Journal of Intelligent Decision Making and Information Science

Journal homepage: www.jidmis.org
eISSN: 3079-0875

Transformer-based Mobile Traffic Prediction in Internet of Vehicular Networks

Weiwei Jiang[1,*], Ziteng Wang[2], P. Thamilselvan[3], Shashi Kant Gupta[4,5], and Sushil Kumar Singh[6]

[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2] School of Telecommunications Engineering, Xidian University, Xi'an, 710126, China
[3] Department of Artificial Intelligence and Data Science, KL Deemed to be University, Greenfields, Vaddeswaram, Guntur District, Andhra Pradesh, 522502, India
[4] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India
[5] Computer Science and Engineering, Lincoln University College, Malaysia
[6] Department of Computer Engineering, Marwadi University, Rajkot, Gujarat, India

ARTICLE INFO

Article history:
Received 27 April 2025
Received in revised form 1 June 2025
Accepted 3 June 2025
Available online 4 June 2025

Keywords:
Internet of Vehicular Things, Mobile Traffic Prediction, Deep Learning, Transformer

ABSTRACT

This paper presents a Transformer-based mobile traffic prediction model for the Internet of Vehicular Things (IoVT), addressing the challenges of handling long-term dependencies and improving computational efficiency in mobile traffic forecasting. The proposed model, which integrates a gated residual attention unit (GRAU) and channel embedding (CE) technique, leverages the strengths of the Transformer architecture to enhance predictive accuracy and efficiency while maintaining recurrent dynamics. Through experiments on a real-world dataset, the model demonstrated superior performance over existing baselines in terms of root mean square error (RMSE) and mean absolute error (MAE), showcasing its effectiveness in capturing complex temporal patterns in mobile traffic data. The study contributes to the IoVT by providing a robust prediction tool that can optimize network resource allocation and support the development of intelligent transport systems within smart city frameworks.

1. Introduction

The rapid advancement of the Internet of Things (IoT) has revolutionized various industries, with transportation emerging as a key beneficiary. In this context, the concept of Internet of Vehicular Things (IoVT) has gained prominence, representing a specialized subset of IoT focused on connecting vehicles, infrastructure, and external systems through advanced communication and sensing technologies. IoVT enhances transportation efficiency, safety, and convenience by enabling advanced functionalities such as autonomous driving, predictive maintenance, traffic optimization, and seam-

*Corresponding author.
E-mail address: jww@bupt.edu.cn

https://doi.org/10.59543/jidmis.v2i.14303

less vehicle-to-everything (V2X) communication. It integrates technologies like artificial intelligence, edge computing, and 5G to support dynamic decision-making and improve mobility experiences. By leveraging IoVT, smart cities can reduce congestion, lower emissions, and create a more sustainable transportation ecosystem. IoVT serves as the backbone for Internet of Things-aided Intelligent Transport Systems (IoT-ITS), which aim to enhance the efficiency, safety, and sustainability of transportation networks. By integrating real-time data from interconnected vehicles, road infrastructure, and smart devices, IoT-ITS facilitates dynamic traffic management, predictive maintenance, autonomous vehicle operations, and enhanced commuter experiences. This synergy between IoT and transportation transforms traditional systems into intelligent ecosystems, enabling seamless V2X communication and supporting the vision of smart cities with optimized mobility solutions.

Mobile traffic networks, also known as mobile telecommunications networks, are vital infrastructure for modern communication and connectivity in IoVTs [1]. They enable the transmission of voice, data, and multimedia content between mobile devices, facilitating a wide range of applications and services [2]. Mobile traffic networks play a crucial role in connecting people and devices, enabling communication and collaboration across geographic boundaries. Mobile networks contribute significantly to the global economy by facilitating commerce, enabling remote work, and driving innovation in industries such as e-commerce, finance, and healthcare [3]. Mobile traffic networks have undergone significant evolution over the years, from early analog systems to advanced digital networks capable of high-speed data transmission [4]. The deployment of 5G technology represents the latest milestone in mobile network evolution, offering unprecedented data speeds, low latency, and massive connectivity to support emerging technologies like the Internet of Things (IoT) and augmented reality (AR) [5]. The proliferation of smartphones, tablets, and other connected devices has led to exponential growth in mobile data traffic, placing increasing demands on mobile network infrastructure [6].

