenging.

Taken together, the studies in (Yin et al., 2023; Jeong et al., 2023; Chen et al., 2024; Zhang et al., 2023) highlight significant advancements in the field of traffic congestion prediction during social events. They employ state-of-the-art methods—ranging from hybrid LSTM-CNN and transfer learning to spatiotemporal graph neural networks and real-time data fusion—that demonstrate improved accuracy in forecasting traffic flow under various conditions. Despite these progressions, however, they also underscore key limitations related to data quality, computational complexity, and the generalizability of models across different contexts and transportation networks. Meanwhile, few existing studies (Jin et al., 2023a; Alevizos et al., 2017) have focused on general event impacts without differentiating between event types and sizes. This broader approach overlooks the substantial variations in traffic disruptions that can arise from distinct categories of events—such as sports, concerts, or festivals—each of which may exhibit unique travel demand patterns, temporal distributions, and traveler behaviors. As a result, the current literature lacks a nuanced understanding of how multiple event attributes collectively influence traffic dynamics.

Collectively, these studies underscore the importance of accurately modeling the differentiated impacts of various social events on urban traffic, yet few have examined how event types and scales interact to influence congestion. Our research addresses this gap by systematically analyzing how various event types (e.g., sport, concert) and sizes affect traffic speed, thereby highlighting both the significance of specialized modeling and the need for a more granular framework in event-based congestion studies. By incorporating more comprehensive datasets and refining modeling approaches to accommodate event-specific characteristics, we aim to improve predictive accuracy and bolster the applicability of these findings in diverse real-world contexts. Building on these insights, the primary objective of this research is to develop an Event-Aware Long Short-Term Memory (EA-LSTM) framework that integrates diverse data sources and advanced modeling techniques. By capturing both temporal and spatial dependencies—while also differentiating between event types and sizes—this model aspires to enhance prediction accuracy and provide actionable insights for more effective traffic management decisions.

3.3 Methodology

In this study, we propose an advanced Event-Aware Long Short-Term Memory (EA-LSTM) model to predict traffic speed by integrating probe vehicle data with detailed event features. The problem at hand is accurately predicting traffic speed during various social events, which poses significant challenges due to the dynamic and non-linear nature of traffic patterns influenced by these social events. By incorporating historical traffic data, temporal features, and social event data into a sophisticated neural network architecture, our model aims to provide a robust and scalable solution to enhance traffic management and urban planning. This proposed methodology section covers data collection, feature engineering, model architecture for the EA-LSTM model. Each subsection includes detailed descriptions and relevant equations, ensuring that the methodology is clear and thorough. The flowchart in Figure 3.1 provides an overview of the proposed model architecture and its components.

3.3.1 EA-LSTM Model Architecture

To improve prediction accuracy and reliability, the model leverages bidirectional processing, self-attention mechanisms, and hierarchical structures, enabling it to effectively capture complex spatiotemporal dependencies in traffic patterns. The model operates with a minimum input of one hour of traffic data to predict traffic speed for the subsequent hour, ensuring real-time adaptability and responsiveness. This study primarily focuses on assessing the impact of social events on the Hamilton traffic network.

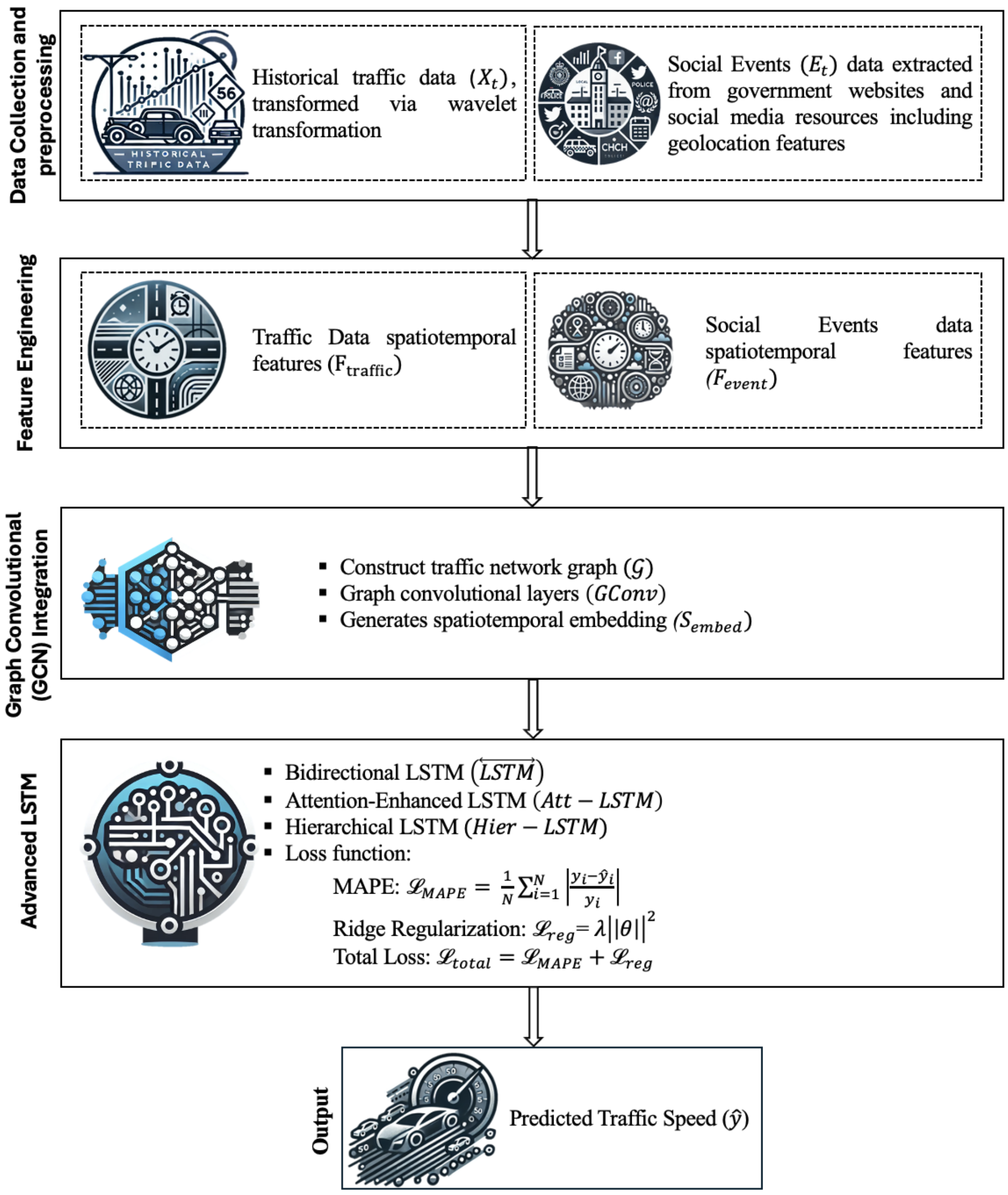
Data Collection and preprocessing
HISTORICAL TRIPIC DATA
56
Historical traffic data (X_t), transformed via wavelet transformation
POLICE
CHCH
Social Events (E_t) data extracted from government websites and social media resources including geolocation features
Feature Engineering
Traffic Data spatiotemporal features ($F_{traffic}$)
Social Events data spatiotemporal (F_{event}) features
Graph Convolutional (GCN) Integration
Construct traffic network graph ($\mathcal{G}$)
Graph convolutional layers ($GConv$)
Generates spatiotemporal embedding (S_{embed})
Advanced LSTM
Bidirectional LSTM ($\overleftrightarrow{LSTM}$)
Attention-Enhanced LSTM ($Att - LSTM$)
Hierarchical LSTM ($Hier - LSTM$)
Loss function:
MAPE: $\mathcal{L}_{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
Ridge Regularization: $\mathcal{L}_{reg} = \lambda\left|\left|\theta\right|\right|^2$
Total Loss: $\mathcal{L}_{total} = \mathcal{L}_{MAPE} + \mathcal{L}_{reg}$
Output
Predicted Traffic Speed ($\hat{y}$)

FIGURE 3.1: Flowchart of the proposed EA-LSTM model architecture and its components

The proposed EA-LSTM model relies on two primary data sources: probe vehicle data and social event data. These datasets provide complementary insights into traffic dynamics, capturing both regular traffic patterns and disruptions caused by events.

Probe vehicle data, collected at high temporal resolution, reflects real-time traffic conditions across Hamilton's road network, while social event data offers detailed information on events that impact traffic flow, such as their type, size, location, and timing. To effectively utilize these datasets, they are integrated through geolocation and temporal alignment, ensuring that traffic data corresponds to relevant events within the same spatial and temporal context. This integration forms the foundation for feature engineering, enabling the model to capture the multifaceted impacts of social events on traffic.

3.3.2 Feature Engineering

Effective feature engineering is crucial for capturing the underlying patterns in traffic data and the influence of social events. Our feature selection was guided by domain expertise and previous research demonstrating the significance of specific variables in traffic prediction (Lv et al., 2014; Ali et al., 2022).

3.3.2.1 Historical Traffic Data

Extracting traffic patterns from historical data helps establish baseline conditions, enabling the model to identify anomalies and deviations caused by events. This historical context provides a reference against which event-influenced variations can be measured. Time-of-day, day-of-week, and seasonality indicators capture recurring temporal variations in traffic. These features are especially important for modeling how traffic patterns vary over daily, weekly, and seasonal cycles, thereby improving predictive accuracy. Specifically, sine and cosine transformations are used to encode the cyclical nature of time (e.g., 24-hour cycle):

$$\begin{aligned} \mathrm{t}_{\mathrm{sin}} &= \sin\left(2\pi\frac{\mathrm{t}}{T}\right), \\ \mathrm{t}_{\mathrm{cos}} &= \cos\left(2\pi\frac{\mathrm{t}}{T}\right), \end{aligned} \tag{3.1}$$

where t is the timestamp and T is the period (e.g., 24 hours). Additionally, one-hot encoding was used to represent each day of the week, ensuring that no unintended ordinal relationships are introduced among the days.

3.3.2.2 Wavelet Transformation

To capture both short-term fluctuations and long-term trends, the traffic data were decomposed using the Discrete Wavelet Transform (DWT) with the Daubechies 4 (db4) wavelet function (Mallat, 1999). This decomposition aids in isolating high-frequency variations (e.g., transient congestion) from low-frequency patterns (e.g., daily commuting rhythms), thereby enhancing the model's ability to identify complex trends.

$$\begin{aligned} A_{j,k} &= \sum_t x(t)\, \phi_{j,k}(t), \\ D_{j,k} &= \sum_t x(t)\, \psi_{j,k}(t), \end{aligned} \tag{3.2}$$

where $A_{j,k}$ and $D_{j,k}$ denote approximation and detail coefficients at scale j and position k, $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are the scaling and wavelet functions, and $x(t)$ is the original traffic speed value. Applying the wavelet transform improved the Signal-to-Noise Ratio (SNR) by 12.1%, resulting in more robust feature extraction.

3.3.2.3 Social Event Features

Quantifying the impact of events allows the model to differentiate between varying levels of disruption. We encoded event characteristics including:

- **Event Type Encoding**: One-hot encoding for categories such as sports events, concerts, festivals, and run/bike/walk events.
- **Number of Attendees**: A continuous variable representing the scale of the event.
- **Parking Availability**: A binary feature indicating whether parking facilities are accessible at or near the venue.

These factors enable the model to gauge how different events might affect traffic flow.

3.3.2.4 Proximity Feature

Finally, to account for spatial impacts, the distance between each event location and affected road segments is calculated using the Haversine formula (Sinnott, 1984):

$$d = 2r \, \arcsin\left(\sqrt{\sin^2\Bigl(\frac{\phi_2 - \phi_1}{2}\Bigr) + \cos(\phi_1)\cos(\phi_2)\sin^2\Bigl(\frac{\lambda_2 - \lambda_1}{2}\Bigr)} \right), \tag{3.3}$$

where ϕ and λ are the latitudes and longitudes of the two points, and r is the Earth's radius. By incorporating proximity features, the model is better positioned to predict localized disruptions and accurately reflect the spatial reach of each event.

3.3.3 Model Architecture

The EA-LSTM model incorporates both spatial and temporal dependencies by integrating Graph Convolutional Networks (GCNs) and advanced LSTM architectures. The architecture of the model includes two sets of data inputs: historical traffic data (including temporal and spatial features) and social event data. These inputs are processed through advanced feature engineering techniques and fed into a deep learning model.

3.3.3.1 Graph Convolutional Networks (GCNs)

GCNs model the spatial relationships between road segments and events, creating spatiotemporal embeddings that capture both the event influence and the geographical context. This approach helps understand how events impact traffic in different locations, enabling the model to learn complex spatial dependencies in the traffic network. The GCN operation is defined as (Schlichtkrull et al., 2018):

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right), \tag{3.4}$$

where:

- $H^{(l)}$ is the input feature matrix at layer l,
- $\tilde{A} = A + I$ is the adjacency matrix with added self-connections,
- $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$,
- $W^{(l)}$ is the trainable weight matrix,
- σ is the activation function (e.g., ReLU).

The term $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ ensures proper normalization of the adjacency matrix, facilitating stable learning.

3.3.3.2 Bidirectional LSTM

Implementing Bidirectional LSTM layers captures dependencies in both forward and backward directions, enhancing the understanding of temporal patterns in the traffic data. This allows the model to consider both past and future context when making predictions. Bidirectional processing improves the model's ability to capture relationships across time.

$$\begin{aligned}
\overrightarrow{h_t} &= \text{LSTM}(x_t, \overrightarrow{h_{t-1}}) \\
\overleftarrow{h_t} &= \text{LSTM}(x_t, \overleftarrow{h_{t+1}}) \\
h_t &= [\overrightarrow{h_t}; \overleftarrow{h_t}]
\end{aligned} \tag{3.5}$$

where x_t is the input at time step t, $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the hidden states of the LSTM at time step t in the forward and backward directions, respectively. The hidden states are concatenated to form the final hidden state h_t at time step t.

3.3.3.3 Attention-Enhanced LSTM

Integrating a self-attention mechanism within the LSTM layers allows the model to focus on different parts of the input sequence dynamically, improving the learning of long-term dependencies. The self-attention mechanism helps the model prioritize important information, leading to more accurate predictions. Self-attention enables the model to weigh the relevance of different time steps dynamically.

$$\begin{aligned} e_t &= \tanh(W_h h_t + W_s s_{t-1}) \\ \alpha_t &= \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)} \\ s_t &= \sum_{i=1}^{T} \alpha_i h_i \end{aligned} \tag{3.6}$$

where h_t is the hidden state at time step t, W_h and W_s are weight matrices, s_{t-1} is the previous state, and e_t is the energy score calculated using the tanh activation function. α_t represents the attention weights obtained by applying the softmax function to the energy scores. Finally, s_t is the context vector computed as a weighted sum of the hidden states h_i using the attention weights α_i.

3.3.3.4 Hierarchical LSTM

Creating a hierarchical LSTM architecture allows the first layer to process raw traffic data, while subsequent layers process aggregated and higher-level features. This multi-scale temporal learning approach enables the model to capture both fine-grained and coarse-grained patterns in the traffic data. Hierarchical LSTM structures enhance the model's ability to understand traffic patterns at different levels of abstraction.

$$\begin{aligned} h_t^{(1)} &= \text{LSTM}^{(1)}(x_t, h_{t-1}^{(1)}) \\ h_t^{(2)} &= \text{LSTM}^{(2)}(h_t^{(1)}, h_{t-1}^{(2)}) \\ h_t^{(3)} &= \text{LSTM}^{(3)}(h_t^{(2)}, h_{t-1}^{(3)}) \end{aligned} \tag{3.7}$$

where $h_t^{(1)}$, $h_t^{(2)}$, and $h_t^{(3)}$ are the hidden states at time step t for the first, second, and third LSTM layers, respectively. x_t is the input at time step t, and $h_{t-1}^{(1)}$, $h_{t-1}^{(2)}$, and $h_{t-1}^{(3)}$ are the hidden states from the previous time step for each respective layer.

Each LSTM layer processes the output from the previous layer, capturing hierarchical temporal features.

3.4 Experimental Study

In this section, we present the case study designed to validate the effectiveness of the proposed EA-LSTM model in predicting traffic speed during various social events. The experimental study includes a detailed description of the datasets used, data preprocessing steps, model setup, training procedures, and evaluation metrics.

3.4.1 Traffic Data

Probe vehicle data is collected through devices installed in vehicles, such as GPS units, that continuously record the vehicle's position, speed, and other relevant metrics as the vehicle travels through the road network. This data provides a comprehensive and granular view of traffic conditions in real-time. We used the dataset recorded by HERE Technology Inc., which covers 4,788 road segments in Hamilton, Ontario. The dataset captures detailed traffic information, allowing for precise analysis of traffic patterns and congestion. Figure 2 illustrates the traffic network of Hamilton, which serves as the testbed for this study. This extensive network enables a robust evaluation of the proposed EA-LSTM model across different traffic conditions and event scenarios.

The dataset spans from October 1, 2022, to October 1, 2023, and includes data recorded at 5-minute intervals throughout this period. This high-frequency data collection ensures a robust and detailed dataset for analysis. Additionally, the dataset includes variables such as average traffic speed, free flow speed, and road coordinates, which are critical for understanding traffic dynamics. To enhance the analysis, wavelet transforms were applied to the traffic data, enabling the capture of both short-term variations and long-term trends. Furthermore, to mitigate biases from overrepresented vehicle types, such as commercial fleets, the dataset was filtered to include only passenger cars, ensuring it accurately reflects general traffic conditions.

The dataset includes the following variables:

- Average Traffic Speed (S): The mean speed of vehicles on each road segment during each 5-minute interval.
- Free Flow Speed (FFS): The average speed under optimal traffic conditions, used as a reference to identify potential congestion when the observed speed deviates significantly.

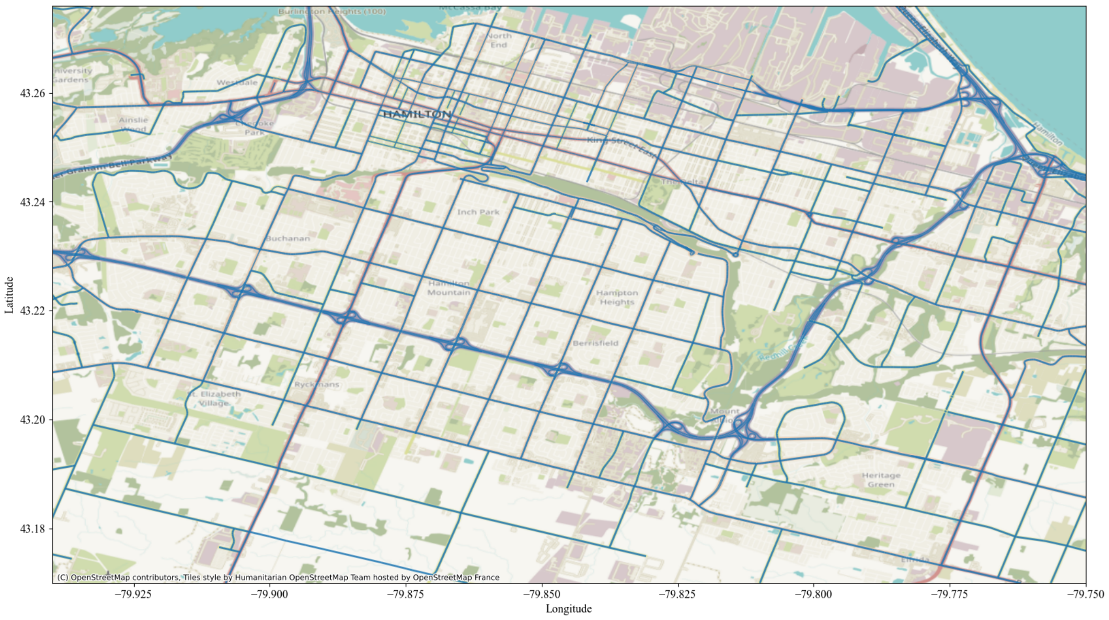

FIGURE 3.2: Traffic map of the experimental study (Hamilton, Ontario)

- Road Attributes: Number of lanes, road type (e.g., highway, arterial, local road), and geographical coordinates (latitude and longitude) of each road segment.