In this study, we consider the mobile traffic prediction problem, in which the future traffic amount is predicted with the historical data. There are many practical benefits of mobile traffic prediction [7]. Analyzing traffic patterns helps in determining when to activate or deactivate base stations, optimizing energy consumption and operational costs [8]. Predicting traffic allows for better allocation of network resources such as bandwidth and spectrum, ensuring optimal performance during peak usage times [9]. Anticipating traffic surges enables proactive measures to manage congestion, such as rerouting traffic or adjusting data transmission priorities [10]. Understanding future demand trends aids in long-term capacity planning, guiding investments in infrastructure upgrades and expansions to accommodate growing user needs. Identifying potential network issues before they occur allows for proactive maintenance, reducing downtime and improving overall network reliability. By predicting traffic patterns, content providers can strategically cache or pre-fetch popular content closer to users, reducing latency and improving content delivery speeds [11]. Efficiently managing network resources based on traffic predictions can lead to cost savings for cellular providers, both in terms of operational expenses and capital investments. Ultimately, all of these applications contribute to a better overall user experience by ensuring reliable, high-performance cellular connectivity [12].

Deep learning is the latest solution to mobile traffic prediction. The development of deep learning has seen significant advancements in various types of neural networks, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), and generative adversarial networks (GANs) [13–15]. CNNs revolutionized image recognition tasks by leveraging the concept of local connectivity and shared weights through convolutional layers [16]. LeNet-5, proposed by Yann LeCun in 1998, is considered one of the earliest successful CNN architectures. AlexNet, introduced by Alex Krizhevsky et al. in 2012, demonstrated the effectiveness of deep CNNs in large-scale image classification tasks, winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a significant margin. Since then, various architectures like VGG, GoogLeNet (Inception), ResNet, and EfficientNet have been proposed, achieving improved performance and efficiency in image recognition tasks [17]. RNNs are designed to process sequential data by maintaining internal states or memory. Long short-term memory (LSTM) and gated recurrent unit (GRU) are popular variants of RNNs, addressing the vanishing gradient problem and enabling the model to capture long-range dependencies. Applications of RNNs include natural language processing (NLP), speech recognition, and time series prediction. However, traditional RNNs suffer from difficulties in capturing long-term dependencies due to vanishing or exploding gradients, leading to the development of more advanced architectures like LSTM and GRU [18]. GNNs are specialized neural networks designed to operate on graph-structured data, such as social networks, biological networks, and recommendation systems [19]. They leverage message passing algorithms to aggregate information from neighboring nodes in a graph, enabling tasks like node classification, link prediction, and graph classification. Graph convolutional networks (GCNs) are one of the foundational architectures in this domain, extending the concept of convolutional layers to graph-structured data. GANs consist of two neural networks, a generator and a discriminator, engaged in a minimax game to generate realistic data samples [20]. Introduced by Ian Goodfellow et al. in 2014, GANs have been widely used for generating images, music, text, and other types of data. Variants of GANs include Conditional GANs (cGANs), which condition the generation process on additional information, and Progressive GANs, which progressively grow the generated images in resolution [21].

Deep learning has made significant strides in time series prediction, offering powerful tools for modeling complex temporal relationships and making accurate predictions, including mobile traffic prediction [22]. RNNs are well-suited for time series prediction due to their ability to process sequential data. They can capture temporal dependencies by maintaining hidden states, allowing them to remember past information while processing new inputs [23]. However, traditional RNNs suffer from the vanishing gradient problem, which can hinder their ability to capture long-term dependencies. LSTMs are a variant of RNNs designed to address the vanishing gradient problem. They incorporate gated mechanisms to selectively remember or forget information over long sequences, making them effective for modeling long-range dependencies in time series data. LSTMs have been successfully applied to various time series prediction tasks, including stock market prediction, energy consumption prediction, and weather prediction [24]. GRUs are another variant of RNNs similar to LSTMs but with a simplified architecture. They have fewer parameters than LSTMs, making them faster to train and more computationally efficient while still being effective for capturing temporal dependencies in time series data. GRUs are commonly used in applications where efficiency is critical, such as real-time prediction and online prediction systems [25]. While CNNs are more commonly associated with image processing tasks, they have also been applied to time series prediction, particularly for tasks involving spatial-temporal data. CNNs can learn hierarchical representations of time series data, capturing patterns at different levels of abstraction. They have been used in applications such as traffic flow prediction, where spatial dependencies between different locations play a crucial role in prediction future traffic patterns.