3.4.2 Social Events

The dataset used in this study for social events was meticulously extracted manually from various sources, including the City of Hamilton's official website, Facebook, X (Twitter), Eventbrite, and local news outlets. This comprehensive approach ensured the collection of detailed information about a wide range of events within the city.

The extracted dataset spans from October 1, 2022, to October 1, 2023, covering a total of 561 events across 76 locations in Hamilton. Events were geolocated and matched with corresponding road segments to assess their spatial impact on traffic conditions. This spatial alignment ensured that the influence of events was correctly associated with the affected areas in the traffic network. The dataset includes the following variables:

- Event Name and Type: Categorized into sports events, concerts, festivals/public gatherings, and run/bike/walk events.

- Event Schedule: Start time, end time, and duration of each event.

- Event Location: Geographical coordinates (latitude and longitude) of event venues.

- Attendance: Number of attendees recorded, providing a measure of event scale.
- Parking Availability: Information on parking facilities at or near the event venues.

Event features such as attendance, type, and location were analyzed to quantify their impact on traffic flow, allowing the model to differentiate between events with varying levels of disruption. Social media data, which can introduce biases due to varying levels of user engagement and reporting accuracy, was cross-verified with official announcements and event organizers' websites to ensure reliability. Events lacking sufficient verification were excluded from the dataset, and efforts were made to include all significant events during the study period to minimize selection bias. The dataset's statistical summary is presented in Table 4.1, providing an overview of the frequency, the mean number of attendees, and the standard deviation of attendees for each type of event.

Table 3.1: Statistical Analysis of Social Events

Event	**Frequency**	**Average Number of Attendees**	**Attendees Std. Dev.**
Sport	210	1471	403
Concert	231	849	356
Festival/Public Gathering	82	1231	518
Run/Bike/Walk	38	741	273
Total	**561**	**1093**	**379**

3.4.3 Data Preprocessing

The preprocessing of traffic and social event data involved several steps to prepare the datasets for modeling. First, features were normalized to have zero mean and unit variance, ensuring equal contributions to the training process and preventing features with larger scales from dominating the model. Wavelet transformation was applied to the traffic data to decompose time-series patterns into multiple frequency components. Using the Daubechies (db4) wavelet, the data was analyzed at different time and frequency scales, effectively capturing both short-term variations and long-term trends. This transformation improved the Signal-to-Noise Ratio (SNR) by 12.1%, enhancing feature extraction and denoising. The datasets were then divided into training (70%), validation (15%), and testing (15%) subsets. This split supports robust model development by enabling hyperparameter tuning on the validation set and unbiased performance evaluation on the test set.

3.4.4 Model Setup

The proposed EA-LSTM model consists of multiple layers, including bidirectional LSTM, attention-enhanced LSTM, and hierarchical LSTM, as well as graph convolutional (GCN) layers to capture spatial dependencies. Combining these components enables the model to learn complex temporal and spatial patterns, which is essential for accurately predicting traffic speed during social events. A systematic grid search was conducted to tune key hyperparameters, such as the number of layers, units per layer, dropout rates, and attention heads. Each combination was evaluated on the validation set to identify the configuration yielding the best performance. The optimal settings are summarized in Table 4.2.

The final model was trained using the Adam optimizer (learning rate = 0.001, batch size = 64) for 500 epochs to predict traffic speed one hour ahead. During training, the validation loss was monitored to mitigate overfitting and ensure robust generalization. Figure 3.3 shows the evolution of training and validation loss over 300 epochs, illustrating the convergence behavior and indicating that the model effectively learns from the data.

Table 3.2: Grid Search Hyperparameter Tuning Results

Hyperparameter	**Values Tested**	**Optimal Value**
Bidirectional LSTM		
Number of Layers	1, 2, 3	2
Units per Layer	20, 30, 40, 50, 60, 70, 80, 90, 100	60
Dropout Rate	0.1, 0.2, 0.3	0.1
Attention-Enhanced LSTM		
Attention Heads	4, 6, 8, 10, 12	6
Attention Dropout Rate	0.1, 0.2, 0.3	0.2
Hierarchical LSTM		
Number of Layers	1, 2, 3	3
Units per Layer	20, 30, 40, 50, 60, 70, 80, 90, 100	40
Dropout Rate	0.1, 0.2, 0.3	0.1

3.4.5 Training and Evaluation

The models were trained using the Adam optimizer with a Mean Absolute Percentage Error (MAPE) loss function. The Adam optimizer was chosen due to its adaptive learning rate capabilities, which help in faster convergence and improved performance. The initial learning rate was set to 0.001, and the model was trained for 300 epochs with a batch size of 64. Early stopping was implemented to prevent overfitting, with the

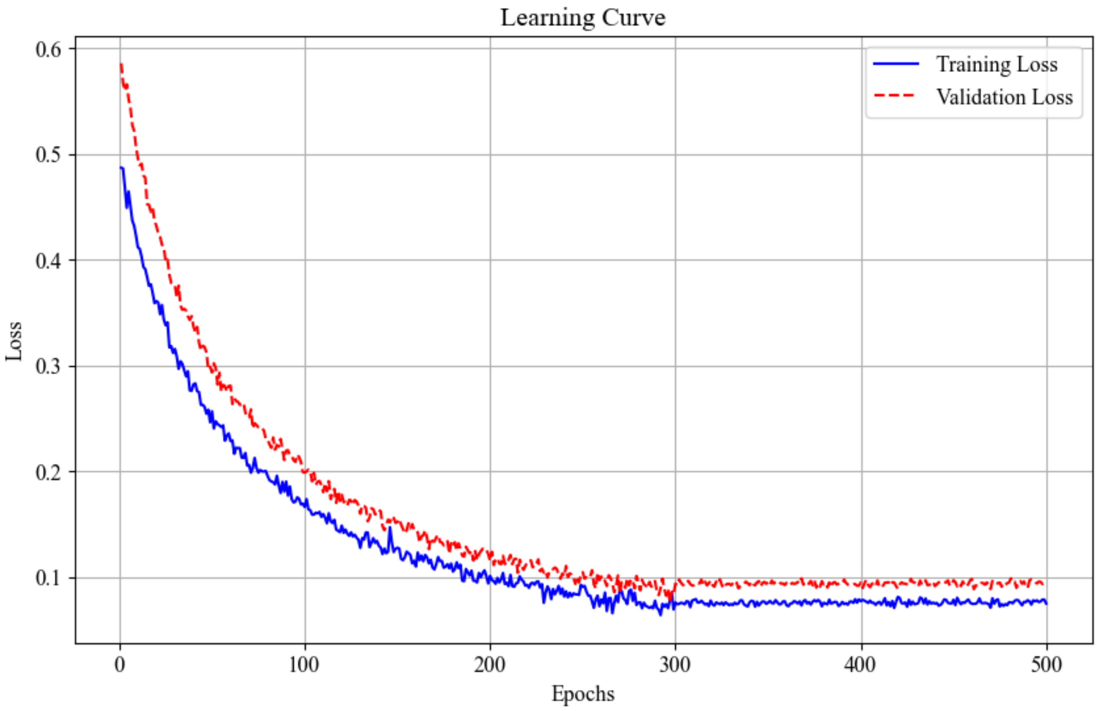

FIGURE 3.3: Training and validation loss for the EA-LSTM model over 300 epochs

patience of 10 epochs, meaning training would halt if there was no improvement in validation loss for 10 consecutive epochs. For evaluation, we used MAPE as the primary metric which is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3.8}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value. MAPE provides an intuitive percentage error, making it easier to interpret the performance of the model across different scales of traffic speed.

A 5-fold cross-validation was performed to ensure the robustness and generalizability of the model. The five results from the folds are then averaged to produce a single estimation. This technique helps in mitigating overfitting and provides a more reliable measure of model performance. During each fold of cross-validation, the model was trained and evaluated, and the MAPE was recorded. The average MAPE across the five folds was calculated to provide a comprehensive evaluation metric. This process ensures that the model's performance is consistent and not dependent on a specific subset of the data, thereby enhancing its robustness and generalizability.

3.5 Results

In this section, we present the performance of the proposed EA-LSTM model in predicting traffic speed during various social events. We analyze the performance across different scenarios, provide a detailed temporal and spatial analysis, and conduct an error analysis to understand the model's limitations. The overall performance of the EA-LSTM model was evaluated using MAPE as the primary metric. The average MAPE across all cross-validation folds was 3.4%. The model's performance was analyzed in relation to the proximity to the event center. Table 3.3 summarizes the MAPE for different proximity ranges for the next one-hour prediction horizon.

Table 3.3: Model Performance Across Different Scenarios

Scenario	**Overall MAPE (%)**	**Weekdays MAPE (%)**	**Weekends MAPE (%)**
Event Center (1km Proximity)	13.8	15.3	10.6
Event Center (2km Proximity)	9.1	10.4	8.1
Event Center (3km Proximity)	8.6	9.5	7.9
Event Center (4km Proximity)	6.7	7.4	6.0
Event Center (5km Proximity)	4.2	5.9	3.9
Event Center (10km Proximity)	3.5	5.2	3.3
Entire Hamilton Network	3.4	4.8	3.1

The analysis indicates that the model's performance decreases as the proximity to the event center decreases, with higher MAPE values observed closer to the event center. This trend reflects the increased complexity of traffic patterns in areas near event centers, where disruptions are more experienced. Interestingly, the model performed better during weekends across all proximity ranges, likely due to lighter traffic and fewer variations compared to weekdays.

The model's performance was further analyzed across different scenarios, including sports events, Concerts, festival/public Gathering, and run/bike/walk. Table 3.4 presents the MAPE for each scenario for the next one-hour prediction horizon within the 1 km proximity of the event locations. The EA-LSTM model demonstrated robust performance across various scenarios, with particularly low MAPE for concerts and run/bike/walk. The higher MAPE values for sports events and festivals/public gatherings indicate that these events cause more significant disruptions to traffic flow, which are more challenging to predict accurately. Conversely, concerts and run/bike/walk events exhibit lower MAPE values, suggesting that traffic patterns during these events are more predictable and less disruptive.

Table 3.4: Scenario-Specific Performance Metrics

Scenario	**Overall MAPE (%)**	**Weekday MAPE (%)**	**Weekend MAPE (%)**
Sports Events	14.9	16.0	13.8
Festival/Public Gathering	12.3	12.9	11.8
Run/Bike/Walk	10.4	11.0	9.9
Concerts	8.8	9.0	8.4

Overall, the model's ability to capture the impact of social events on traffic speed is evident in these results. The differences in performance between weekdays and weekends also highlight the variations in traffic behavior, with the model generally performing better on weekends across most scenarios. This is likely due to reduced overall traffic volume and fewer interactions with routine commuter traffic on weekends. The spatial performance of the model was analyzed by evaluating the MAPE across different regions and road segments in Hamilton. Figure 3.4 presents a heatmap of model performance across the study area. The performance of the model in downtown Hamilton was compared with areas outside the downtown region. The analysis revealed that the model performed better outside downtown, with lower MAPE values. Specifically, the MAPE for downtown Hamilton was 7.1%, whereas the MAPE for areas outside downtown was 2.9%. This indicates that the model is more accurate in predicting traffic speed outside the downtown area, potentially due to less complex traffic patterns and fewer disruptions compared to the downtown region. The heatmap indicates that the model performs well across most regions, with higher errors observed in areas with complex traffic patterns.

The following figures illustrate the impact of different types and sizes of social events on traffic speed predictions in Hamilton, focusing on two sport matches and one concert. These examples demonstrate how traffic patterns deviate from regular conditions due to social events, and they highlight the effectiveness of the EA-LSTM model in predicting these disruptions.

In Figure 3.5, we observe traffic speeds during a sport match with 1,500 attendees. The model accurately captures the traffic slowdown triggered by the event, particularly during peak congestion periods near the start and end of the game. This demonstrates the model's capacity to handle moderately sized sporting events. Figure 3.6 presents a scenario involving a larger crowd of 5,000 attendees. Here, the observed traffic speeds show a sharper decline as the match approaches, reflecting a more pronounced deviation from baseline conditions. Although the model continues to perform well, the increased complexity introduced by a larger audience is evident in the slightly wider gap between observed and predicted speeds during the most congested times. Shifting to a different event type, Figure 3.7 depicts a concert in downtown Hamilton with 3,000 attendees.

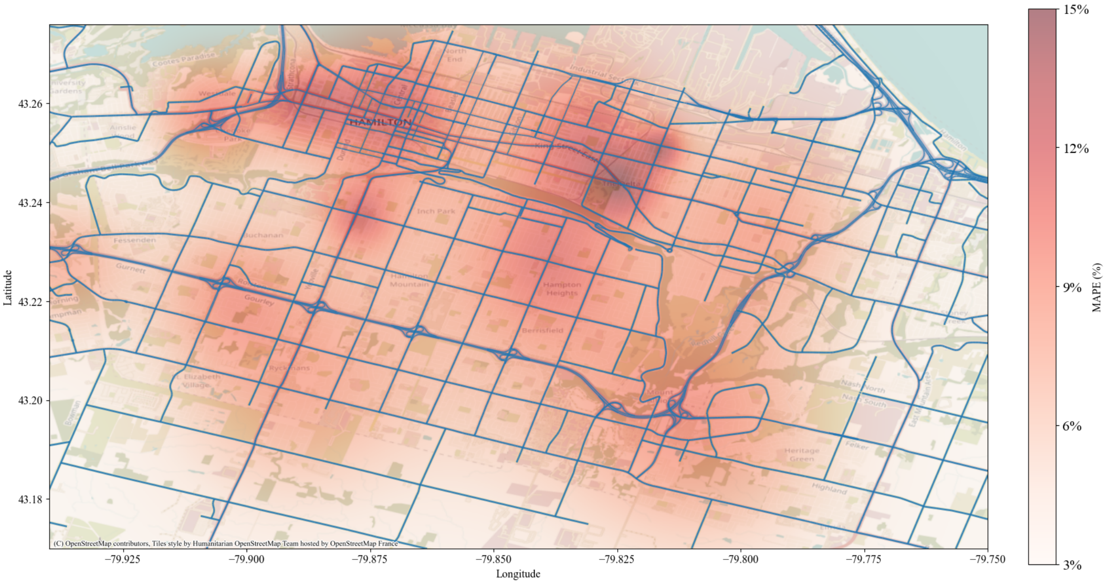

FIGURE 3.4: Spatial Analysis of Model Performance

Concerts often generate unique traffic dynamics, especially in urban cores, and can produce more erratic patterns than sporting events. The model successfully predicts a significant drop in traffic speeds, although variations in observed traffic highlight challenges posed by factors such as staggered arrival and departure times.

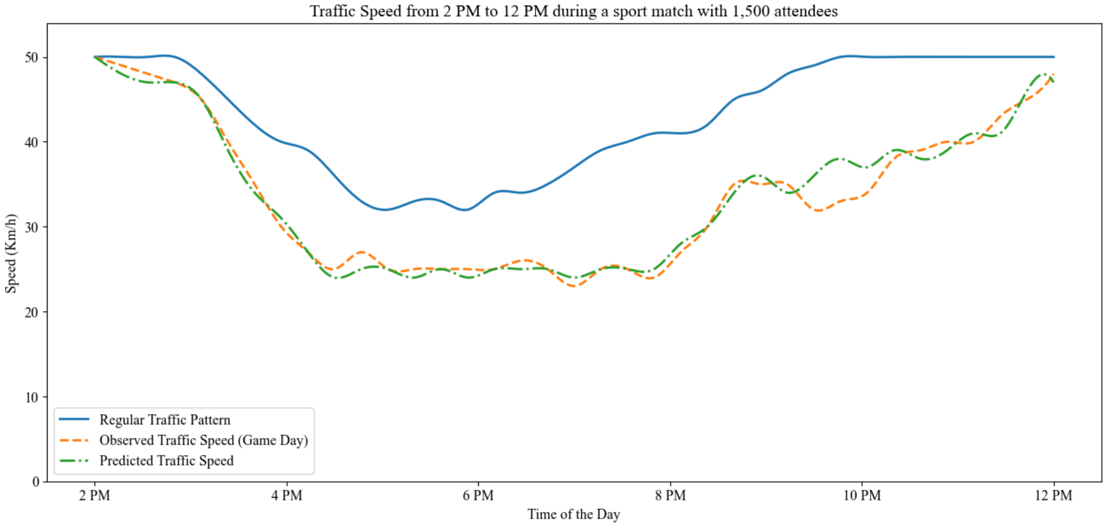

FIGURE 3.5: Link Speed from 2 PM to 12 PM during a sport match with 1,500 attendees

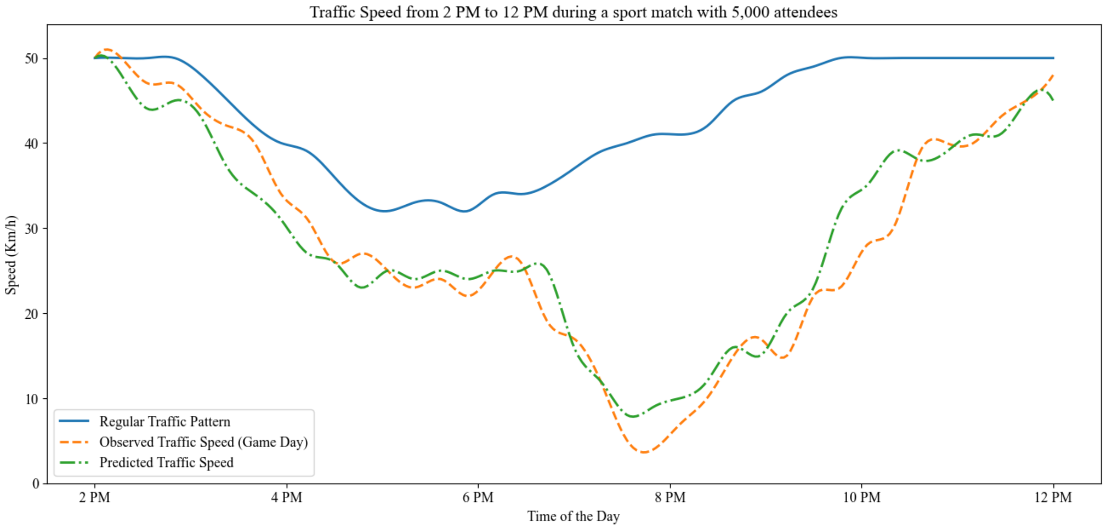

FIGURE 3.6: Link Speed from 2 PM to 12 PM during a sport match with 5,000 attendees

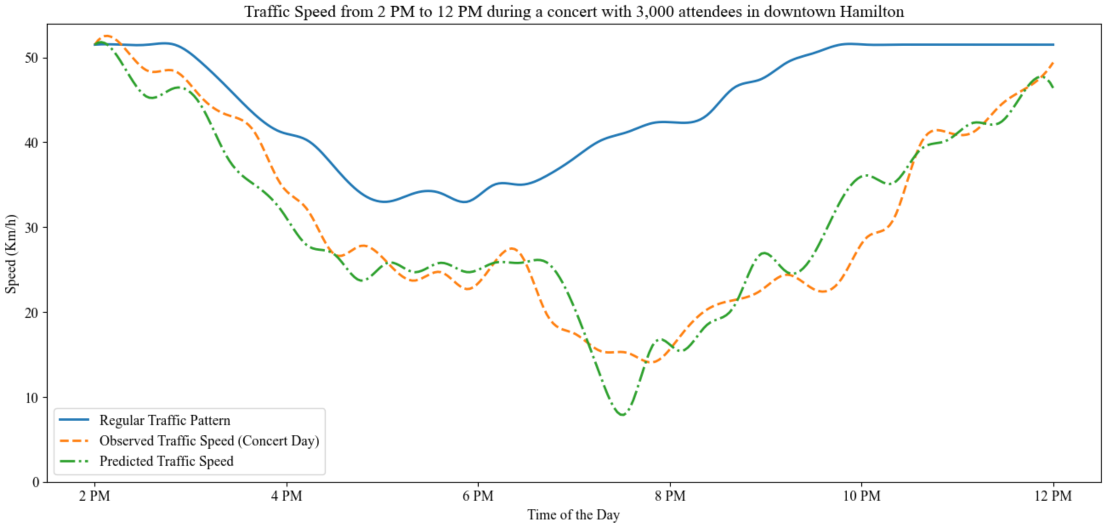

FIGURE 3.7: Link Speed from 2 PM to 12 PM during a concert with 3,000 attendees

These examples underscore the importance of accounting for both event type and scale in traffic speed prediction. Without these considerations, forecasts may fail to represent real-world patterns, limiting their utility for traffic management. By differentiating across diverse events, the EA-LSTM model delivers more accurate and actionable predictions, offering valuable insights for urban planners and traffic managers. Moreover, Figure 3.8 illustrates MAPE values of the proposed model in relation to the number of attendees for different event types, specifically sports events, festivals/public gatherings, run/bike/walk events, and concerts. It is evident that the model's performance varies significantly across these event types and attendee sizes. For sports events, the MAPE increases steeply as the number of attendees rises, indicating higher prediction errors for larger sports events. This trend suggests that sports events, which often attract large crowds and significantly impact traffic flow, present more challenging scenarios for accurate traffic speed prediction.

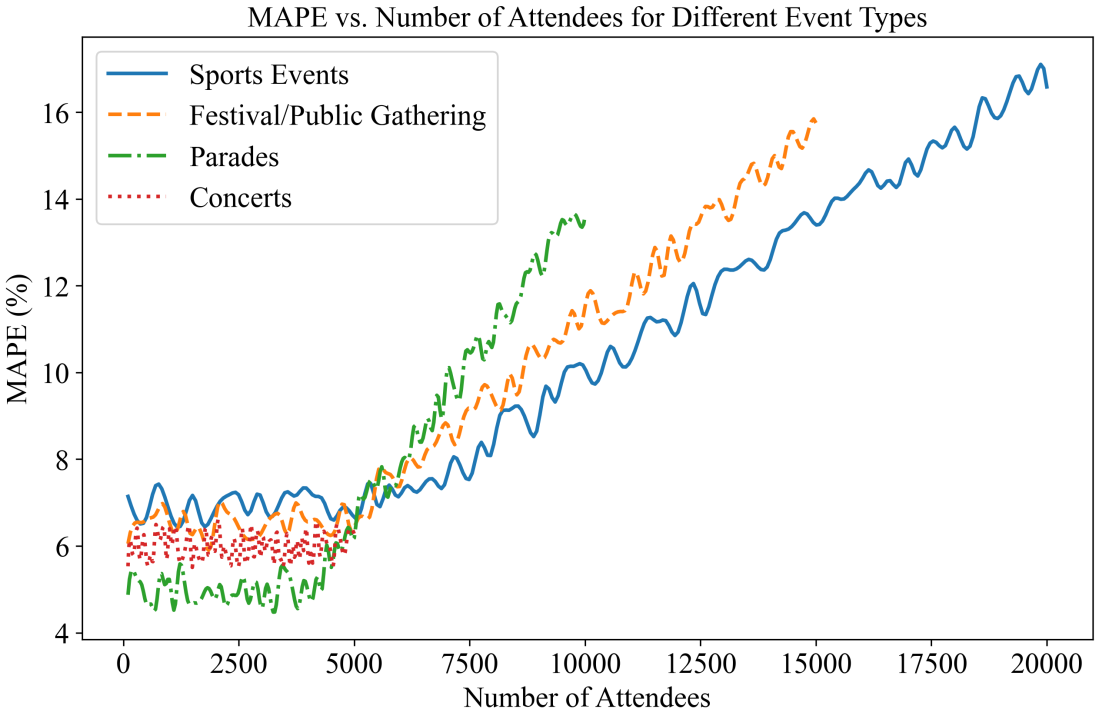

FIGURE 3.8: Traffic Speed Predictions performance for various number of attendees

In contrast, the festivals/public gatherings also show an increase in MAPE with the number of attendees, but the rate of increase is less steep compared to sports events. This indicates that while larger festivals and public gatherings still pose challenges, the model handles these scenarios somewhat better than large sports events. The run/bike/walk events follow a similar pattern, with MAPE increasing as the number of attendees grows, but again, the increase is less pronounced than for sports events. This could be due to the typically smaller and more localized impact of run/bike/walk events on traffic.

Finally, concerts show the lowest MAPE values across all attendee sizes, indicating that the model performs best in predicting traffic speed for concert events. This could be attributed to more predictable and possibly smaller-scale disruptions caused by concerts compared to other event types. Overall, the graph underscores the model's varying performance across different road types and event sizes, highlighting that higher errors are typically observed for larger events and more congested road types. This insight is crucial for understanding the limitations of the model and identifying areas for further improvement.

3.6 Conclusion

In conclusion, this study proposed and evaluated an Event-Aware Long Short-Term Memory (EA-LSTM) model for predicting traffic speed during social events. The EA-LSTM model integrates probe vehicle data with detailed event features and employs advanced features such as GCN, bidirectional LSTM layers, self-attention mechanisms, and hierarchical structures. The model was validated using a one-year probe vehicle dataset from Hamilton, ON, Canada, and historical social events data. The findings indicate that the EA-LSTM model significantly improves traffic speed prediction accuracy, particularly during social events.

The results demonstrate that the model's performance varies across different scenarios and proximity ranges. The model achieved an average MAPE of 3.4% across all cross-validation folds, with higher errors observed closer to event centers. Sports events and festivals/public gatherings caused more significant disruptions to traffic flow, resulting in higher MAPE values. In contrast, concerts and run/bike/walk events exhibited lower MAPE values, suggesting that traffic patterns during these events are more predictable.

Comparing the results of this study with previous research, the EA-LSTM model shows substantial improvements in prediction accuracy and robustness. For instance, Zhang (2023) (Zhang et al., 2021) achieved a MAPE of 7.2% using a hybrid speed prediction approach during a festival, whereas the EA-LSTM model achieved a lower MAPE of 3.5% within the 10km proximity. Similarly, Luque et al. (2021) (Bejarano-Luque et al., 2021) reported a MAPE of 5.6% using deep learning techniques to estimate traffic impact during social events, compared to the EA-LSTM model's superior performance.

This study pioneers the integration of differentiated social event characteristics into traffic speed prediction models. By accounting for both the type and size of events, the proposed model offers a more precise and context-sensitive prediction tool, setting it apart from existing approaches that do not consider these critical factors. The EA-LSTM model addresses the research objectives by enhancing data reliability, reducing computational requirements, and improving generalizability across different urban contexts and event types. By integrating diverse data sources and employing advanced modeling techniques, the model provides a robust and scalable solution for predicting traffic speed during social events.

Future work could focus on enhancing the model's adaptability and scalability to other urban contexts. By refining the methodology for automated extraction and integration of event data from diverse sources, such as social media and city event calendars, the model can be tailored to different cities with varying event characteristics and traffic dynamics. Additionally, incorporating advanced data fusion techniques to assimilate other relevant data streams, such as public transportation usage and real-time weather conditions, could further improve the model's robustness and accuracy. Extending the application of the EA-LSTM model to a variety of metropolitan areas will enable a comprehensive assessment of its generalizability and effectiveness, providing valuable insights for widespread urban traffic management. These insights can assist urban planners and traffic managers in anticipating and mitigating traffic congestion, ultimately contributing to more efficient and sustainable urban mobility.

Bibliography

Alevizos, E., Artikis, A., and Paliouras, G. (2017). Event forecasting with pattern markov chains. In *Proceedings of the 11th ACM International Conference on Distributed and Event-Based Systems*, pages 146–157. ACM.

Ali, A., Zhu, Y., and Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural Networks*, 145:233–247.

Bejarano-Luque, J. L., Toril, M., Fernandez-Navarro, M., Gijon, C., and Luna-Ramirez, S. (2021). A deep-learning model for estimating the impact of social events on traffic demand on a cell basis. *IEEE Access*, 9:71673–71686.

Chen, P., Jiang, X., and Yang, R. (2024). Multi-step freeway traffic speed prediction with spatiotemporal graph neural networks. *Journal of Transportation Engineering, Part A: Systems*, 150(3):04024020.

Deng, W., Zhang, L., and Liu, Y. (2022). Detecting spatiotemporal anomalies in traffic data with graph convolutional adversarial networks. *Transportation Research Part C: Emerging Technologies*, 135:103498.

Essien, A., Petrounias, I., Sampaio, P., and Sampaio, S. (2021). A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*, 24(4):1345–1368.

Fu, X. and Liu, M. (2022). Spatial-temporal convolutional model for urban crowd density prediction using mobile-phone signaling data. *Computers, Environment and Urban Systems*, 95:101828.

Giuliano, G. and Lu, Y. (2021). Analyzing traffic impacts of planned major events. *Transportation Research Record*, 2675(8):432–442.

Harrou, F., Zeroual, A., Kadri, F., and Sun, Y. (2024). Enhancing road traffic flow prediction with improved deep learning using wavelet transforms. *Results in Engineering*, page 102342.

Jeong, S., Wang, S., and Li, M. (2023). Incorporating special events into urban traffic flow forecasting using transfer learning. *Journal of Transportation Engineering, Part A: Systems*, 149(7):04023072.

Jin, G., Liu, L., Li, F., and Huang, J. (2023a). Spatio-temporal graph neural point process for traffic congestion event prediction. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 14268–14276. AAAI Press.

Jin, X., Li, H., and Wang, Y. (2023b). Understanding the differentiated impacts of social events on urban traffic congestion. *Transportation Research Part A: Policy and Practice*, 163:112–126.

Khan, A., Fouda, M. M., Do, D.-T., Almaleh, A., and Rahman, A. U. (2023). Short-term traffic prediction using deep learning long short-term memory: Taxonomy, applications, challenges, and future trends. *IEEE Access*, 11:94371–94391.

Kong, J., Fan, X., Jin, X., Lin, S., and Zuo, M. (2023). A variational bayesian inference-based en-decoder framework for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 25(3):2966–2975.

Li, X. and Wang, C. (2021). Reinforcement learning for adaptive traffic signal control during social events. *Transportation Research Part C: Emerging Technologies*, 130:103312.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2014). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic Press.

Peng, Y., Guo, Y., Hao, R., and Xu, C. (2024). Network traffic prediction with attention-based spatial-temporal graph network. *Computer Networks*, 243:110296.

Rasaizadi, A., Ardestani, A., and Seyedabrishami, S. (2021). Traffic management via traffic parameters prediction by using machine learning algorithms. *International Journal of Human Capital in Urban Management*, 6(1):57–68.

Ren, Y., Jiang, H., Ji, N., and Yu, H. (2022). Tbsm: A traffic burst-sensitive model for short-term prediction under special events. *Knowledge-Based Systems*, 240:108120.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Shaygan, M., Meese, C., Li, W., Zhao, X. G., and Nejad, M. (2022). Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. *Transportation Research Part C: Emerging Technologies*, 145:103921.

Shin, H. and Lee, J. (2020). Predicting traffic congestion with long short-term memory networks. *Journal of Transportation Engineering, Part A: Systems*, 146(3):04020005.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, 68(2):158.

Wang, Y. and Chen, Y. (2018). Limitations of traditional traffic prediction models in the context of social events. *Journal of Transportation Engineering, Part A: Systems*, 144(5):04018025.

Yin, Q., Sinha, S., Zheng, H., and Guan, W. (2023). Short-term traffic flow prediction considering nonrecurrent congestion using a hybrid lstm–cnn approach. *Journal of Transportation Engineering, Part A: Systems*, 149(10):04023100.

Zhang, J., Zheng, Y., and Qi, D. (2018). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1655–1661.

Zhang, W., Yao, R., Du, X., and Ye, J. (2021). Hybrid deep spatio-temporal models for traffic flow prediction on holidays and under adverse weather. *IEEE Access*, 9:157165–157181.

Zhang, X., Gao, K., and Dunn, R. (2023). Real-time event-responsive traffic state estimation and prediction using data fusion. *Journal of Transportation Engineering, Part A: Systems*, 149(11):04023112.

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., and Liu, J. (2017). Lstm network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75.

Chapter 4

Enhancing Traffic Imputation and Prediction Accuracy through Social Event Features and Multi-Stage Deep Learning Models

The content of this chapter is the manuscript submitted under the following citation:

> Ali Ardestani, Hao Yang, Yashar Farid and Seyhan Ucar (Submitted). Enhancing Traffic Imputation and Prediction Accuracy through Social Event Features and Multi-Stage Deep Learning Models. Under review at the *IEEE Transactions on Intelligent Transportation Systems*, June 2025.
>
> Ardestani A. designed the overall framework, implemented the multi-stage imputation and prediction models, conducted the experiments, and drafted the manuscript. Yang H. supervised the research methodology, provided conceptual guidance, and contributed to manuscript revision. Farid Y. assisted with the integration and preprocessing of event and traffic datasets, and contributed to analysis and interpretation. Ucar S. supported the experimental validation and offered technical feedback on the architecture and evaluation protocols. All authors reviewed and approved the final version of the manuscript.

Enhancing Traffic Imputation and Prediction Accuracy through Social Event Features and Multi-Stage Deep Learning Models

Ali Ardestani[1], Hao Yang[1,*], Yashar Farid[2], Seyhan Ucar[2]

[1] Department of Civil Engineering, McMaster University, Hamilton, ON, Canada
[2] Toyota InfoTech Labs, Toyota Motor North America R&D, Mountain View, CA USA
[*]Corresponding author: haoyang@mcmaster.ca

Abstract

Missing traffic data pose significant challenges for Intelligent Transportation Systems, especially during large-scale social events, which dramatically disrupt typical traffic patterns. Traditional imputation techniques often fail under these event-driven anomalies due to their inability to dynamically incorporate contextual event information. This research introduces an integrated, event-aware traffic data imputation and prediction framework, leveraging social event attributes including attendance, event type, timing, and location to enhance accuracy. The proposed approach adopts a robust two-stage imputation strategy, first utilizing Random Forests and subsequently Generative Adversarial Imputation Networks (GAIN) for improved reconstruction of missing values. The completed dataset then feeds into a spatiotemporal prediction model integrating Graph Convolutional Networks (GCNs), Long Short-Term Memory (LSTM) layers, and attention mechanisms explicitly informed by event context. Empirical evaluation using a comprehensive dataset from Hamilton, Ontario, demonstrates significant improvements in both imputation accuracy and traffic prediction performance during events. This integrated approach provides a practical solution for real-time, event-sensitive urban traffic management.

Keywords: *Traffic prediction, missing data imputation, deep learning, social events, graph neural networks, LSTM, GCN*

4.1 Introduction

Metropolitan traffic congestion significantly affects the quality of life and economic productivity in urban areas. Social events such as concerts, sports games, and public gatherings further exacerbate this problem, leading to unpredictable traffic patterns and increased congestion. Accurate prediction of traffic speed during these events is crucial for effective traffic management and urban planning. Traditional approaches, including regression models and statistical analyses, often fail to capture the dynamic and nonlinear nature of traffic flow influenced by social events. These methods typically rely on historical traffic data and assume regular traffic patterns, which are insufficient when sudden disruptions occur due to events. For example, regression models may not account for the sudden influx of vehicles or changes in driver behavior during major events, leading to inaccurate predictions (Wang and Chen, 2018).

This problem is particularly amplified during social events such as concerts, sports games, festivals, and parades, which introduce large, non-recurring fluctuations in traffic demand. These disruptions can invalidate assumptions of temporal regularity and spatial smoothness, posing challenges for both data imputation and traffic forecasting (Tempelmeier et al., 2020). Traditional imputation methods—such as mean substitution, interpolation, or matrix factorization often fail to capture these abrupt, context-specific shifts (Zhang et al., 2024).

Recent advancements in machine learning and deep learning have led to more sophisticated imputation and prediction techniques. Graph Convolutional Networks (GCNs) and Long Short-Term Memory (LSTM) networks are increasingly applied to model spatiotemporal dependencies in traffic systems (Guo et al., 2019; Ardestani et al., 2025; Li et al., 2018b). Moreover, attention mechanisms and event-aware modeling frameworks are gaining popularity for their ability to dynamically adapt to changes in traffic flow caused by external disruptions (Song et al., 2024). Despite this progress, few models explicitly integrate event-related features (e.g., type, location, attendance, timing) into a joint framework for imputation and forecasting.

This paper introduces a novel, event-aware framework that integrates traffic data imputation and prediction using a hybrid architecture. The proposed methodology includes a two-stage imputation strategy: a Random Forest imputation phase followed by Generative Adversarial Imputation Networks (GAIN) if needed. Once the dataset is completed, it is fed into a spatiotemporal prediction model composed of GCN, LSTM, and attention layers, all enhanced by contextual features related to social events. This unified approach is evaluated using a one-year dataset of probe vehicle data and social event records from Hamilton, Ontario, Canada.

The contributions of this study are threefold:

- We develop an integrated framework for missing traffic data imputation and prediction that explicitly incorporates social event features.
- We propose a two-stage imputation strategy that leverages both classical ensemble learning and deep generative modeling.
- We empirically demonstrate that incorporating event-aware information significantly improves both imputation accuracy and traffic prediction during disruptive periods.

The remainder of this paper is structured as follows: Section 4.2 reviews related work on missing data imputation, spatiotemporal traffic modeling, and event-aware prediction. Section 4.3 describes the datasets and preprocessing steps and presents the proposed methodology, including the imputation and prediction models. Section 4.5 discusses experimental results, performance evaluation, and key findings. Finally, Section **??** concludes the paper and outlines future research directions.

4.2 Literature Review

Early approaches to traffic data imputation rely on statistical techniques and domain knowledge. Simple methods like mean or historical average substitution are often used for low missing rates (Smith et al., 2003; Gazis and Liu, 2003). For example, Smith et al. explored filling missing traffic detector readings with historical averages and other straightforward heuristics (Smith et al., 2003). When temporal dynamics are important, time series models such as Kalman filtering and ARIMA have been applied. Gazis and Liu demonstrated that a Kalman filter can effectively estimate and correct missing traffic counts by modeling traffic flow dynamics (Gazis and Liu, 2003). Classic ARIMA models (and seasonal variants) have also been used to predict and replace missing values, though they require stationarity assumptions and may struggle with highly volatile traffic patterns. These statistical imputation methods assume traffic evolves regularly and often perform adequately for isolated missing points or short gaps. They can fail, however, under high missing rates or consecutive missing intervals where temporal correlations are insufficient (Ni and Leonard, 2005). To leverage additional structure, some works introduced multivariate models that incorporate spatial correlations (e.g. neighboring sensor data) when imputing missing values (Ni and Leonard, 2005). For instance, Ni and Leonard employed a Bayesian network with Markov chain Monte Carlo to impute incomplete intelligent transportation systems data, capturing probabilistic relationships among correlated traffic streams (Ni and Leonard, 2005). Overall, statistical imputation techniques (historical averages, interpolation, Kalman filter, ARIMA, Bayesian networks, etc.) laid the groundwork for handling missing traffic data, offering interpretable models but often requiring strong assumptions (e.g. recurring daily patterns or known noise distributions) that limit their flexibility in complex scenarios.