While many models have been proposed for mobile traffic prediction, research gaps still exist and the prediction performance is still below satisfaction. Traditional models like RNNs struggle with long-term dependencies due to issues like the vanishing gradient problem, while Transformer-based models often suffer from high computational complexity in the feedforward layer. These limitations hinder the accuracy and efficiency of mobile traffic prediction. Addressing these limitations is of vital practical importance as it can significantly enhance the ability to forecast traffic patterns, which in turn aids in optimizing network resource allocation, improving the efficiency of intelligent transport systems, and supporting the development of smart cities. Accurate traffic predictions allow for better planning and management of network infrastructure, ensuring reliable and high-performance cellular connectivity, which is essential for the growing demands of modern communication and transportation systems.

To amend the shortcomings of existing time series prediction models, the Transformer structure is introduced by extracting long-term dependencies, e.g., GRAformer [26] is a gated residual attention transformer for multivariate time series prediction and addresses the challenges of RNNs in handling long-term dependencies and the inefficiency of transformer-based models in the feedforward layer. By integrating a gated residual attention unit and channel embedding technique, a GRAformer-based mobile traffic prediction model is proposed in this study, which enhances predictive accuracy and computational efficiency while maintaining recurrent dynamics.

In an IoVT environment, the proposed Transformer-based mobile traffic prediction model can be deployed in several ways. The model can be integrated with vehicle-mounted sensors and road infrastructure devices to collect real-time data on traffic conditions. This data can then be transmitted to cloud or edge computing platforms for processing and analysis using the proposed prediction model. The model can also be deployed in a distributed manner across different nodes within the IoVT network to enable localized traffic predictions and support dynamic decision-making for traffic management. Additionally, the model can be incorporated into V2X communication systems, allowing vehicles and infrastructure to exchange traffic data and predictions for coordinated traffic optimization.

The contributions of this study are summarized as follows:

- To addresses the challenges of RNNs in handling long-term dependencies and the inefficiency of transformer-based models in the feedforward layer, a gated residual attention transformer for mobile traffic prediction in IoVT is introduced.

- By integrating a gated residual attention unit and channel embedding technique, a GRAformer-based mobile traffic prediction model is proposed in this study, which enhances predictive accuracy and computational efficiency while maintaining recurrent dynamics.

- Based on a real-world mobile traffic dataset, the proposed GRAformer-based prediction model is proven effective and outperforms baselines in terms of prediction error metrics.

The rest of this paper is organized as follows. In Section 2, a short literature review is presented. In Section 3, the proposed traffic prediction model is explained. In Section 4, numerical experiments and results are discussed. And in Section 5, the conclusion is drawn.

2. Related Work

2.1 Mobile Traffic Prediction Overview

The landscape of mobile traffic prediction is evolving rapidly with innovative methodologies aimed at enhancing prediction accuracy, computational efficiency, and applicability across various network architectures [27, 28]. In particular, mobile traffic prediction stands out as a critical focus area within this dynamic landscape [29]. Recent advancements have seen the development of novel techniques tailored specifically for mobile networks, addressing the unique challenges and demands of this environment [30]. From lightweight models designed to minimize computational overhead to sophisticated deep learning approaches harnessing the power of clustering and convolutional neural networks, researchers are striving to deliver accurate forecasts while mitigating the complexities inherent in mobile traffic dynamics [31, 32]. These efforts are further bolstered by the integration of multi-feature prediction networks and attention-based spatial-temporal graph architectures, allowing for a comprehensive understanding of mobile traffic patterns [33, 34]. Additionally, the application of digital twin frameworks offers promising avenues for network management and prediction, leveraging real-world data to inform predictive models [35]. As mobile networks continue to evolve alongside emerging technologies like 5G, the pursuit of efficient and reliable mobile traffic prediction remains paramount in ensuring optimal network performance and user experience [36].

An adaptive model for optimal traffic flow prediction using adaptive wildfire optimization (AWO) and spatial pattern super learning (SPSL) to enhance forecast accuracy is presented in [37], which categorizes and predicts traffic flow from a database, outperforming conventional classification models. A method for core network traffic prediction using vertical federated learning and split learning is proposed in [38]. The proposed method allows multiple edge clients to collaboratively train high-quality prediction models using diverse traffic data while maintaining raw data confidentiality. The outcomes of individual dimension-specific prediction models are aggregated through collaboration, forming a partially global model shared among clients to address statistical heterogeneity in distributed machine learning. T-For [39] is an adaptable prediction model for throughput performance in networks, based on neural networks and time series analysis, which estimates future network performance in specific time periods, based on past throughput measurements. Traffic Matrix (TM) records traffic volumes among network nodes and is essential for network operation and management, especially for Traffic Engineering (TE) applications. Directly measuring and collecting TMs in real-time is often not feasible due to cost and operation issues. Prophet is a TM prediction solution for TE that adopts matrix normalization and a TE-centric angle loss function to maintain the critical property and scale invariance of TMs, improving link-level and path-level TE by up to 45.4% and 52.8%, respectively [40]. An interpretable user-behavior-based (UBB) network traffic prediction (NTP) method is proposed in [41], which matches the practical wireless traffic demand very well and improves computational efficiency and predictive accuracy compared with existing methods. The UBB NTP method is based on user behavior, dividing traffic patterns into three components in three time periods for each category of weekday, Saturday, and Sunday. Each component is modeled as a normal-distributed signal, and numerical results show the UBB NTP method matches practical wireless traffic demand well and improves computational efficiency and predictive accuracy compared with existing methods.

2.2 Deep Learning for Mobile Traffic Prediction

More recently, deep learning methods have been proven effective for mobile traffic prediction, including CNN, RNN, GNN and GAN networks [42, 43]. Cellular traffic prediction is a key enabler for automatic network management, but existing studies mainly focus on complex deep neural network models which suffer from extensive computation cost and large model size. A multiservice and multimodal feature fusion network for super-lightweight cellular network traffic prediction is proposed in [44], which consists of a dual feature extraction channel based on grouped 3D convolution for capturing multiservice feature and multimodal feature. A novel method for mobile traffic prediction combining a clustering strategy based on daily traffic peak time with a multi temporal convolutional network and a LSTM model is proposed to address the limitations of recent deep learning studies that only exploit spatiotemporal features, causing high complexity and erroneous predictions [45]. A method for network traffic prediction using an Attention-based Spatial-Temporal Graph Network (ASTGN) is proposed in [46], which integrates temporal and spatial information of network traffic data through a Spatio-Temporal Embedding module in an encoder–decoder architecture and better captures both the temporal and spatial relations between the network traffic. 5GT-GAN-NET [47] is a novel approach to precise internet traffic prediction in 5G smart cities using GANs to create synthetic traffic data that closely mimics real-world statistics, enhancing a new 5GT-GAN-NET-based prediction model and significantly reducing mean square error (MSE) and mean absolute error (MAE) compared to benchmarks.

Hybrid models are also proposed to achieve a better performance than single-structure deep learning models. An intelligent network traffic prediction method is proposed in [48], which integrates the Butterworth filter, convolutional neural network and long short-term memory network to predict network traffic data. The method processes network traffic data to frequency domain and extracts low-frequency component using the Butterworth filter. The residual component is generated by subtracting the low-frequency component from the network traffic sequence. CNN-LSTM prediction models are employed to capture the spatial and temporal features of the data in different frequency bands. The prediction results of the two models are linearly summed to represent the final prediction value. A multi-feature traffic prediction model for cellular networks based on signaling information is proposed in [49], which overcomes the limitations of existing deep learning-based methods by using a multi-feature prediction network that can learn temporal correlations and extract features of signaling information.

To protect user data privacy, federated learning is proposed as a effective distributed learning scheme [50]. A method for optimizing energy efficiency in mobile networks with small base stations (SBSs) by using deep reinforcement learning (DRL) is proposed to create a BS switching operation strategy, which is based on a federated long short-term memory (LSTM) model to predict user traffic demands [51]. A multi-step internet traffic prediction model with variable forecast horizons is proposed for proactive network management, incorporating outlier detection and mitigation techniques with advanced gradient descent and boosting algorithms, which allows multiple edge clients to collaboratively train prediction models using diverse traffic data while maintaining raw data confidentiality [52]. A novel framework for constructing digital twin networks for 4G and 5G cellular systems is presented in [53], using measurement data from user equipment and geographical information and characteristics of 4G and 5G base stations within a specific observation area. The study demonstrates the usefulness of digital twin enabled cellular network management and prediction through representative case studies.