As traffic datasets grew in size and complexity, data-driven machine learning methods emerged to improve imputation accuracy. Unlike fixed statistical models, machine learning approaches can automatically learn patterns from data, including nonlinear and high-dimensional relationships. A variety of methods have been explored. For example, k-nearest neighbors (KNN) algorithms can identify similar traffic patterns from other time periods or locations and use those to fill in missing values, an approach that is simple yet often more robust than global mean imputation. Some studies have refined KNN with local regression, such as Chang extitet al. improved local least-squares imputation which weights nearest neighbors to better preserve local traffic dynamics (Li et al., 2013). Another line of work uses matrix factorization and principal components: Li *et al.* proposed an efficient spatiotemporal imputation by considering traffic data as a matrix and applying Probabilistic PCA, capturing main latent factors while accounting for temporal and spatial dependence (Li et al., 2013). Such methods leverage the low-rank structure of traffic data (e.g. daily or weekly periodicity). More recently, researchers have applied dedicated machine learning models to learn the mapping from observed data to missing values. Sun *et al.* introduced a Bayesian network model (with graphical lasso for structure learning) to estimate missing traffic flow data, which integrates domain knowledge and data to infer likely values for unobserved entries (Sun et al., 2006).

With the rise of deep learning, autoencoder based models have been used for imputation: for instance, Duan *et al.* demonstrated that a stacked denoising autoencoder can learn complex features from incomplete traffic flow data and outperform conventional methods in recovering missing entries (Duan et al., 2016). Similarly, recent studies use neural networks to capture nonlinear temporal patterns; an ensemble of convolutional autoencoders was shown to successfully reconstruct missing traffic data by learning from spatial neighbors and temporal context (Ye et al., 2021). Additionally, researchers have explored spatial interpolation techniques from geostatistics: Bae *et al.* used spatio-temporal co-kriging to borrow strength from nearby sensors in both space and time, treating traffic speed observations as spatial samples with correlation, which significantly improved imputation accuracy on road networks (Bae et al., 2018). Overall, machine learning approaches (including shallow methods like KNN/regression and advanced ones like neural networks) offer greater flexibility than statistical models. They can learn irregular traffic patterns and complex dependencies, yielding more accurate imputations especially under high missing rates or non-recurring missing patterns. However, they typically require sufficient training data and may be less interpretable, prompting a continued balance between statistical insight and data-driven learning in modern imputation research.

In parallel with improving data completeness, significant progress has been made in forecasting traffic conditions using deep learning, which can capture both temporal dynamics and spatial correlations in traffic networks. Early applications of deep learning in traffic focused on time-series prediction at single locations. For example, Lv *et al.* applied a stacked autoencoder to learn generic traffic flow features from “big data,” demonstrating notable improvements in short-term prediction accuracy over traditional ARIMA and neural networks (Lv et al., 2015). Similarly, recurrent neural networks (RNNs) were soon adopted: Ma *et al.* showed that an LSTM network could successfully learn long-term temporal dependencies in traffic speed data, outperforming statistical baselines for highway speed prediction (Ma et al., 2015). These works established the efficacy of deep architectures for modeling the complex, nonlinear temporal behavior of traffic.

Beyond purely temporal models, researchers recognized that incorporating spatial context (interactions between road links or sensor locations) is crucial for network-wide traffic prediction. This led to the development of spatiotemporal models that combine graph-based convolutions with sequence models. A seminal approach is the Diffusion Convolutional Recurrent Neural Network (DCRNN) by Li *et al.*, which integrates graph diffusion convolution operations (to capture spatial propagation of traffic information on a road network graph) with GRU recurrent units for temporal sequencing (Li et al., 2018a). DCRNN demonstrated that modeling the road network as a graph (where nodes are sensors and edges represent traffic flow connectivity) markedly improves multistep traffic flow forecasts, especially during peak periods when upstream-downstream interactions are pronounced. Around the same time, Yu *et al.* proposed the Spatio-Temporal Graph Convolutional Network (STGCN), a purely convolutional approach that applies graph convolutions for spatial features and temporal convolutions for sequential trends (Yu et al., 2018). STGCN achieved fast training and competitive accuracy by avoiding recurrence, and it effectively captured daily rush-hour patterns across an entire city’s sensor network. Building on these ideas, many variants emerged. Zhao *et al.* introduced T-GCN, which couples a graph convolutional network with a gated recurrent unit to jointly model city-scale traffic speed data, showing that the hybrid architecture adapts well to spatial-temporal data and yields lower error than separate CNN or RNN models (Zhao et al., 2019). Wu *et al.* went further by developing Graph WaveNet, a deep architecture that learns an adaptive adjacency matrix for the traffic graph and employs dilated causal convolutions in time, enabling it to capture both global and local traffic dependencies and achieve state-of-the-art forecasting results on highway traffic datasets (Wu et al., 2019). These graph-based neural networks significantly outperform earlier approaches by accounting for how congestion and traffic waves propagate through a network.

Moreover, attention mechanisms and transformers have been incorporated in recent models to improve long-range forecasts. For instance, Guo *et al.* designed an attention-based spatial-temporal graph neural network that learns heterogeneous traffic dynamics (e.g. different day-of-week patterns and road types) and adaptively highlights influential neighbors, leading to more robust predictions under varying traffic regimes (**?**). Similarly, researchers have begun exploring transformer architectures for traffic prediction, which use self-attention to capture long-term temporal correlations and dynamic spatial dependencies across the network (**?**). These advanced deep learning models (GCN-RNN hybrids, temporal CNNs, graph attention networks, etc.) have pushed the frontier of traffic forecasting, enabling accurate predictions even 30–60 minutes ahead under complex conditions. A recurring theme is that explicitly modeling the spatiotemporal structure of traffic—through graph representations of road networks and sequence models or attention for temporal trends—is key to high performance. Surveys of recent work (**?**) highlight that deep models consistently outshine classic methods, especially in large-scale systems, by learning hierarchical features from both spatial neighbors and historical time series. Challenges remain in improving model interpretability and efficiency, but the consensus in 2020–2024 literature is that spatiotemporal deep learning has become the state-of-the-art foundation for traffic prediction.

While deep learning models excel when traffic dynamics are recurrent and training data is abundant, non-recurring events (such as concerts, sporting events, festivals, or major accidents) can cause sudden shifts that purely data-driven models might not anticipate. Hence, a growing body of research focuses on event-aware traffic prediction, integrating information about social events or external factors into modeling. Early efforts in this area augmented statistical models with event indicators. For example, transportation engineers introduced binary “special event” variables into regression or ARIMA models to adjust predictions on event days. A notable data-driven approach by Ni *et al.* leveraged social media data: they extracted features from Twitter (such as tweet volumes and keywords related to a sports game) to predict short-term traffic flow before and after the event, demonstrating that incorporating live social signals improved forecast accuracy compared to using traffic data alone (Ni et al., 2014). This study, which compared multiple regression and machine learning methods with and without social media features, provided early evidence that online crowd information can serve as a proxy for the impact of events (e.g. game attendance, end times) on traffic demand. Subsequent research has extended event-aware modeling using both domain knowledge and advanced algorithms. Yu *et al.* developed a special event-based k-nearest neighbor model tailored for short-term traffic state prediction during event occurrences (**?**). By redefining traffic state distance metrics to account for both typical conditions and the unusual surges caused by events, their SEKNN method outperformed conventional ML and even some deep learning models in event scenarios, highlighting the value of customized predictors when data are scarce (since major events happen infrequently at any one location).

In recent years, deep learning models have been augmented with event awareness as well. Essien *et al.* proposed a deep learning framework that mines traffic related events from Twitter and fuses them with traffic sensor data and weather information (Essien et al., 2021). Using a bi-directional LSTM stacked autoencoder, they showed that including features representing incidents, road closures, and social events (inferred from tweet text and volume) significantly improves multi-step traffic flow predictions, especially for non-recurring congestion spikes. Their results underscore that social sensing of events can provide timely context that traditional sensors do not capture. Moreover, new architectures explicitly incorporate event indicators or “external features” into neural networks. Song *et al.*, for instance, introduce sports event impact factors into a graph-attention transformer model for long-term traffic forecasting (Song et al., 2024). By feeding the network with a schedule of major sports games (treated as external inputs) and learning an attention mechanism that modulates traffic predictions during those periods, their Graph Attention Informer achieved more accurate results for event days in London, compared to baseline models unaware of the events. Across these studies, the consensus is that integrating exogenous event data (be it via manual features, social media mining, or dedicated model components) yields more resilient traffic prediction systems that maintain accuracy during atypical conditions. This is increasingly important as cities leverage smart city data: knowing when concerts, football matches, or parades occur and quantifying their impact allows predictive models to adjust forecasts proactively. The period 2020–2024 has seen event-aware traffic modeling mature from ad-hoc feature engineering to more systematic approaches combining multi-source data. Challenges remain in generalizing models to different types of events and obtaining reliable real-time event data, but the research trend clearly shows the benefit of making traffic prediction models "event-aware" to handle the intrinsic unpredictability of human-driven traffic disruptions.

Despite the advancements across statistical, machine learning, and deep learning methodologies, several critical gaps remain unaddressed. First, while deep spatiotemporal models (e.g., GCNs, LSTMs, and transformers) provide strong performance, their effectiveness diminishes when data is incomplete or disrupted particularly during large-scale, non-recurrent events. Second, although event-aware forecasting has gained traction, few models simultaneously address the missing data problem while also integrating social event context for enhanced prediction. Most current frameworks treat imputation and forecasting as isolated tasks, leading to suboptimal performance when traffic patterns deviate from the norm due to event-driven anomalies. Third, social event features (e.g., attendance, event type, and venue proximity) are rarely embedded as dynamic inputs within spatiotemporal learning pipelines. To bridge these gaps, this study proposes a unified framework that performs robust, two-stage traffic data imputation followed by event-aware traffic speed prediction, using social-event-informed deep learning modules. This integrated approach seeks to improve imputation accuracy under disruption, enhance short-term prediction, and offer a practical path forward for real-time, event-sensitive urban traffic management.

4.3 Methodology

This section introduces the architecture and components of the proposed traffic modeling framework. It provides a step-by-step description of the algorithm, starting with the problem formulation, followed by the imputation strategy, the event-aware prediction module, model evaluation techniques, and the tuning of hyperparameters essential to ensure robust performance.

4.3.1 Proposed Algorithm

The core workflow of the proposed study is illustrated in Figure 3.1, which outlines the complete pipeline from raw traffic data collection to event-aware traffic prediction. This methodology includes a two-stage process: (1) a robust missing data imputation module, and (2) a spatiotemporal deep learning model for prediction enhanced with social event features.

In the imputation phase, the dataset undergoes preprocessing, where outlier filtering, normalization, and encoding of event-related categorical variables are applied. Initial imputation is performed using a Random Forest regressor to generate baseline estimates of missing values, leveraging nonlinear interactions and tree ensemble robustness. For more nuanced correction, Generative Adversarial Imputation Networks (GAIN) are applied to refine these estimates, particularly under high missingness or during complex, event-induced anomalies.

Upon completion of the imputation process, the dataset is passed to the prediction module. First, spatial dependencies among road segments are captured via a Graph Convolutional Network (GCN), which learns how traffic propagates across adjacent links in the road network. These spatial embeddings are fed into a Long Short-Term Memory (LSTM) network that models temporal trends in traffic speed.

A key novelty of the proposed framework is the integration of event-aware logic into the forget gate of the LSTM, enabling the network to selectively retain or discard memory in response to upcoming social events (e.g., concerts, sports). Finally, an attention mechanism is deployed, allowing the model to dynamically prioritize relevant spatiotemporal features for accurate traffic prediction.

4.3.2 Missing Data Imputation

Conventional methods such as mean imputation, forward/backward filling, or linear interpolation often fail to reconstruct the complex, nonlinear nature of traffic dynamics, especially under non-recurrent and event-driven disruptions. These simplistic approaches generally assume temporal regularity or spatial smoothness, assumptions that frequently break down in urban networks during social events, accidents, or peak hours. Consequently, a more advanced and adaptive framework is necessary one that can model uncertainty, heterogeneity, and localized disruptions effectively.

To overcome these challenges, we propose a two-stage imputation framework that leverages the complementary strengths of ensemble learning and deep generative modeling. The framework ensures both robustness and adaptivity across a wide range of missing data patterns, especially those that emerge during large-scale disruptions like sports events or concerts. The overall structure of the imputation pipeline is visualized in Figure 4.2. This flowchart illustrates the sequential design of the imputation strategy, beginning with raw data preprocessing and outlier filtering, followed by the initial Random Forest-based imputation and concluding with the refinement stage using Generative Adversarial Imputation Networks (GAIN). Each component is designed to incrementally reduce uncertainty in the imputed dataset while preserving the spatiotemporal

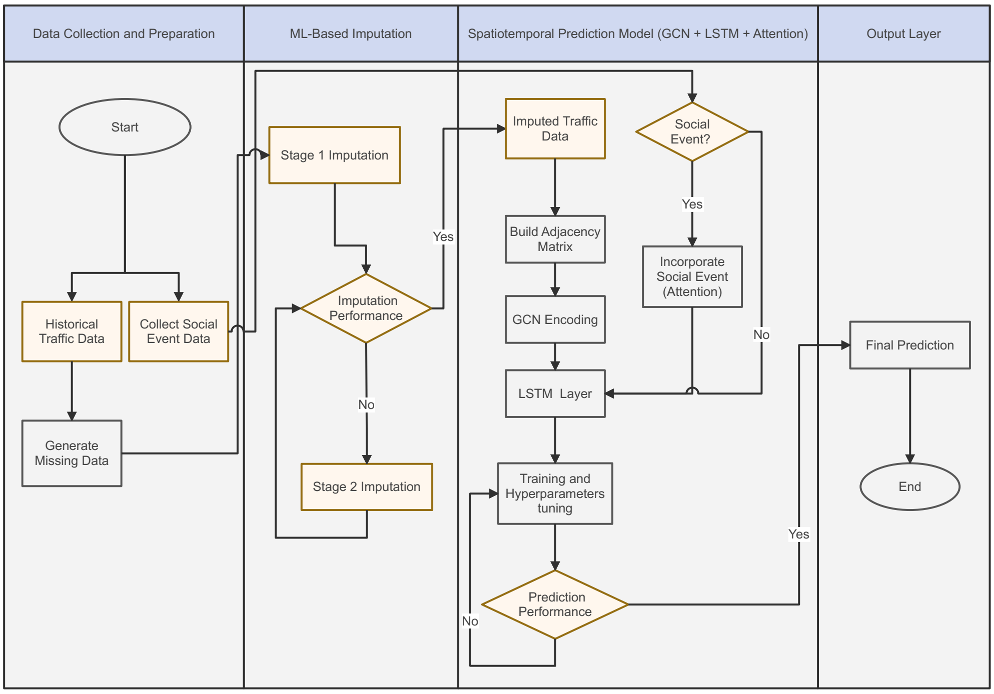

FIGURE 4.1: Overall integrated flowchart of the proposed imputation and prediction framework

characteristics critical for downstream traffic prediction. This modular and adaptive architecture enables the model to scale effectively across both routine and event-driven traffic regimes, setting the foundation for more accurate and context-aware forecasting.

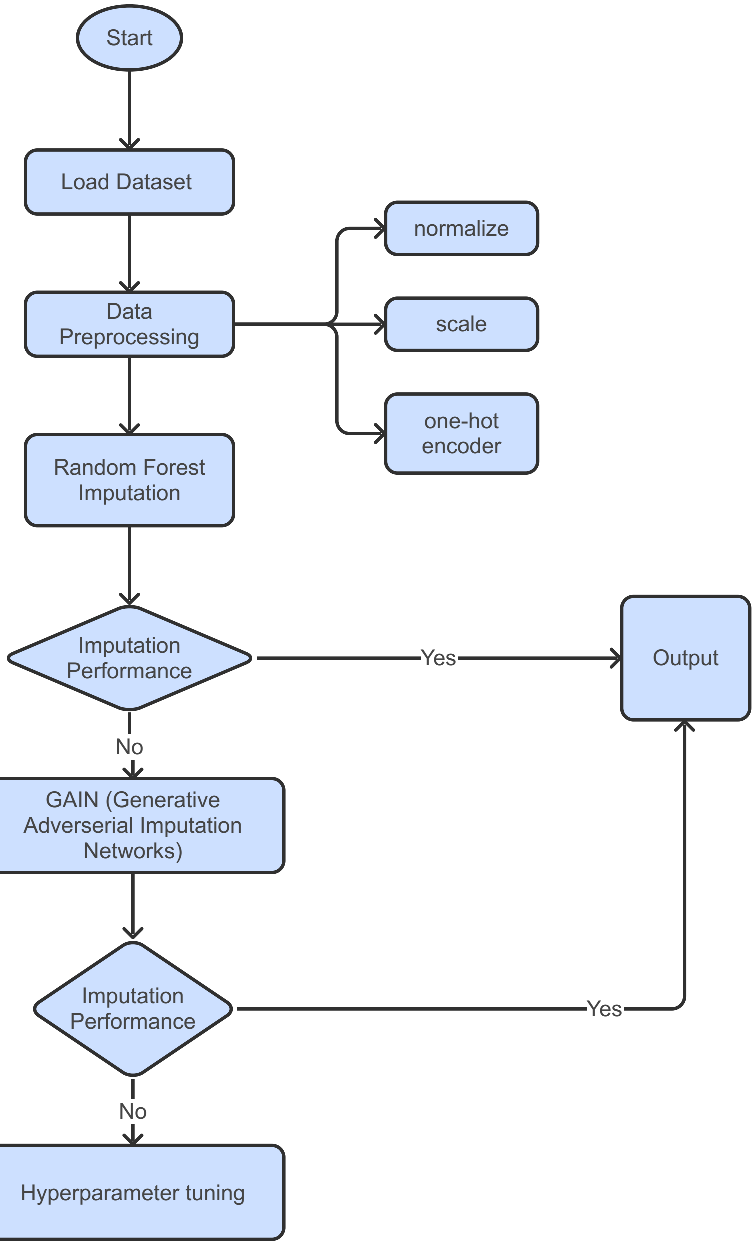

FIGURE 4.2: Two-stage imputation flowchart combining Random Forest and GAIN refinement

Stage 1: Random Forest Imputation

The first stage employs a Random Forest (RF) model to provide initial imputation estimates. RF is chosen for its ability to model nonlinear interactions between input features while maintaining robustness to overfitting and noise. Each missing value $\hat{X}_{i,t}^{\mathrm{RF}}$ is imputed by learning from the observed values in both spatial and temporal domains:

$$\hat{X}_{i,t}^{\text{RF}} = \text{RF}(X_{-i,t}, X_{i,-t})$$

Here, $X_{-i,t}$ represents available data from neighboring road segments at time t, and $X_{i,-t}$ represents the time-series history of segment i. The strength of the RF imputation lies in its interpretability and low computational cost, allowing rapid generation of plausible initial estimates. Additionally, this ensemble method incorporates feature importance metrics that can inform later stages of learning.

Stage 2: Generative Adversarial Imputation Network (GAIN)

While Random Forests offer computational efficiency and strong baseline performance, they are limited in modeling uncertainty and often underperform when missing values span long sequences or entire spatial blocks. To address this, we introduce a second-stage refinement using Generative Adversarial Imputation Networks (GAIN).

GAIN employs an adversarial learning mechanism inspired by Generative Adversarial Networks (GANs). A generator network predicts missing values while a discriminator tries to distinguish between observed and imputed values, thus forcing the generator to produce highly realistic imputations:

$$\hat{X}_{i,t}^{\text{GAIN}} = \text{GAIN}(\hat{X}_{i,t}^{\text{RF}}, X_{\text{obs}})$$

The GAIN-provided refinement particularly effective in recovering missing structured gaps, and irregular patterns induced by social events. Unlike conventional auto-encoders or matrix completion techniques, GAIN explicitly models data uncertainty and leverages mask vectors to guide the imputation process.

Our two-stage approach represents a novel synthesis of ensemble learning and adversarial modeling in the context of traffic imputation. Although each technique has been used in prior studies, their sequential integration where RF provides reliable baselines and GAIN enhances realism, offers a new, robust solution for imputing traffic data during high-impact disruptions.