2.3 *Applications of Mobile Traffic Prediction*

Mobile traffic prediction offers numerous practical applications, notably in optimizing base station sleep strategies. By analyzing historical data and real-time traffic patterns, predictive models can forecast future cellular network demand with remarkable accuracy. This foresight enables network operators to dynamically adjust the sleep schedules of base stations, activating or deactivating them based on anticipated traffic volumes. By intelligently managing base station sleep cycles, operators can significantly reduce energy consumption and operational costs while ensuring adequate network coverage and quality of service. Additionally, cellular traffic prediction facilitates proactive maintenance and capacity planning, allowing operators to anticipate and address potential congestion points before they impact user experience.

A base station sleeping strategy in heterogeneous cellular networks based on user traffic prediction is presented in [54], which uses a bidirectional long short-term memory (BLSTM) neural network to predict the future traffic of each user and switches user connections from underutilized micro base stations to other base stations, then switches off the idle micro base stations. To enhance the prediction accuracy of cellular network traffic to provide reliable support for base station sleep control or the identification of malicious traffic, a cellular network traffic prediction method based on multi-modal data feature fusion is proposed in [55], using an attributed K-nearest node (KNN) graph and a dual branch spatio-temporal graph neural network with an attention mechanism (DBSTGNN-Att). Predictive dynamic bandwidth allocation (PDBA) [56] is proposed for for minimizing communication delays in the Internet of Things (IoT) context, which uses adaptive predictive algorithms to forecast resource needs and network conditions, allowing for efficient bandwidth allocation. An efficient adaptive slice flowing prediction in contention random access scheme for optimizing virtual/physical resource in B5G/5G New Radio and core network is proposed in [57], in which the dFRC phase dynamically determines the preamble configuration modes and dynamic flow backoff, while the aRPF phase adaptively differentiates collision domains for different types of flows and minimizes collision probability and determines the optimal virtual/physical resource for achieving optimal SFC corresponding to network slices.

2.4 Research Gaps

While many progresses have been made for mobile traffic prediction, research gaps still exist. Although models based on Transformer structures have begun to be applied to time series prediction, their performance in mobile traffic prediction has not been fully evaluated [58]. Recently, GRAformer [26] is a gated residual attention transformer for multivariate time series prediction. GRAformer addresses the challenges of RNNs in handling long-term dependencies and the inefficiency of transformer-based models in the feedforward layer. By integrating a gated residual attention unit and channel embedding technique, GRAformer enhances predictive accuracy and computational efficiency while maintaining recurrent dynamics. The authors demonstrate the competitive predictive accuracy and accelerated processing through extensive experiments on real-world data for GRAformer. In this study, we aim to fill in the research gap by proposing an efficient Transformer-based structure with GRAformer for mobile traffic prediction in IoVT.

3. Methodology

In this paper, we propose a Transformer-based prediction model for mobile traffic prediction in IoVT. The proposed prediction model integrates a gated residual attention unit (GRAU) and channel embedding (CE) technique to improve predictive accuracy and computational efficiency in multivariate time series prediction [26]. The gated residual attention unit enhances predictive accuracy and computational efficiency, while the channel embedding technique differentiates between series and boosts performance. Recurrent dynamics is maintained, ensuring its suitability for long-term mobile traffic prediction tasks.

The GRAformer model is a novel transformer-based architecture designed for long-term multivariate time series forecasting. In the GRAformer model, the gated residual attention unit (GRAU) and channel embedding (CE) are key components that enhance its predictive capabilities. The GRAU replaces the standard feedforward layers in traditional transformers, combining attention mechanisms with gating for better parameter efficiency and predictive accuracy. It incorporates residual attention to enhance the modeling of complex temporal patterns in time series data by accumulating and adding attention scores from previous layers as residuals before applying the softmax operation. The feedforward network in GRAU is redefined to include element-wise multiplication with the residual attention, ensuring efficient computation while retaining high accuracy. The channel embedding technique, on the other hand, allows the model to distinguish and adapt to the unique characteristics of each variable in multivariate time series data. Each variable (channel) is treated as a separate univariate series, and channel-specific embeddings are learned using a trainable embedding matrix. By embedding channels separately, the model captures variable-specific characteristics and dependencies, improving its ability to generalize across multivariate datasets.