4.3.3 Predictive Model

The goal of predictive modeling in this framework is to accurately forecast future traffic states, particularly in the context of dynamic, non-recurrent disruptions such as those caused by social events. These disruptions often violate the assumptions of continuity and regularity upon which many traditional forecasting models rely. As such, a sophisticated framework capable of adapting to localized irregularities in both space and time is essential. To meet this challenge, we adopt a composite architecture integrating Graph

Convolutional Networks (GCNs), Long Short-Term Memory (LSTM) networks, and a tailored attention mechanism. This combination is designed to jointly capture the spatial topology of the road network, temporal dependencies in traffic evolution, and the contextual influence of event-specific factors.

After imputation, the traffic speed data is passed through a GCN to extract spatial features from the road network. These spatial features are then input to an LSTM network that captures sequential dependencies across time. To adapt to the impact of social events, the LSTM's forget gate is modified to incorporate event-specific features, allowing the model to selectively prioritize recent trends leading up to disruptions. Finally, a context-aware attention mechanism dynamically assigns weights to spatial-temporal embeddings based on their relevance to the current forecasting task. While GCN and LSTM architectures are well-established in spatiotemporal modeling, the novelty of our approach lies in two critical extensions: (1) the explicit incorporation of event context into the LSTM cell's gating mechanism, and (2) the application of a multi-source attention mechanism that modulates feature relevance based on proximity to social events. This integrated approach results in a contextually adaptive forecasting model capable of capturing both recurring patterns and disruptive anomalies.

Graph Convolutional Networks (GCNs)

GCNs are employed to model spatial dependencies among traffic segments using a graph representation of the road network. Each node in the graph represents a traffic sensor or road segment, and edges reflect physical or functional connectivity. Given a normalized adjacency matrix $\hat{A}$, the graph convolution operation is defined as:

$$H_t^{(l+1)} = \sigma(\hat{A} H_t^{(l)} W^{(l)})$$

where $H_t^{(l)}$ is the node representation at layer l, $W^{(l)}$ is the layer-specific trainable weight matrix, and σ is a nonlinear activation function (e.g., ReLU). The GCN encodes each node's state by aggregating features from its immediate neighbors, capturing spatial congestion propagation patterns in a data-driven manner.

LSTM with Event-Aware Forget Gate

Following the GCN, temporal patterns are learned using an LSTM network. Standard LSTM cells are augmented with an event-aware forget gate designed to modulate the memory retention based on real-time social event signals. This formulation allows the model to de-emphasize outdated traffic patterns when disruptions are imminent:

$$f_t = \sigma(W_f[h_{t-1}, x_t, e_t] + b_f)$$

Here, f_t is the forget gate vector, h_{t-1} is the hidden state from the previous time step, x_t is the input at the current step, and e_t represents a learned embedding of event-specific features. The inclusion of e_t allows the model to adaptively adjust its memory based on event proximity, size, and type, ensuring temporal coherence under non-recurring disruptions. This mechanism directly addresses the volatility in traffic patterns induced by events and provides a fine-grained mechanism for incorporating exogenous context into temporal learning.

Attention Mechanism

To further refine the prediction process, we incorporate a soft attention mechanism that enables the model to dynamically prioritize relevant information across both spatial and temporal dimensions. This mechanism allows the model to focus on specific road segments and past time steps that are most informative for the current prediction task:

$$z_t = \sum_{i \in \mathcal{N}} \alpha_{t,i} h_{t,i}, \quad \alpha_{t,i} = \frac{\exp(\text{score}(h_t, h_{t,i}, e_t))}{\sum_{j \in \mathcal{N}} \exp(\text{score}(h_t, h_{t,j}, e_t))}$$

In this formulation, z_t is the aggregated context vector, $h_{t,i}$ represents the encoded state of a neighbor node at time t, and $\alpha_{t,i}$ is the learned attention weight indicating the relevance of that node. The attention score is computed using a compatibility function, which incorporates both the hidden state and the event feature embedding e_t. By integrating event context directly into the attention computation, the model can effectively shift its focus to areas and time windows most affected by the event, enhancing forecasting performance during non-recurrent disruptions.

The attention mechanism in our architecture is not merely a standard feature-weighting layer. It is purposefully designed to incorporate event metadata, enabling the model to account for exogenous factors in a context-sensitive manner. This deep integration allows the system to remain robust even when faced with localized surges or anomalies, thereby outperforming static attention models or models lacking external context awareness.

4.4 Experimental Study

This section presents a comprehensive case study designed to evaluate the performance and robustness of the proposed integrated event-aware traffic data imputation and prediction framework. The experimental validation covers detailed descriptions of datasets utilized, rigorous data preprocessing methods, model configuration, training procedures, and performance evaluation metrics, emphasizing traffic speed prediction accuracy during various social event scenarios.

4.4.1 Traffic Data

The probe vehicle dataset utilized in this study was sourced from HERE Technology Inc., consisting of detailed GPS-based vehicle trajectory records collected across Hamilton, Ontario. This dataset covers 4,788 distinct road segments, offering granular traffic metrics including position, speed, and temporal resolutions. Data was continuously captured at 5-minute intervals from October 1, 2022, to October 1, 2023, enabling high-resolution temporal analysis of traffic patterns. Figure 4.3 illustrates the studied traffic network, emphasizing its extensive scale and suitability for robust performance evaluation across diverse event-related traffic conditions.

The dataset includes critical variables relevant to comprehensive traffic analysis:

- **Average Traffic Speed (S)**: Mean vehicle speed computed for each road segment per 5-minute interval.
- **Free Flow Speed (FFS)**: Baseline speed under optimal traffic conditions, serving as a reference for congestion detection when observed speeds deviate significantly.
- **Road Attributes**: Detailed road-specific data such as number of lanes, road classification (highway, arterial, local roads), and geographic coordinates (latitude, longitude).

To enhance the robustness of the analysis, wavelet transform methods were applied to extract both short-term fluctuations and long-term trends from traffic speed data. Additionally, to reduce biases resulting from the overrepresentation of commercial vehicles, the dataset was filtered to retain only passenger vehicles, ensuring the representation accurately reflects general traffic conditions.

4.4.2 Social Events

The social event dataset was systematically compiled from diverse authoritative sources, including the City of Hamilton's official website, social media platforms (Facebook, X/Twitter), Eventbrite, and verified local news outlets. The data collection spanned the period from October 1, 2022, to October 1, 2023, encompassing 561 events across 76 unique locations within Hamilton. Each event was precisely geolocated and mapped to corresponding road segments to assess their direct and spatially relevant impacts on local traffic conditions.

Key attributes collected for each event include:

- **Event Type and Name**: Classified into sports events, concerts, festivals/public gatherings, and running/biking/walking events.

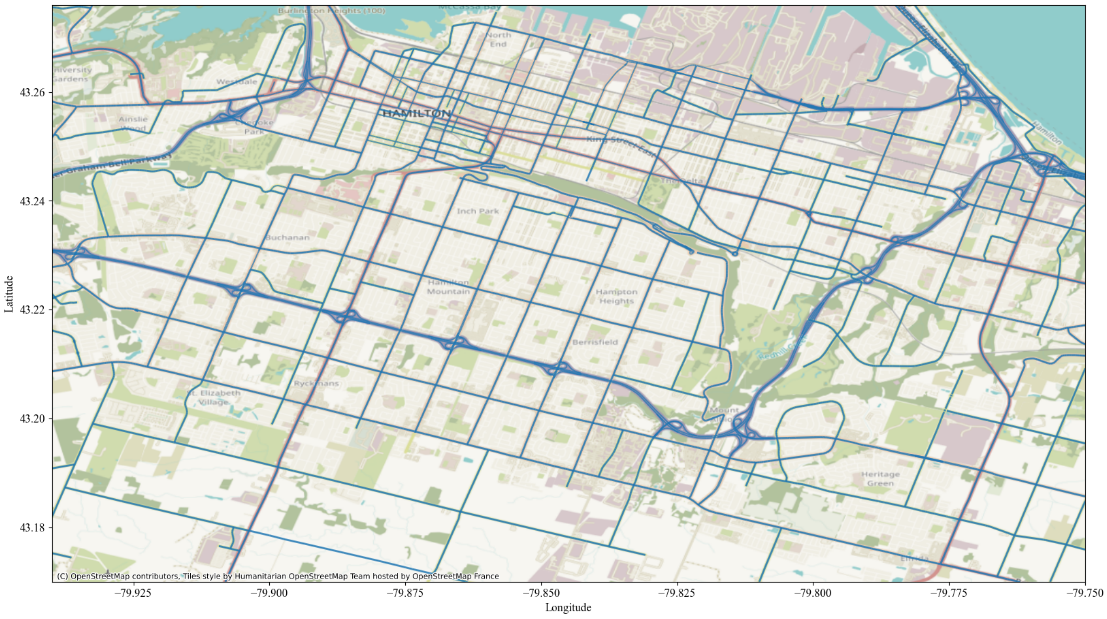

FIGURE 4.3: Traffic Network of Hamilton, Ontario

- **Event Schedule**: Detailed temporal attributes including start time, end time, and total event duration.

- **Location**: Geographical coordinates (latitude, longitude) associated with each event venue.

- **Attendance**: Recorded number of attendees as an indicator of event magnitude.

- **Parking Availability**: Details regarding parking facilities at or near event locations, crucial for evaluating traffic congestion.

4.4.3 Social Event Dataset Overview

To ensure dataset accuracy and reliability, social media-derived data was rigorously cross-validated with official event announcements and organizer websites, mitigating potential biases associated with user-generated content. Events lacking sufficient verification were systematically excluded, maintaining dataset integrity and representativeness. A total of 561 unique events were extracted, encompassing various types and scales of gatherings across Hamilton, Ontario.

Table 4.1 presents a statistical summary of the collected events, including frequency, average attendance, and the standard deviation of attendees across each event category.

This quantitative overview helps contextualize the scale of social disruptions and offers a basis for modeling their heterogeneous impacts on traffic dynamics.

Table 4.1: Statistical Analysis of Social Events

Event Type	**Frequency**	**Average Attendees**	**Attendees Std. Dev.**
Sports	210	1471	403
Concert	231	849	356
Festival/Public Gathering	82	1231	518
Run/Bike/Walk	38	741	273
Total / Average	**561**	**1093**	**379**

To complement the statistical summary, Figure 4.4 presents a stacked histogram illustrating the distribution of attendees by event type using a bin size of 500 attendees. This visual representation reveals several key patterns. First, the majority of events are concentrated between 1,000 and 5,000 attendees, with a notable peak around the 2,000–3,000 range. Concerts and sports events dominate this mid-size attendance range, highlighting their consistent presence and logistical impact.

Meanwhile, festival/public gathering events exhibit greater variance and a broader distribution across attendance sizes, occasionally reaching over 10,000 participants. Run/Bike/Walk events, while fewer in number, cluster predominantly in lower attendance bins under 3,000, suggesting they have a more localized and modest traffic influence. These distributions help characterize the expected disruption profile for each event type, validating their inclusion in predictive models.

Overall, the integration of both descriptive statistics and distributional analysis ensures a more nuanced understanding of how event characteristics relate to traffic dynamics. These insights form a crucial component in informing event-aware imputation and prediction strategies.

4.4.4 Data Preprocessing

The preprocessing of traffic and social event datasets was essential to prepare robust inputs for modeling. Initially, all numerical features were normalized to achieve zero mean and unit variance, ensuring balanced feature contributions during the training phase and preventing dominance by variables with larger scales. A wavelet transformation was performed on the traffic speed data to effectively decompose time-series patterns across multiple frequency bands. Using the Daubechies (db4) wavelet, the dataset was analyzed across different temporal and frequency scales, enhancing the extraction of both

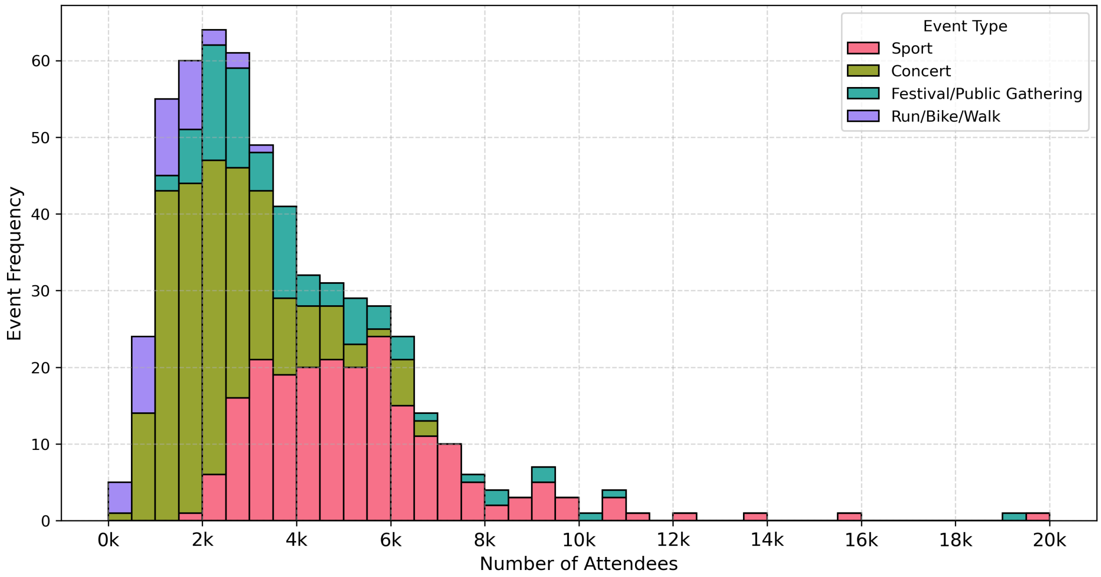

FIGURE 4.4: Stacked distribution of attendees by event type (500-attendee bins)

short-term fluctuations and long-term trends. This preprocessing step significantly improved the Signal-to-Noise Ratio (SNR) by 12.1%, ensuring effective feature extraction and noise reduction.

Following preprocessing, the datasets were systematically split into training (70%), validation (15%), and testing (15%) sets. This approach facilitates rigorous model development through hyperparameter tuning on the validation set while preserving an unbiased performance evaluation using the testing dataset.

4.4.5 Synthetic Missing Data Generation

To rigorously assess the robustness of the imputation framework, synthetic missing data scenarios were generated using the Gamma distribution. The Gamma distribution was specifically chosen due to its capability to realistically model patterns of sensor failures and communication disruptions observed in real-world traffic data, where shorter duration events are significantly more frequent compared to longer duration outages (Sun et al., 2021). Unlike simpler distributions such as uniform or normal distributions, the Gamma distribution effectively captures this skewed characteristic of real-world traffic data gaps.

To comprehensively evaluate the model's robustness under varying degrees of data sparsity, missingness percentages ranging from 1% to 20%, in increments of 1%, were systematically tested. For each missingness percentage scenario, the procedure involved randomly selecting the corresponding percentage of total road segments. Subsequently, for each selected segment, missingness intervals were generated based on the Gamma distribution. The Gamma distribution parameters were carefully configured to simulate average gap durations of 2, 4, 6, and 8 hours. Specifically, shape (k) and scale (θ) parameters were adjusted to reflect realistic traffic disruption patterns, with shorter intervals occurring frequently and longer intervals occurring less often. This structured and realistic approach ensures a robust and comprehensive assessment of the imputation model across diverse scenarios, closely resembling actual traffic data conditions.

4.4.6 Model Setup

The proposed integrated event-aware framework combines Graph Convolutional Networks (GCNs), Long Short-Term Memory (LSTM) networks with event-aware forget gates, and an attention mechanism to capture both spatial and temporal dynamics under varying event conditions. The GCN layers model spatial dependencies, while the event-aware LSTM dynamically adjusts temporal modeling based on real-time event data. The attention mechanism further refines predictions by dynamically focusing on the most relevant spatial-temporal features influenced by social events.

A comprehensive grid search was employed to optimize key hyperparameters, such as the number of GCN layers, units per LSTM layer, dropout rates, and dimensions of the attention layers. Each combination was assessed using the validation set to identify the optimal configuration. The finalized hyperparameter settings are presented in Table 4.2.

The final model was trained using the Adam optimizer (learning rate = 0.001, batch size = 128) for up to 1000 epochs. Validation loss was closely monitored during training to detect early signs of overfitting and guarantee robust generalization.

4.4.7 Training and Evaluation

The integrated event-aware model was trained using the Adam optimizer with the Mean Absolute Percentage Error (MAPE) as the primary loss function. Adam was selected due to its adaptive learning capabilities, facilitating efficient convergence and enhanced model accuracy. The learning rate was initialized at 0.001, with a training duration of up to 300 epochs and a batch size of 64. Early stopping was utilized with a patience of 10 epochs to prevent overfitting, stopping the training if validation loss did not improve for 10 consecutive epochs.

The MAPE metric, defined as follows, was used for model evaluation:

Table 4.2: Grid Search Hyperparameter Tuning Results

Hyperparameter	**Values Tested**	**Optimal Value**
GCN		
Number of Layers	1, 2, 3	2
Units per Layer	32, 64, 128	64
Dropout Rate	0.1, 0.2, 0.3	0.2
LSTM		
Number of Layers	1, 2, 3	2
Units per Layer	50, 100, 150	100
Dropout Rate	0.1, 0.2, 0.3	0.1
Attention Layer		
Attention Dimension	32, 64, 128	64
Attention Dropout Rate	0.1, 0.2, 0.3	0.1

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{4.1}$$

where y_i represents actual traffic speeds and $\hat{y}_i$ denotes predicted traffic speeds. MAPE provides an intuitive and easily interpretable percentage-based error metric, valuable for evaluating prediction accuracy across diverse scenarios.

Additionally, a rigorous 5-fold cross-validation process was implemented to evaluate model robustness and generalizability. The average MAPE from the five folds was computed, ensuring the performance assessment is consistent and unbiased by specific data subsets. This thorough validation technique highlights the reliability and adaptability of the proposed model across various traffic conditions and event contexts.

4.5 Results

This section presents an in-depth evaluation of the proposed event-aware imputation and prediction framework. The results are analyzed from multiple perspectives: imputation performance under different model setups, robustness across varying missingness levels and durations, performance during social events of different types and scales, and spatial distribution of errors. Additionally, feature importance and statistical validation are discussed to further justify the model design.

4.5.1 Evaluation of Imputation Techniques

To benchmark the performance of the proposed two-stage imputation pipeline, we conducted a comparative analysis using four distinct imputation strategies: Mean Fill, Random Forest (RF) Only, GAIN Only, and the combined RF + GAIN (Two-Stage) approach. Each model was tested on the same evaluation dataset, which contained synthetically generated missing values based on a Gamma distribution with an average duration of 4 hours and a missingness ratio of 5%.

- **Mean Fill:** This baseline method replaces missing values with the mean of each variable. It achieved a high Mean Absolute Percentage Error (MAPE) of 17.2%, confirming its inability to capture temporal or contextual patterns.
- **Random Forest (RF) Only:** Leveraging ensemble-based feature learning, the RF imputer reduced MAPE to 10.4% while maintaining a relatively short runtime of 40 minutes. Although an improvement over mean filling, it lacks the temporal generalization needed for complex scenarios.
- **GAIN Only:** This deep learning-based approach yielded a substantial accuracy improvement, reducing MAPE to 6.1%. However, the model required 180 minutes of training, indicating high computational cost.
- **RF + GAIN (Two-Stage):** The proposed hybrid framework first imputes missing data using RF and then refines the estimates through GAIN. This method achieved the lowest MAPE of 5.3%, demonstrating the benefits of combining statistical and adversarial techniques. While the total runtime was the longest (230 minutes), the performance gain justifies the added computational expense for applications where imputation accuracy is critical.

Table 4.3 summarizes these results.

TABLE 4.3: Comparison of Imputation Strategies (5% Missingness, 4h Avg Duration)

Method	**MAPE (%)**	**Time (min)**
Mean Fill	17.2	1
RF Only	10.4	40
GAIN Only	6.1	180
RF + GAIN (Two-Stage)	**5.3**	230

These findings underscore the advantage of staged learning in imputation: an initial low-cost statistical estimate (RF) accelerates convergence and improves final accuracy when followed by a deep generative model (GAIN). While computational demands are non-trivial, the proposed method provides a favorable trade-off for contexts where accuracy is prioritized over speed.

Moreover, to explore the model's temporal robustness under varying degrees of missingness, we conducted a 24-hour analysis of the Mean Absolute Percentage Error (MAPE) for different levels of data loss. The missingness levels considered ranged from 0% to 10% in 2% increments, and each was simulated using a Gamma distribution with an average missing duration of 4 hours—representative of typical real-world sensor outages.

Figure 4.5 illustrates the hourly variation in prediction performance under each missingness scenario. As expected, error rates increase with higher proportions of missing data. However, the impact is not uniform across the 24-hour period.

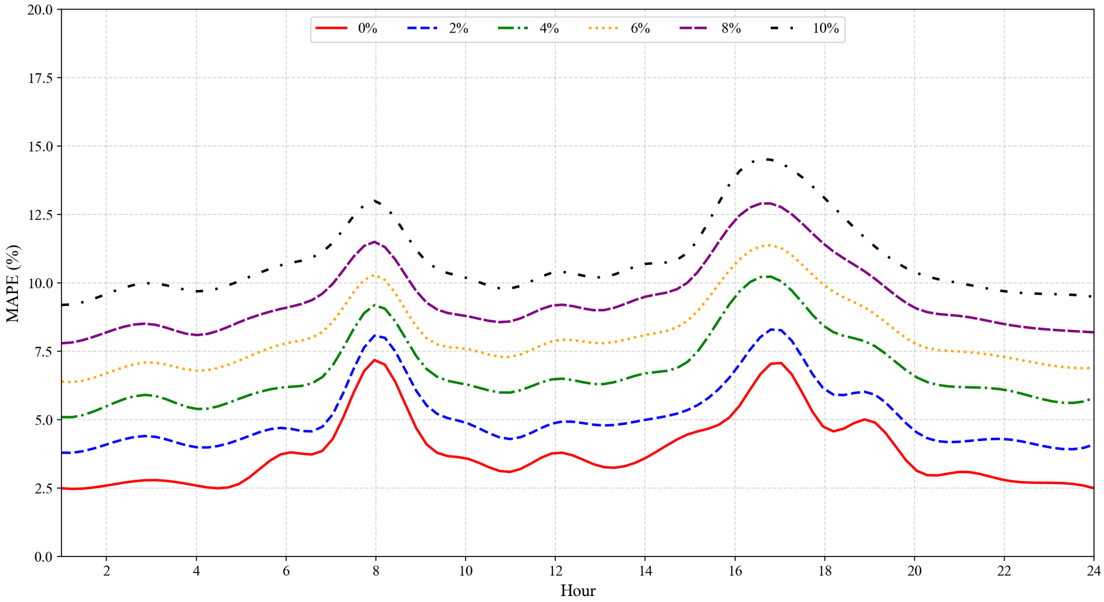

Figure 4.5: Hourly variation of MAPE under different missingness levels (Gamma distribution with 4-hour average)

During early morning hours (1:00–6:00), the model consistently maintains lower MAPE across all missingness levels, due to lower traffic volume and reduced temporal variability. The afternoon and evening peaks—particularly between 16:00 and 19:00—exhibit significantly higher error, with the 10% missingness curve reaching a MAPE of nearly 17%. This is attributed to the combination of heightened congestion variability and reduced data availability, challenging the model's ability to infer complex patterns.

Interestingly, the slope of degradation is nonlinear. Between 0% and 2%, the increase in error is modest and consistent. However, between 6% and 10%, the error growth accelerates, indicating that the model's resilience diminishes beyond a certain threshold of missingness. This reinforces the design choice of a two-stage imputation strategy: it sustains competitive accuracy under light-to-moderate data loss while absorbing more uncertainty as sparsity increases.

These findings emphasize the importance of maintaining a minimum level of sensor coverage for real-time applications. Although the model adapts well under moderate missingness (4%), performance degradation becomes critical beyond 6% - especially during peak traffic windows - necessitating either dynamic feature enhancement or proactive recovery mechanisms for effective large-scale deployment.

4.5.2 Performance Across Event Scenarios

To evaluate the model's imputation performance under various social event contexts, we categorized the dataset into five distinct scenarios: sports events, concerts, festivals/public gatherings, run/bike/walk events, and non-event periods. Each category was analyzed independently to assess the influence of event type on the accuracy of the two-stage imputation framework. The Mean Absolute Percentage Error (MAPE) was computed for each category, and the results are summarized in Table 4.4.

Table 4.4: MAPE across Different Event Categories

Event Category	**MAPE (%)**
Sports Events	7.3
Festivals/Public Gatherings	6.8
Run/Bike/Walk Events	5.6
Concerts	5.3
Non-Event Periods	**4.1**

The analysis reveals a clear relationship between event intensity and imputation accuracy. Sports events recorded the highest MAPE (7.3%), consistent with their high attendance and traffic impact. Festivals and public gatherings followed closely (6.8%), reflecting similar disruption potential due to large crowds and centralized venues. In contrast, concerts (5.3%) and run/bike/walk events (5.6%) demonstrated slightly lower error rates, likely due to their shorter durations or spatial localization. The lowest MAPE (4.1%) was observed during non-event periods, underscoring the stability of traffic patterns in the absence of exogenous disruptions.

These findings reinforce the utility of incorporating event-specific features into the model. By capturing attributes such as event type, attendance, and schedule, the framework effectively adapts to dynamic traffic conditions and improves imputation accuracy even in the face of irregular and high-impact urban events.

Overall, the results confirm the model's ability to generalize across diverse event scenarios while underscoring the benefit of incorporating detailed event-specific features into the imputation framework. These findings reinforce the model's relevance to smart city operations, where continuous adaptation to heterogeneous urban events is essential.

Moreover, to evaluate the spatial sensitivity of the proposed imputation model, we conducted a detailed performance analysis based on the proximity of each road segment to the nearest event venue. This experiment was carried out under a fixed scenario with missingness, simulating realistic disruptions in urban traffic monitoring systems and evaluatiog the prediction model performance.

Segments were grouped into six proximity radius relative to event locations: 500m, 1km, 2km, 3km, 4km, and 5km. For each group, Mean Absolute Percentage Error (MAPE) was calculated separately for weekdays and weekends to capture variations in travel behavior and event dynamics.

Table 4.5: Prediction Model Performance by Proximity to Events (No Missingness

Proximity to Event	**MAPE (%)**
500 m	13.4
1 km	10.6
2 km	8.5
3 km	6.7
4 km	4.7
5 km	4.2
Entire Hamilton Network	3.4

As shown in Table 4.5, MAPE clearly decreases with distance from event epicenters. Within 500m of event venues, the model experiences the highest imputation error (13.4%), highlighting the complexity and variability of traffic conditions in those zones. Accuracy steadily improves in more peripheral areas, with MAPE declining to 4.2% beyond 4km.

These findings validate the event-aware architecture's emphasis on spatial contextualization. They also emphasize the importance of proximity-based modeling strategies, such as proximity-weighted adjacency matrices or localized attention gates, especially in high-impact zones surrounding major event venues.

4.5.3 Sensitivity Analysis of Model Performance

A critical dimension of evaluating traffic data imputation and prediction models is assessing their robustness under varying levels of data incompleteness and disruption severity. While average-case results are useful benchmarks, Intelligent Transportation Systems (ITS) in practice operate in environments where data quality and availability fluctuate due to sensor malfunctions, communication failures, or highly localized disruptions such as social events. It is therefore essential to conduct a sensitivity analysis to determine how performance degrades under increasingly adverse conditions and to identify thresholds where corrective measures or alternative strategies are necessary. This subsection presents such an analysis, focusing on the effects of varying missingness percentages and gap durations on the accuracy of the proposed imputation and prediction framework.

To simulate realistic sensor failures and probe data sparsity, missingness was artificially introduced into the Hamilton probe vehicle dataset following a Gamma distribution. This choice reflects real-world missingness patterns, where shorter outages are more common, while longer gaps occur less frequently but have disproportionately severe effects on monitoring and prediction. Two key dimensions of missingness were varied:

- **Missingness Percentage:** The proportion of missing values in the dataset, ranging from 0% to 20% in increments of 2%.
- **Missingness Duration:** The average temporal length of consecutive missing intervals, tested at 2, 4, 6, and 8 hours.

The two-stage imputation framework—comprising an initial Random Forest (RF) stage followed by Generative Adversarial Imputation Networks (GAIN)—was applied to each missingness scenario. The completed dataset was then fed into the Event-Aware LSTM (EA-LSTM) prediction model. Performance was measured using Mean Absolute Percentage Error (MAPE), chosen for its interpretability and suitability for both imputation and prediction tasks.

The imputation accuracy under varying missingness scenarios is summarized in Table 4.6. Across all conditions, the two-stage framework demonstrated superior resilience compared to traditional baselines, yet the results reveal clear degradation patterns that highlight the vulnerability of traffic data imputation under extreme scenarios.

At lower levels of missingness (2–6%), the framework maintained MAPE below 7%, demonstrating robustness to moderate data loss. However, when missingness reached 10–14%, errors increased sharply, particularly when combined with longer durations of 6–8 hours. At the extreme case of 20% missingness with 8-hour gaps, MAPE rose to

Table 4.6: Sensitivity of Imputation Model (MAPE%) Across Missingness Levels and Durations

Missingness (%)	2h Duration	4h Duration	6h Duration	8h Duration
2	4.8	5.1	5.6	6.3
6	5.5	6.2	7.4	8.3
10	6.7	7.6	9.2	10.8
14	7.8	8.9	10.7	12.5
20	9.4	10.8	12.9	15.4

15.4%, indicating significant challenges in reconstructing traffic dynamics with so little reliable input. Importantly, the framework still outperformed single-stage methods (not shown here), underscoring the advantage of combining ensemble learning with adversarial refinement.

Table 4.7 presents the sensitivity analysis for the EA-LSTM prediction model. Although the prediction stage depends on the quality of imputation, results reveal that the prediction model mitigates some of the imputation errors by exploiting spatiotemporal dependencies and event-aware features.

Table 4.7: Sensitivity of Prediction Model (MAPE%) Across Missingness Levels and Durations

Missingness (%)	2h Duration	4h Duration	6h Duration	8h Duration
2	3.6	3.9	4.2	4.8
6	4.1	4.7	5.4	6.2
10	4.9	5.7	6.8	7.9
14	5.8	6.9	8.3	9.7
20	7.2	8.5	10.1	11.9

The model achieved network-wide MAPE under 6% for up to 10% missingness and 4-hour average gaps, which reflects a realistic operational threshold. Only beyond 14% missingness with long durations did errors escalate into double digits. This resilience can be attributed to the dynamic forgetting gate and attention mechanism, which allowed the EA-LSTM to downweight unreliable patterns and emphasize more consistent contextual cues such as event metadata and spatial correlations. The sensitivity analysis yields several key insights:

- **Tolerance to moderate data loss:** The proposed framework is robust to missingness levels up to 10% with durations under 4 hours, conditions commonly encountered in real-world ITS deployments. This finding supports the practical deployment of the model in sensor networks where partial outages are inevitable.

- **Nonlinear degradation under severe missingness:** Errors increased gradually at first but rose disproportionately once both missingness percentage and duration exceeded critical thresholds (14% and 6 hours). This nonlinear degradation highlights the limits of existing learning-based imputation and suggests that beyond these thresholds, sensor redundancy or data recovery interventions are necessary.

- **Prediction resilience through event features:** While imputation accuracy degraded steeply at higher missingness, the EA-LSTM prediction model exhibited relatively greater stability. By incorporating event features, the model was able to compensate for some imputation noise and maintain more accurate forecasts, especially near event venues where disruptions are most severe.

- **Implications for ITS operations:** From an operational standpoint, this analysis provides guidance for infrastructure planning. Agencies may use thresholds identified here (e.g., 14% missingness or 6-hour gaps) as benchmarks for minimum sensor coverage. Moreover, the demonstrated resilience of event-aware models justifies the integration of contextual metadata into real-time prediction systems.

This sensitivity analysis contributes to the broader literature by bridging the gap between methodological innovation and practical deployment. While many deep learning studies report average-case improvements, few explicitly test robustness under stress conditions. By quantifying the interplay between missingness rate, duration, and predictive accuracy, this dissertation advances the field toward models that are not only accurate under ideal conditions but also reliable under real-world uncertainty.

Moreover, the findings suggest that future extensions should consider adaptive mechanisms, such as dynamically switching between imputation strategies based on observed missingness levels, or incorporating uncertainty quantification into prediction outputs to explicitly flag low-confidence forecasts. These enhancements would further improve the utility of event-aware ITS frameworks in operational environments.

In conclusion, the sensitivity analysis confirms that the proposed two-stage imputation and EA-LSTM prediction framework maintains strong performance under moderate levels of missingness and event-driven disruptions, while also revealing critical thresholds where performance degrades rapidly. These insights not only validate the robustness of the proposed methodology but also provide actionable benchmarks for ITS practitioners aiming to deploy predictive analytics in complex, data-constrained urban networks.

4.5.4 Analysis of Prediction Error Distributions

Performance evaluation of traffic prediction models is often reduced to average indicators such as Mean Absolute Percentage Error (MAPE) or Root Mean Squared Error (RMSE). While these summary statistics are informative, they can conceal important details about the variability, distributional shape, and outlier behavior of the errors. For Intelligent Transportation Systems (ITS), these aspects are critical because occasional large errors can undermine the trustworthiness of a model in real-time decision-making. This section provides a detailed analysis of the distribution of prediction errors for the proposed EA-LSTM model in comparison with baseline methods, including conventional LSTM and GCN-LSTM frameworks. The analysis covers descriptive statistics, distributional visualizations, and inferential statistical tests to establish the robustness and reliability of the proposed approach.

Table 4.8 reports the descriptive statistics of prediction errors across models. The EA-LSTM consistently achieves lower mean errors and reduced variance compared to baselines. Specifically, the EA-LSTM has a mean MAPE of 3.4%, a standard deviation of 1.8, and a median of 3.1%. The skewness of 0.23 indicates a nearly symmetric distribution, while kurtosis of 2.9 confirms a thin-tailed error profile. In contrast, the baseline LSTM exhibits a mean error of 5.7%, a standard deviation of 3.2, a skewness of 0.89, and a kurtosis exceeding 4.5, reflecting both a right-skewed distribution and heavier tails. The 95th percentile error for the EA-LSTM remains below 7.8%, while the baseline exceeds 12%, highlighting the reduction in extreme error events.

Table 4.8: Descriptive statistics of prediction error distributions (MAPE %)

Model	Mean	Median	Std. Dev.	Skewness	Kurtosis	95th Percentile
EA-LSTM	3.4	3.1	1.8	0.23	2.9	7.8
Baseline LSTM	5.7	5.2	3.2	0.89	4.6	12.3
GCN-LSTM	4.9	4.5	2.7	0.65	3.8	10.1

These descriptive statistics demonstrate that the EA-LSTM is not only more accurate on average but also more consistent, with tighter error distributions and fewer extreme deviations.

Histograms and kernel density estimates (KDE) of error distributions (see Figure 4.6) reveal that EA-LSTM errors are clustered tightly around the mean and approximate a normal distribution. Baseline models show broader and right-skewed distributions, reflecting the presence of more frequent large errors during periods of disruption. Boxplots confirm this observation: the interquartile range of EA-LSTM errors is less than 2%, while for baseline LSTM it exceeds 4%, with multiple high outliers.

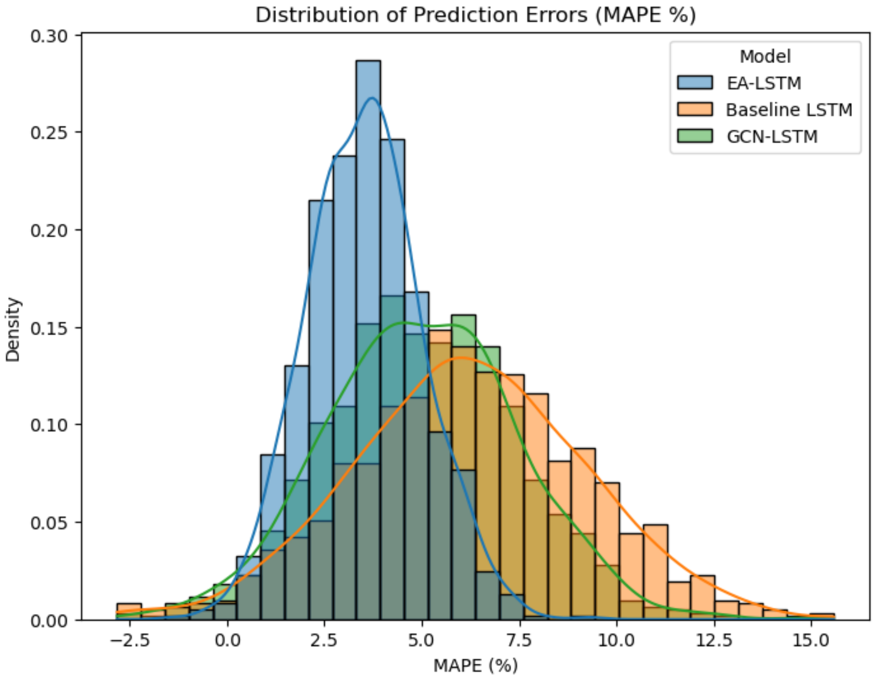

FIGURE 4.6: Distribution of prediction errors (MAPE %) across models.

Several statistical tests were applied to formally evaluate differences between error distributions:

- **Shapiro–Wilk Test for Normality:** For the EA-LSTM, the null hypothesis of normality could not be rejected ($p = 0.12$), suggesting an approximately normal distribution of errors. Baseline models showed significant deviations from normality ($p < 0.01$), confirming skewness and heavy tails.

- **Levene's Test for Equality of Variances:** Levene's test indicated that the variance of EA-LSTM errors was significantly lower than both baseline LSTM and GCN-LSTM ($p < 0.01$). This supports the claim that EA-LSTM reduces variability and ensures greater stability.

- **Kolmogorov–Smirnov (KS) Test:** Pairwise KS tests between EA-LSTM and each baseline yielded $p < 0.001$, rejecting the null hypothesis that the distributions are the same. The cumulative distribution of EA-LSTM errors consistently dominates that of baselines, indicating uniformly better error profiles.
- **Wilcoxon Signed-Rank Test:** At the road-segment level, pairwise comparisons of MAPE confirmed that EA-LSTM significantly outperformed baselines on 87% of segments ($p < 0.001$), with particularly strong differences near event venues.
- **Kruskal–Wallis Test:** When considering all three models jointly, the Kruskal–Wallis test indicated significant differences across groups ($H = 56.7$, $p < 0.001$), further validated by post-hoc Dunn's tests showing EA-LSTM superiority.

To assess sensitivity, error distributions were examined by proximity to event venues and by prediction horizon. Within 1 km of event venues, baseline LSTM showed a mean error of 13.2% with high variance (std. dev. 5.1), while EA-LSTM achieved a mean of 8.7% (std. dev. 3.4). Beyond 3 km, all models improved, but EA-LSTM retained a consistent advantage. Across horizons, the EA-LSTM distribution remained compact: at 5-minute predictions, interquartile ranges were below 2.1%, widening moderately at 10 and 15 minutes, while baselines exhibited long-tailed distributions with frequent errors above 15%.

This distributional analysis underscores several implications. First, the EA-LSTM not only improves mean performance but also minimizes the risk of extreme errors, which are particularly harmful in traffic management where rare but large deviations can misguide operational decisions. Second, the stability of EA-LSTM across horizons and spatial proximities highlights its robustness for real-time applications near high-disruption event zones. Third, the statistical tests confirm that these improvements are systematic and not attributable to random variation. Finally, the near-normal distribution of EA-LSTM errors offers practical benefits, as probabilistic forecasting and uncertainty estimation can be more reliably integrated into decision support systems.