Specifically, GRAU replaces the standard feedforward layers in traditional transformers, combining attention mechanisms and gating for better parameter efficiency and predictive accuracy. GRAU incorporates residual attention to enhance the modeling of complex temporal patterns in time series data. Attention scores from previous layers are accumulated and added as residuals before applying the softmax operation, allowing for better flow and representation of temporal information. The feedforward network (FFN) is redefined to include element-wise multiplication with the residual attention, ensuring an efficient computation while retaining high accuracy. The redefined FFN in GRAU is shown as follows:

$$\mathsf{FFN}_{\mathsf{GRAU}}(x, W_1, W_2) = (xW_1 \odot \mathsf{ResidualAttention}(x))W_2$$

where $\odot$ represents element-wise multiplication, and the attention mechanism integrates residual components for robust temporal learning.

To address the limitations of channel-independent strategies commonly used in multivariate time series models, the channel embedding is introduced to allow the model to distinguish and adapt to the unique characteristics of each variable while retaining computational efficiency and scalability. Each variable (channel) is treated as a separate univariate series. Channel-specific embeddings are learned using a trainable embedding matrix. By embedding channels separately, the model captures variable-specific characteristics and dependencies, improving its ability to generalize across multivariate datasets.

For an input time series, channel embeddings are added to the position encoding and the patch embedding to generate enriched representations for each channel:

$$x_d^{(i)} = W_d x_p^{(i)} + W_{\mathsf{pos}} + W_{\mathsf{channel}}^{(i)}$$

The GRAformer backbone network is shown in Figure 1. The overall architecture adopts a simple encoder-only structure with modifications that include patching, instance normalization, and the above enhancements. The input to the model includes patched and normalized time series data with added channel embeddings. Vanilla multi-head self-attention is used for encoding temporal dependencies in the patched inputs. Instance normalization normalizes each channel using statistics (mean and variance) computed over the look-back window, ensuring consistent input scaling. Inspired by the Vision Transformer (ViT), each univariate time series is divided into patches of fixed size and stride. This patching technique helps reduce input size and improves training efficiency.

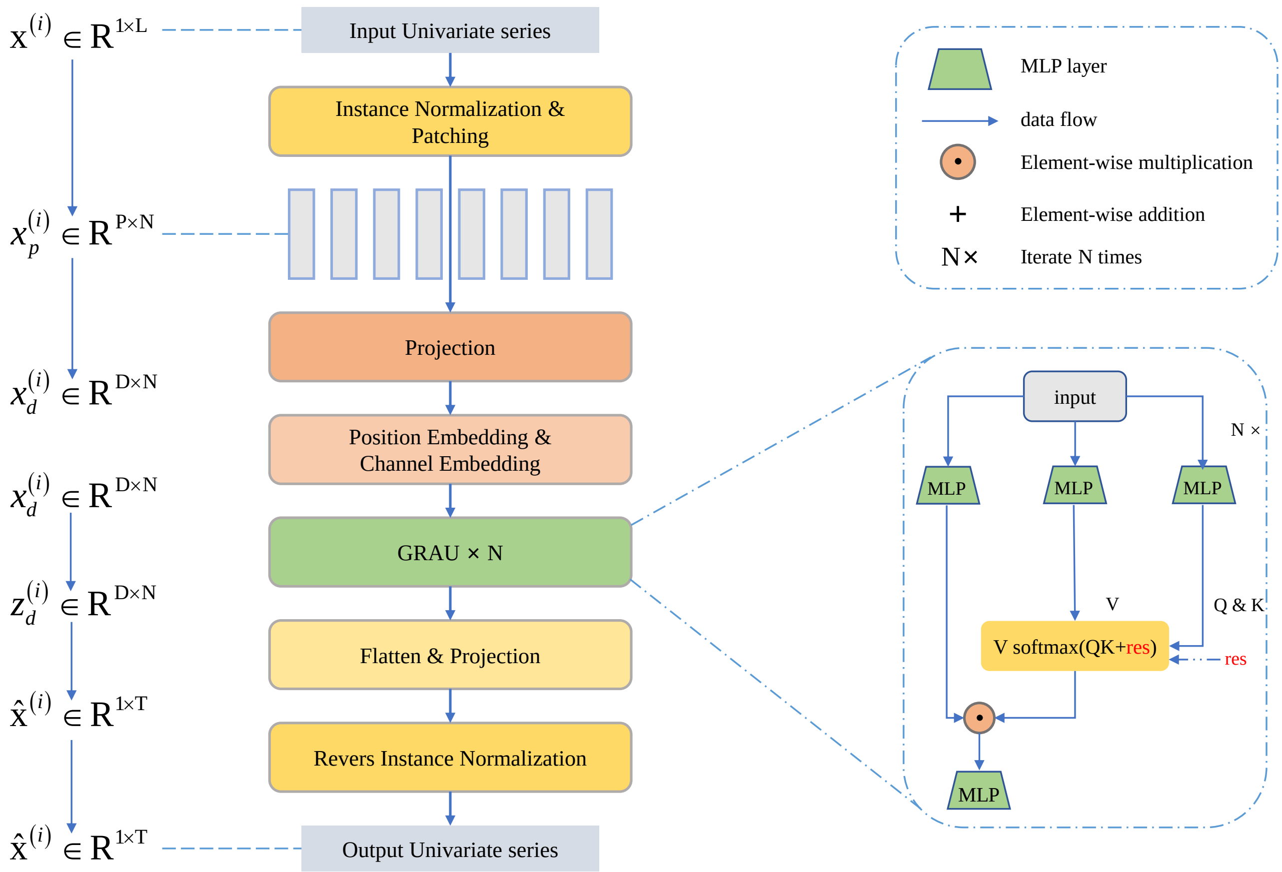
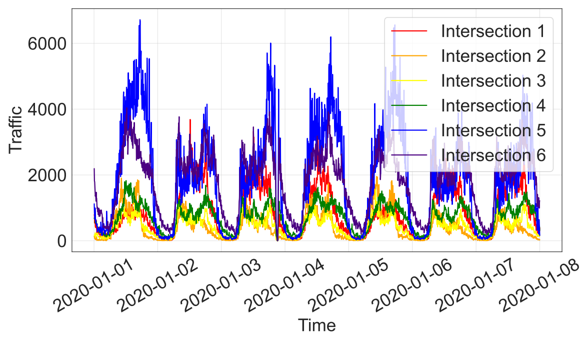