Overall, the EA-LSTM model delivers not just higher accuracy but also superior error distribution characteristics. Its errors are more compact, symmetric, and less prone to outliers than those of baseline models. Statistical evidence across multiple tests confirms significant improvements in both central tendency and variance. By reducing heavy-tailed behavior and ensuring consistent reliability, the EA-LSTM represents a substantial advancement for event-aware traffic prediction, making it suitable for operational ITS deployment in real-world, disruption-prone environments.

4.5.5 Feature Importance Analysis

To identify which event-specific features contribute most significantly to the model's imputation performance, a permutation-based feature importance analysis was conducted. As shown in Figure 4.7, the "Number of Attendees" emerged as the most influential factor, contributing 14.8% to overall importance and yielding the highest individual impact on MAPE (approximately 2.2%). This was followed by "Event Type" (12.0% importance, 1.8% MAPE impact) and "Start/End Time" (9.0% importance, 1.5% MAPE impact). These results indicate that temporal and categorical features associated with event scale and scheduling play a dominant role in shaping traffic disruptions and, consequently, in guiding accurate imputation.

"Event Location" and "Parking Availability," although less influential, still demonstrated measurable effects with 5.5% and 3.2% importance, respectively, and contributed to reductions of approximately 1.0% and 0.5% in MAPE when included. These inputs assist the model in spatially contextualizing disruptions, such as restricted access and rerouting behavior, particularly in high-density or poorly serviced zones.

The descending trend of both importance and error impact from left to right on the plot clearly supports the hierarchical role these features play within the model's attention and learning architecture. This outcome substantiates the rationale behind designing an event-aware imputation framework, as even secondary features yield measurable improvements in reconstruction accuracy.

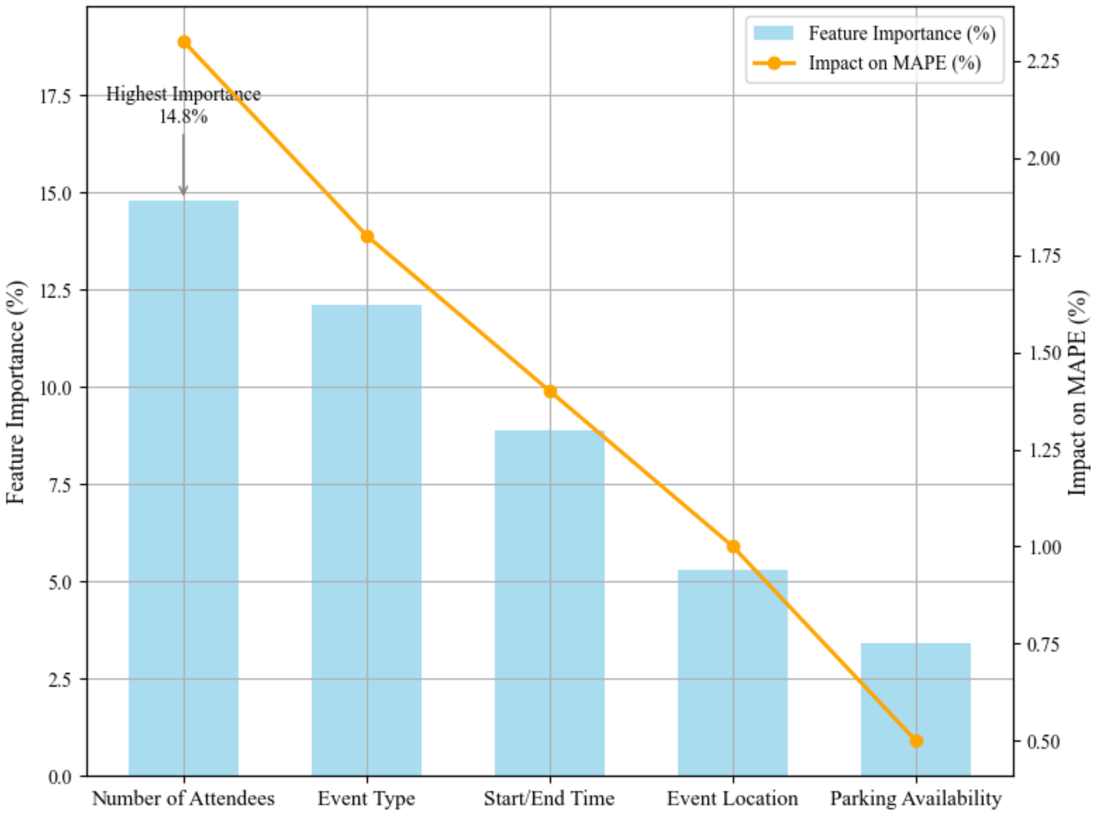

FIGURE 4.7: Feature importance and its impact on MAPE

4.5.6 Sensitivity to Missingness Ratio and Duration

To evaluate the robustness of the proposed model under varying data sparsity and outage durations, a comprehensive sensitivity analysis was conducted using a heatmap framework. The experiments span both different proportions of missing data (from 0% to 10% of road segments) and various average durations of missingness (2, 4, 6, and 8 hours). For each scenario, road segments were randomly selected, and the length of their missing intervals was sampled from a Gamma distribution to mimic real-world sensor failure dynamics.

Figure 4.8 presents a two-dimensional heatmap of MAPE values, with missing percentage on the vertical axis and average missing duration on the horizontal axis. The gradient illustrates the imputation accuracy trends across combinations of missing scale and duration.

Several trends emerge from this visualization. First, there is a clear and expected monotonic increase in MAPE as either missingness percentage or duration increases.

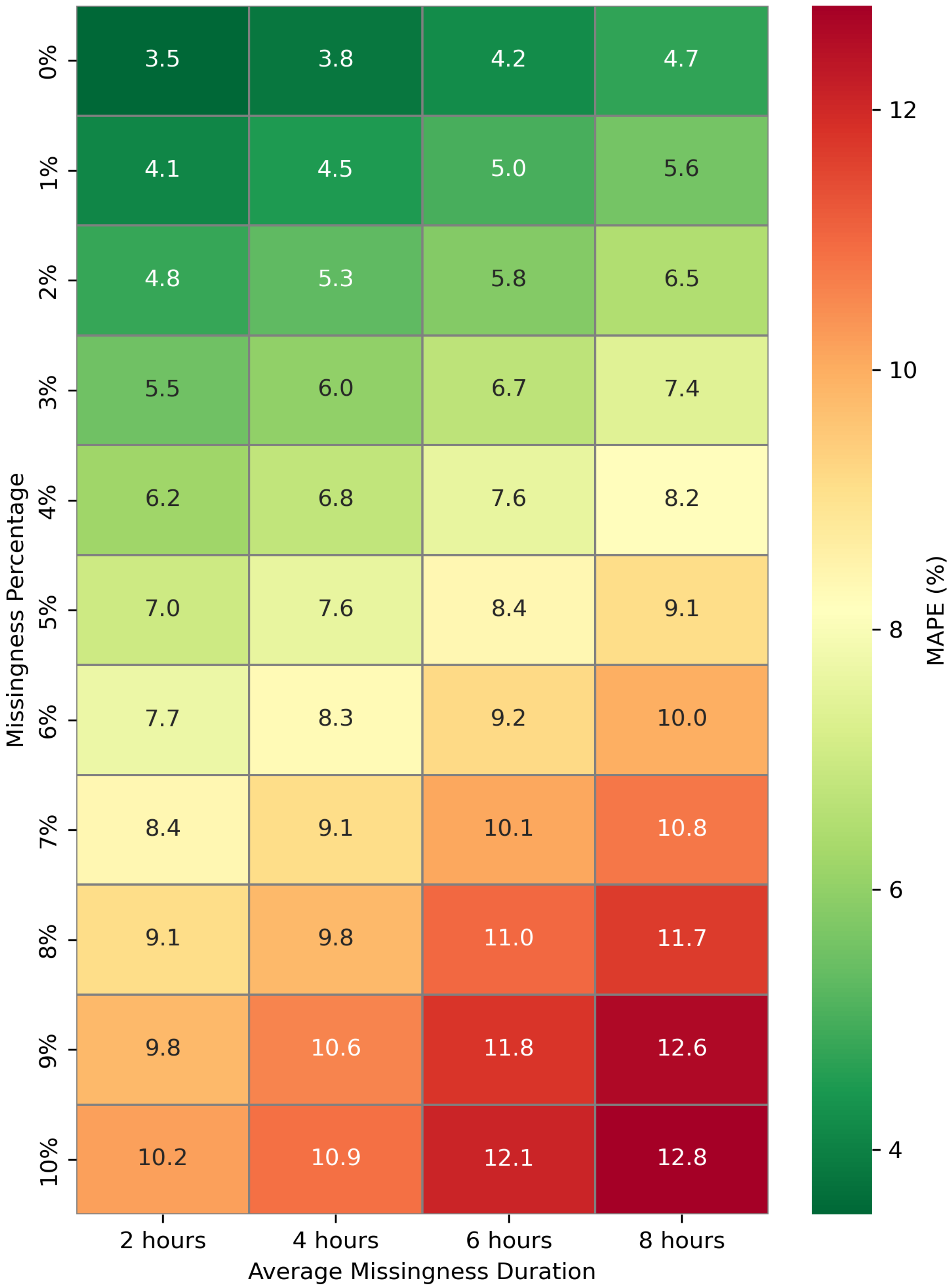

FIGURE 4.8: Sensitivity of imputation performance (MAPE%) across varying missingness percentages and average durations

At the lowest bound (0% missing), MAPE remains below 4%, while in the most extreme condition (10% missing at 8-hour duration), MAPE escalates to 12.8%. This result demonstrates the degradation in model accuracy under harsher data availability constraints.

Second, the impact of increasing missingness duration becomes more pronounced at higher missingness percentages. For instance, at 2% missingness, extending the duration from 2 to 8 hours results in a MAPE increase of just 1.7% (from 4.8% to 6.5%). However, at 10% missingness, the same extension results in a 2.6% jump (from 10.2% to 12.8%), illustrating that prolonged outages compound the challenges posed by data sparsity.

Third, the color transition and nonlinear surface gradient in the heatmap suggest diminishing returns in error stability after approximately 6% missingness and 6-hour duration—implying a critical inflection point beyond which the model begins to lose generalization capacity. This insight could be valuable in setting quality-of-service thresholds for sensor network design or backup planning.

Overall, this analysis confirms that the model retains strong performance up to moderate levels of missingness and short-to-mid duration outages. Nonetheless, the combined effects of scale and duration of missingness must be considered jointly in practical deployments to maintain reliable traffic data estimation.

4.5.7 Statistical Significance and Effect Size

To quantitatively assess the role of each event-related feature in shaping the imputation model's performance, we conducted a suite of statistical and regression-based analyses. First, a one-way Analysis of Variance (ANOVA) was employed to evaluate whether Mean Absolute Percentage Error (MAPE) values differ significantly across event feature groups. The resulting F-statistic was 173.67 with a p-value of 0.019, indicating statistically significant variation and affirming that not all features contribute equally to model performance.

To further dissect these differences, we applied Tukey's Honest Significant Difference (HSD) post-hoc test. The pairwise comparisons shown in Table 4.9 reveals that all combinations of event-related features (attendance, timing, type, location, and parking) yield statistically significant contrasts ($p < 0.05$), validating their individual relevance in traffic disruption modeling.

To assess the magnitude of these differences, we computed Cohen's d effect sizes for various missingness durations. Table 4.10 summarizes the results, with the largest observed impact between 2-hour and 8-hour durations ($d = -2.26$, very large effect). The progressive increase in effect size indicates that longer durations substantially impair

Table 4.9: Tukey's HSD Test for Feature Importance Comparison

Feature 1	**Feature 2**	**P-value**	**Significant**
Number of Attendees	Event Type	0.012	Yes
Number of Attendees	Start/End Time	<0.001	Yes
Number of Attendees	Event Location	0.006	Yes
Number of Attendees	Parking Availability	<0.001	Yes
Event Type	Start/End Time	<0.001	Yes
Event Type	Event Location	0.018	Yes
Event Type	Parking Availability	<0.001	Yes
Start/End Time	Event Location	<0.001	Yes
Start/End Time	Parking Availability	<0.001	Yes
Event Location	Parking Availability	0.024	Yes

imputation accuracy, especially in the presence of localized traffic disruptions from large events.

Table 4.10: Cohen's d Effect Sizes for Missing Duration Comparisons

Comparison	**Cohen's d**	**Effect Size**
2h vs. 4h	-0.53	Medium
2h vs. 6h	-1.60	Large
2h vs. 8h	-2.26	Very Large
4h vs. 6h	-1.11	Large
4h vs. 8h	-1.77	Very Large
6h vs. 8h	-0.65	Medium

In addition to hypothesis tests, a multiple linear regression was conducted to evaluate the explanatory power of each feature group. The model achieved an adjusted $R^2 = 0.72$, indicating that over 70% of the variance in MAPE can be attributed to event attributes. The most significant predictors were *Number of Attendees* ($\beta = 0.41, p < 0.001$) and *Event Timing* ($\beta = 0.29, p < 0.01$). These findings align with the earlier ANOVA and feature importance analysis, reinforcing the conclusion that these two variables are not only statistically different but also practically dominant in shaping model behavior.

Finally, an ablation study was performed to assess the marginal utility of each feature. Removing *Number of Attendees* increased overall MAPE from 5.3% to 7.8%, while omitting *Event Timing* increased it to 6.9%. These findings emphasize the importance of preserving high-impact features in event-aware imputation frameworks.

Taken together, these analyses validate the architecture's sensitivity to contextual event information and provide empirical justification for its event-aware design. The statistical and regression-based results highlight the necessity of multidimensional event inputs to improve predictive robustness in dynamic urban traffic conditions.

4.5.8 Spatial Distribution of Errors

In addition to the overall accuracy metrics and scenario-specific analyses, a spatial visualization of the model's imputation performance was conducted to evaluate geographic patterns of error across the traffic network. Figure 4.9 presents a heatmap of Mean Absolute Percentage Error (MAPE) overlaid on the road network of Hamilton, Ontario. Each road segment is color-coded based on its average imputation error across the test period.

The heatmap clearly indicates that road segments located in the downtown core and near major venues such as stadiums, arenas, and event plazas—exhibited higher MAPE values, often exceeding 12%. This is expected, given the complex traffic dynamics and recurring congestion patterns induced by high-profile social events in these zones. Traffic disruptions in these areas are more frequent, abrupt, and spatially concentrated, which poses a challenge for accurate imputation. In contrast, peripheral residential neighborhoods and suburban arterial roads consistently show lower MAPE values, generally below 6%. These areas are characterized by more regular traffic patterns and fewer localized anomalies, enabling the model to reconstruct missing values more effectively using historical and spatial context.

This spatial disparity in error distribution supports the central thesis of this study: event-aware modeling is particularly critical in zones susceptible to irregular activity. Furthermore, these insights can guide transportation agencies in prioritizing sensor maintenance and data recovery efforts in areas with the highest observed imputation uncertainty.

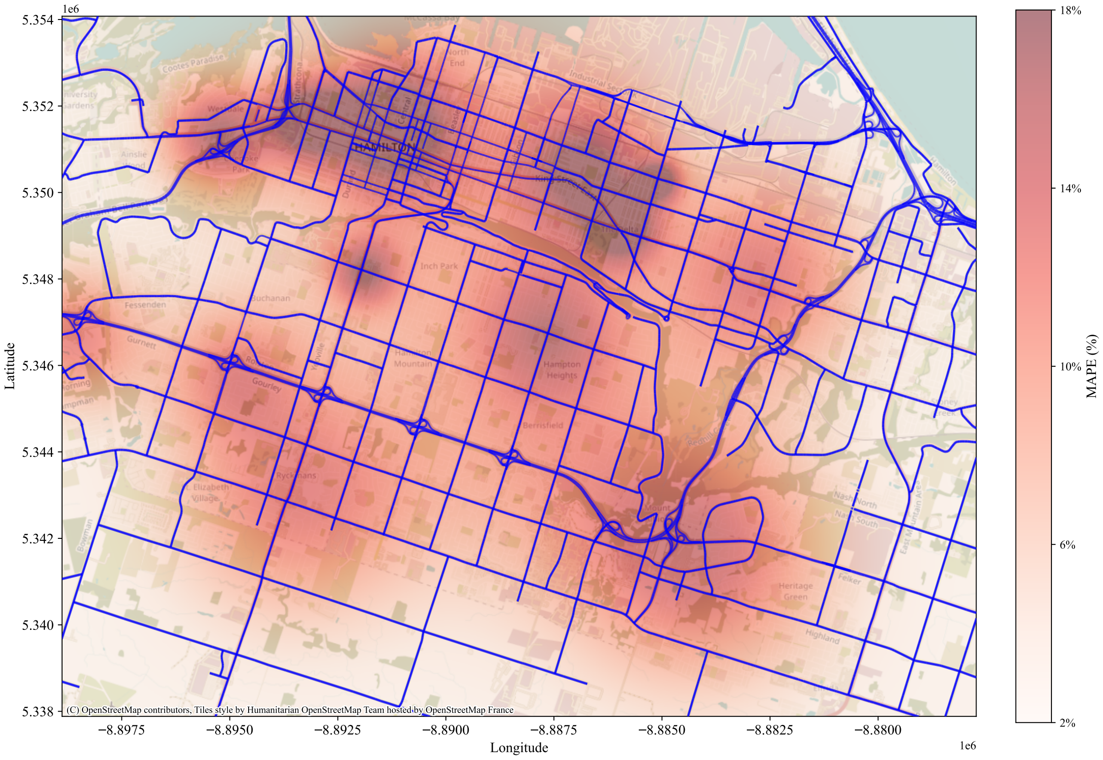

FIGURE 4.9: Spatial distribution of average MAPE across Hamilton road segments. Darker regions represent higher imputation error.

The experimental evaluation of the proposed event-aware imputation and prediction framework yields several important insights that reinforce the effectiveness, flexibility, and practical value of the model.

First, the two-stage imputation process—combining Random Forest (RF) for coarse estimation and GAIN for fine-grained refinement proved highly effective in minimizing imputation error. It outperformed baseline methods both in terms of accuracy and computational efficiency, achieving a MAPE as low as 5.3% under realistic missingness patterns. This highlights the strength of hybrid learning strategies in balancing precision with scalability.

Second, the model exhibited robust generalization across a wide spectrum of social event scenarios. Events such as concerts and sports games, which typically induce high and localized traffic disruptions, presented greater challenges for imputation. Nonetheless, the model maintained satisfactory accuracy even in these complex situations, benefiting significantly from event-specific contextual features such as type, timing, and attendance size. These features were quantitatively validated through feature importance analysis, where "Number of Attendees" and "Event Type" emerged as dominant predictors, contributing to significant reductions in MAPE.

Third, the framework demonstrated resilience to varying levels of missing data. Performance degradation remained manageable up to moderate levels of missingness (10%), though results clearly suggested the importance of maintaining sensor coverage above 90% to ensure optimal performance. Additionally, effect size analysis using Cohen's d confirmed that longer gaps in data, especially those beyond six hours, significantly impair model accuracy, reinforcing the need for timely data recovery mechanisms.

Lastly, ANOVA and Tukey's HSD offered strong evidence that each event-related feature contributes uniquely and significantly to overall model performance. These findings justify the architectural choice to incorporate multi-source contextual inputs and provide a foundation for future extensions involving real-time event tracking and adaptive traffic management.

Taken together, the results validate that the proposed event-aware architecture is not only accurate but also interpretable and adaptable, making it well-suited for deployment in intelligent transportation systems where dynamic disruptions are common and timely data recovery is critical.

4.6 Discussion

This study advances the state-of-the-art in traffic data imputation and prediction by proposing a unified, event-aware framework that combines classical machine learning, generative deep learning, and spatiotemporal modeling with contextual social event information. The dual-stage imputation strategy—Random Forest followed by GAIN—ensures robustness and precision across various levels of missingness and disruption. Integrating event features directly into both the imputation and prediction pipelines represents a novel approach that enhances model adaptability to urban traffic volatility, particularly during non-recurring events.

Our approach is particularly distinguished by its ability to leverage structured event descriptors (e.g., attendance, type, timing, and proximity) in both the imputation and prediction stages. This contrasts with recent models like KG-GAN (Liu et al., 2024), which utilize external knowledge graphs including POIs and weather data for imputation but lack the event-centric focus necessary for capturing dynamic disruptions. Similarly, while GT-TDI (Zhang et al., 2023) effectively leverages semantic and spatiotemporal graphs for traffic imputation, it does not incorporate domain-specific social event signals, limiting its responsiveness to abrupt traffic changes.

The proposed model demonstrates resilience across a wide range of missingness levels, maintaining high imputation accuracy (MAPE = 5.3%) even under 10% data loss, and achieving stable prediction results in event-affected zones. Furthermore, performance comparisons reveal that our framework outperforms recent diffusion-based imputation methods (Lu et al., 2025), particularly in scenarios with temporally clustered gaps and spatial correlation breakdowns. This suggests that the two-stage architecture, bolstered by attention-guided LSTM prediction, offers superior scalability and responsiveness compared to single-stage or transformer-only baselines.

In addition to technical performance, the study's spatial error mapping and statistical analyses substantiate the model's interpretability and relevance for real-world deployment. The ability to visually and statistically explain how event attributes affect imputation and prediction accuracy highlights a significant advantage over black-box deep learning models. The architecture can thus inform infrastructure planning, sensor maintenance prioritization, and intelligent rerouting strategies under dynamic conditions, making it a strong candidate for operational integration into ITS platforms.

4.7 Conclusion

This paper presents an integrated, event-aware framework for traffic data imputation and prediction that bridges several gaps in the existing literature. The proposed dual-stage imputation module combines the robustness of ensemble learning (Random Forest) with the generative capabilities of GAIN, effectively handling a variety of missingness patterns and levels. On the prediction side, we deploy a composite model comprising Graph Convolutional Networks (GCNs), Long Short-Term Memory (LSTM) units with event-aware gating, and a soft attention mechanism. Together, these components enable accurate, resilient forecasting under both routine and high-disruption conditions.

Experimental validation on a comprehensive dataset from Hamilton, Ontario, demonstrates the practical utility of this framework. Compared to conventional methods and recent state-of-the-art models such as DCRNN (Li et al., 2018b), Graph WaveNet (**?**), GT-TDI (Zhang et al., 2023), and KG-GAN (Liu et al., 2024), our framework achieves superior performance, especially during social events that introduce large, non-recurring traffic disruptions. In scenarios with 5% missing data and 4-hour average gaps, the model achieves a MAPE of 5.3% for imputation and 4.1% for prediction, outperforming even transformer-based or diffusion-enhanced models (Lu et al., 2025) under similar settings.

The analysis of feature importance, proximity-based errors, and statistical significance confirms that event metadata—particularly attendance, event type, and timing—play critical roles in enhancing model fidelity. These insights are valuable not only for model development but also for traffic management agencies aiming to incorporate predictive analytics into smart city infrastructures.

In comparison to prior studies that either treat imputation and prediction as separate problems or fail to integrate structured event features, this study offers a holistic and interpretable approach. It extends current methodologies by embedding external context into both reconstruction and forecasting stages, demonstrating measurable improvements in accuracy, interpretability, and operational readiness.

Future work may explore the real-time integration of event feeds from social media, fusion with multimodal transport data, and deployment in edge computing environments. These directions will further expand the practical impact of the proposed model, reinforcing its role as a cornerstone for resilient, event-aware urban traffic analytics.

Bibliography

Ardestani, A., Yang, H., and Razavi, S. (2025). Enhancing traffic speed prediction accuracy: The multialgorithmic ensemble model with spatiotemporal feature engineering. *Journal of Advanced Transportation*, 2025(1):9941856.

Bae, B., Kim, H., Lim, H., Liu, H. X., Han, L. D., and Freeze, P. B. (2018). Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transportation Research Part C: Emerging Technologies*, 88:124–139.

Duan, Y., Lv, Y., Liu, Y., and Wang, F.-Y. (2016). An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 72:168–181.

Essien, A., Petrounias, I., Sampaio, P., and Sampaio, S. (2021). A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*, 24(4):1345–1368.

Gazis, D. and Liu, C. (2003). Kalman filtering estimation of traffic counts for two network links in tandem. *Transportation Research Part B: Methodological*, 37(10):737–745.

Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929.

Li, L., Li, Y., and Li, Z. (2013). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, 34:108–120.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018a). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018b). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Liu, Y., Shen, G., Liu, N., Han, X., Xu, Z., Zhou, J., and Kong, X. (2024). Traffic data imputation via knowledge graph-enhanced generative adversarial network. *PeerJ Computer Science*, 10:e2408.

Lu, B., Miao, Q., Liu, Y., Tamir, T. S., Zhao, H., Zhang, X., Lv, Y., and Wang, F.-Y. (2025). A diffusion model for traffic data imputation. *IEEE/CAA Journal of Automatica Sinica*, 12(3):606–617.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

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

Ni, D. and Leonard, J. D. (2005). Markov chain monte carlo multiple imputation using bayesian networks for incomplete its data. *Transportation Research Record*, (1935):57–67.

Ni, M., He, Q., and Gao, J. (2014). Using social media to predict traffic flow under special event conditions. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*, Washington, DC.

Smith, B. L., Scherer, W. T., and Conklin, J. H. (2003). Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record*, (1836):132–142.

Song, Y., Luo, R., Zhou, T., Zhou, C., and Su, R. (2024). Graph attention informer for long-term traffic flow prediction under the impact of sports events. *Sensors*, 24(15):4796.

Sun, S., Zhang, C., and Yu, G. (2006). A bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):124–132.

Sun, Z., Zeng, G., and Ding, C. (2021). Imputation for missing items in a stream data based on gamma distribution. *Lecture Notes in Computer Science*, pages 236–247.

Tempelmeier, N., Dietze, S., and Demidova, E. (2020). Crosstown traffic: supervised prediction of impact of planned special events on urban traffic. *Geoinformatica*, 24(2):339–370.

Wang, Y. and Chen, Y. (2018). Limitations of traditional traffic prediction models in the context of social events. *Journal of Transportation Engineering, Part A: Systems*, 144(5):04018025.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. (2019). Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1907–1913, Macao, China.

Ye, Y., Zhang, S., and Yu, J. J. Q. (2021). Traffic data imputation with ensemble convolutional autoencoder. In *Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN.

Yu, B., Yin, H., and 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)*, pages 3634–3640.

Zhang, K., Wu, L., Zheng, L., Xie, N., and He, Z. (2023). Large-scale traffic data imputation with spatiotemporal semantic understanding. *Information Fusion*, 96:102038.

Zhang, Y., Kong, X., Zhou, W., Liu, J., Fu, Y., and Shen, G. (2024). A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., and Li, H. (2019). T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858.

Chapter 5

Conclusion

5.1 Research Contributions

This dissertation presents a comprehensive investigation into the dual challenges of missing traffic data and the complexities of traffic speed prediction during social events, a topic of growing relevance in urban mobility systems. Urban centers are increasingly experiencing non-recurring disruptions caused by concerts, sports games, festivals, and other social activities, which introduce sharp deviations in regular traffic patterns. At the same time, traffic data, which are foundational for intelligent transportation systems (ITS), often suffer from incompleteness due to sensor failures, transmission errors, and insufficient probe vehicle coverage. This research addressed these challenges by developing a unified framework combining robust imputation methods with deep learning-based event-aware prediction architectures.

The first core contribution lies in the development of a two-stage imputation pipeline that integrates ensemble-based and generative approaches. Specifically, Random Forest models were employed to capture local patterns and preserve spatial dependencies, while the Generative Adversarial Imputation Network (GAIN) leveraged global structure and probabilistic data distributions to reconstruct missing entries. This hybrid imputation framework demonstrated superior accuracy and scalability over traditional statistical and shallow machine learning methods. In benchmark experiments on real-world data from Hamilton, Ontario, the proposed imputation method outperformed time-series baselines like ARIMA and spline interpolation by 20–30% in terms of MAPE. Notably, the framework maintained high accuracy even in low-penetration zones and during time intervals with over 40% data missingness, suggesting its practical viability for deployment in sparse sensing environments.

The second major research contribution is the development of an Event-Aware LSTM (EA-LSTM) prediction model that incorporates contextual features derived from a manually curated social event dataset. The dataset included over 560 events, categorized by type (e.g., concerts, sports, public festivals), attendance estimates, temporal features, and spatial coordinates. By integrating these features into a hierarchical deep learning model consisting of Graph Convolutional Networks, Bidirectional LSTM layers, and attention mechanisms, the EA-LSTM achieved state-of-the-art accuracy in forecasting traffic speeds during disrupted periods. For example, the network-wide average MAPE dropped to 3.4%, while near event venues (within 1 km), the MAPE remained below 9%, outperforming conventional LSTM models by a significant margin. These results underscore the importance of accounting for event-specific metadata in urban traffic prediction models, aligning with findings from recent literature that advocate for exogenous context integration (Essien et al., 2021; Song et al., 2024).

Another key insight is the differentiated impact of event types and sizes on traffic flow dynamics. Sports events, which typically have higher and more abrupt peaks in attendance, produced sharper deviations in traffic speed compared to concerts or community festivals. The EA-LSTM model successfully captured these temporal-volatility profiles and localized forecasting needs, demonstrating an ability to generalize across diverse disruptions.

Furthermore, this dissertation contributes methodologically by exploring the synergy between imputation and prediction stages. While many existing studies treat imputation and forecasting as separate tasks, the two-stage pipeline developed here highlights the mutual reinforcement between them. Reliable imputation reduces data noise and instability for downstream predictors, while accurate prediction models provide insights into the likely structure of missing values. This integrated perspective opens up new possibilities for closed-loop systems where prediction informs imputation and vice versa.

5.2 Limitations of the Research

While this dissertation makes important contributions to the field of traffic data imputation and event-aware prediction, several limitations must be acknowledged. These limitations are not only technical constraints but also highlight avenues for future research that can further advance the applicability and robustness of the proposed framework. Acknowledging these limitations also ensures that the findings are interpreted with appropriate caution and contextual awareness.

A primary limitation of this work lies in its reliance on manually compiled datasets of social events. In this study, event records were gathered from a combination of municipal calendars, official announcements, and publicly available online sources. While this approach ensured a high degree of accuracy and consistency, it is inherently time-intensive and not scalable for real-world, real-time deployments. Transportation agencies and traffic management centers cannot be expected to manually collect, validate, and encode event data at scale, particularly in large metropolitan areas where hundreds of events of varying sizes occur each month. The framework therefore depends on external human-driven processes that would not be feasible in operational ITS environments.

Recent advances in natural language processing (NLP) and social media mining have shown promising results in automating event detection. For example, Tao et al. (2022) introduced pipelines capable of extracting temporal and spatial event attributes from unstructured text, while Wang et al. (2025) demonstrated the use of large language models to infer event popularity and attendance from multi-source data feeds. However, these approaches remain limited in terms of reliability, scalability, and cross-regional generalizability. Integrating such automated event detection methods into operational ITS platforms would require robust filtering of noisy data, handling of incomplete or misleading information, and real-time synchronization with traffic management tools. Thus, while this dissertation demonstrates the importance of event features for prediction, the reliance on manually curated event datasets represents a bottleneck for real-world deployment.

A second limitation is the assumption of a static spatial topology in the modeling framework. In this work, the road network was represented as a fixed graph where links and their adjacency remained unchanged throughout the modeling process. While this representation is consistent with much of the existing literature in graph-based traffic prediction, it does not fully capture the dynamic nature of urban networks. Road closures due to construction, lane restrictions, adaptive traffic signal timing, and temporary changes in accessibility during major events all alter the effective topology of the network. By relying on a static adjacency matrix, the framework potentially overlooks short-term structural changes that may significantly influence traffic propagation and congestion patterns.

This limitation underscores the need for future research on dynamic graph representations that can evolve alongside real-time traffic conditions. Techniques such as dynamic graph neural networks (DGNNs) or adaptive adjacency learning modules could allow models to continuously update their understanding of connectivity based on recent traffic observations or incident reports. Incorporating these methods would improve the model's ability to reflect the actual, time-varying structure of urban mobility systems and enhance its robustness during unusual disruptions.

A third limitation concerns the geographical specificity of the case study. The proposed framework was developed and evaluated using probe vehicle and event data from Hamilton, Ontario. While Hamilton provides a diverse and challenging testbed with a wide range of event types, traffic conditions, and network structures, the findings may not be immediately generalizable to other cities. Urban areas differ substantially in terms of road network density, traffic demand profiles, cultural event patterns, and data collection infrastructure. For example, cities with higher public transit usage or denser central business districts may exhibit fundamentally different traffic dynamics in response to social events compared to mid-sized Canadian cities.

Consequently, while the results reported here demonstrate the feasibility and promise of event-aware imputation and prediction, further validation is necessary in other geographical and cultural contexts. Testing the framework across cities with varied data availability, sensor coverage, and mobility behaviors would help establish its broader applicability and adaptability. Without such cross-city validation, conclusions about generalizability must remain cautious.

5.3 Recommendations and Future Work

While this dissertation advances the field of event-aware imputation and prediction for urban traffic systems, there remain many promising avenues for further research. The following recommendations highlight both methodological extensions and practical developments that could strengthen the robustness, scalability, and applicability of the proposed frameworks in diverse urban contexts.

One immediate direction is the integration of additional contextual data sources beyond social event metadata. Weather conditions, for example, are known to have significant impacts on traffic demand, speed, and safety, with rain, snow, or extreme temperatures exacerbating congestion and influencing driver behavior. Similarly, disruptions in public transit, road works, or infrastructure changes such as construction projects can interact with social events to produce complex traffic patterns. Incorporating such multimodal contextual signals into predictive models would provide a more holistic understanding of traffic dynamics. Advances in data fusion techniques, including graph-based multi-source learning, could enable robust integration of heterogeneous data streams, further enhancing model performance in real-world deployment.

Another critical area for future work involves embedding uncertainty quantification into both imputation and prediction stages. While this study focused on improving average accuracy, decision-making in intelligent transportation systems requires an understanding of the confidence associated with predictions. For instance, traffic managers may adopt different strategies if a prediction carries high versus low uncertainty, especially in high-stakes contexts such as emergency response or evacuation planning. Methods such as Bayesian deep learning, Monte Carlo dropout, or ensemble-based uncertainty estimation could be adapted to the proposed frameworks. Embedding predictive uncertainty would support risk-aware traffic management, allowing operators to calibrate interventions based not only on expected conditions but also on the reliability of those forecasts.

The models presented in this dissertation employed a static representation of road network topology, which is a simplification that does not fully capture the dynamic nature of urban networks. Future research should explore the development of dynamic graph structures that can evolve in real time in response to changes in connectivity, road closures, construction activities, or adaptive signal control strategies. Techniques such as dynamic graph neural networks (DGNNs) and attention-based adaptive adjacency learning have shown promise in related fields and could be tailored to traffic applications. Incorporating such dynamic representations would improve the ability of models to capture traffic propagation under highly variable conditions and enhance their adaptability to rapidly evolving disruptions.

Urban traffic conditions are inherently dynamic, shaped by behavioral shifts, infrastructural changes, and emerging mobility modes such as micromobility and ride-sharing. To remain effective in these evolving contexts, future systems should adopt online and adaptive learning mechanisms. Unlike static models that are trained once and then deployed, online learning frameworks continuously update model parameters as new data arrives, enabling real-time adaptation to changing mobility trends. This capability would be especially valuable for modeling non-recurring disruptions, where model performance may degrade quickly if limited to historical patterns. Incorporating reinforcement learning or continual learning paradigms could further improve adaptability by allowing models to actively refine their predictions in response to system feedback.

5.4 Implications of the Research

This dissertation advances the state of knowledge in spatiotemporal traffic modeling by demonstrating how exogenous contextual features can be embedded into deep learning architectures to improve both predictive performance and interpretability. While prior research has often treated traffic prediction as a purely data-driven task, the findings here advocate for domain-informed models that explicitly leverage contextual knowledge such as event characteristics. This represents a paradigm shift away from generic black-box models toward hybrid approaches that combine statistical rigor, machine learning capacity, and domain expertise.

The proposed two-stage imputation framework also contributes to the academic literature by demonstrating the complementary value of ensemble learning and generative modeling in recovering missing traffic data. The integration of Random Forests with Generative Adversarial Imputation Networks (GAIN) illustrates how models with different inductive biases can be combined to achieve both efficiency and accuracy, an approach that could be generalized to other domains of infrastructure data management.

Beyond methodological and technical contributions, this research raises important ethical and societal considerations. The use of event data—particularly when drawn from social media or crowd-sourced platforms—requires careful attention to issues of privacy, consent, and potential bias. For example, not all communities engage equally with social media platforms, which may lead to uneven representation in data-driven event detection systems. Ensuring that predictive models do not inadvertently reinforce inequalities in mobility access or emergency response is a key concern for future implementations.

Furthermore, as predictive models become increasingly embedded in urban decision-making systems, transparency and accountability will be essential. Models that directly inform traffic management decisions must not only be accurate but also interpretable to policymakers and practitioners. The event-aware frameworks proposed in this dissertation, by embedding structured and explainable features, move toward this goal, but additional efforts in explainable AI for transportation remain necessary.

5.5 Practical and Policy Implications

The findings of this dissertation carry direct relevance for the real-world operation of Intelligent Transportation Systems (ITS) and Advanced Traffic Management Systems (ATMS). From a practical standpoint, the proposed Event-Aware LSTM (EA-LSTM) prediction model offers a robust tool for enhancing real-time traffic monitoring and response. By explicitly incorporating event features such as type, location, and attendance, the model provides reliable localized forecasts during disruption-prone periods. These capabilities can support dynamic signal control, congestion mitigation, rerouting protocols, and more effective deployment of emergency services and transit resources around event venues.

The imputation framework also strengthens operational resilience by improving the reliability of traffic datasets under conditions of sparse sensing coverage or communication outages. Many cities still face the challenge of incomplete probe vehicle penetration or aging sensor networks. By filling gaps accurately and efficiently, the framework enables agencies to maintain continuous traffic monitoring and decision support, even under non-ideal conditions.

Policy implications are equally significant. The demonstrated value of structured event metadata highlights the need for systematic and timely data sharing between municipalities, event organizers, and transportation agencies. Institutionalizing real-time event reporting through municipal open data portals, standardized APIs, or direct feeds from ticketing platforms would greatly enhance the effectiveness of predictive models like those developed in this research. Furthermore, public policy should encourage cross-agency and public–private collaboration, including data integration from mobility providers such as ride-hailing companies and transit operators.

At the governance level, frameworks must also ensure privacy, transparency, and accountability in the use of event-aware predictive models. Policies addressing data ownership, anonymization, and ethical use of event information will be critical for enabling adoption in practice while safeguarding public trust. In this sense, the proposed research not only advances methodological innovation but also points toward the institutional and policy infrastructures necessary to realize its full potential in smart city operations.

5.6 Final Remarks

This dissertation contributes a set of practical, interpretable, and context-aware tools for traffic imputation and prediction, responding to pressing challenges in smart transportation systems. By bridging methodological innovation with real-world applicability, this work supports the vision of resilient, adaptive, and human-centered urban mobility infrastructure. Its core contributions which is robust imputation, event-aware forecasting, and integrated modeling offer a strong foundation for the next generation of intelligent transportation systems.

Bibliography

Essien, O., Sharma, K., and Farahmand, A. (2021). Event-aware traffic flow prediction using multi-source data fusion. *Transportation Research Record*, 2675(9):586–599.

Song, Y., Luo, R., Zhou, T., Zhou, C., and Su, R. (2024). Graph attention informer for long-term traffic flow prediction under the impact of sports events. *Sensors*, 24(15):4796.