Figure 1: The GRAformer backbone network [26].

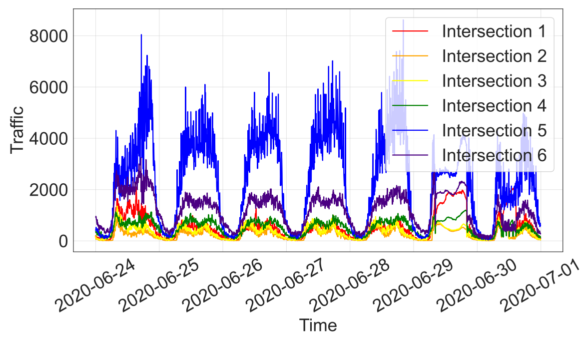
4. Experiments

In this paper, we use a real-world mobile traffic dataset in the IoVT scenario [59]. The dataset used in our experiments was collected from six different intersections located in a major metropolitan area. The data encompasses a six-month period, from January 1 to July 1, 2020. It was gathered by counting vehicles crossing each intersection at five-minute intervals. Vehicle identification was achieved through their International Mobile Equipment Identity (IMEI) information when connecting to base stations positioned at each intersection. Various technologies were employed to collect traffic information, including vehicle-mounted sensors such as GPS, cameras, and LIDAR/RADAR, which gathered real-time data on location, speed, and surrounding conditions. Road infrastructure devices like smart traffic lights, embedded sensors, and closed-circuit television cameras monitored traffic density and flow. V2X communication enables data exchange between vehicles, infrastructure, and other road users. Additionally, mobile and IoT devices, including smartphones and connected dashcams, contribute supplementary data, while drones provide aerial traffic monitoring. This information is aggregated and analyzed using cloud and edge computing platforms, leveraging artificial intelligence and machine learning to deliver actionable insights for traffic management and optimization.

The first week and the last week of the traffic dataset are shown in Figure 2(a) and Figure 2(b), respectively. Mobile traffic exhibits distinct periodical patterns driven by human activities and daily routines, such as commuting, work hours, and leisure time. Peak traffic typically occurs during morning and evening rush hours due to increased mobility and communication needs, while midday usage is often associated with work-related tasks. Late-night traffic tends to decline, except for specific use cases like streaming or gaming. These patterns are influenced by regional and cultural factors, such as holidays or local events, which can temporarily alter the usual trends. Understanding these patterns is critical for optimizing network performance, as it enables resource allocation, capacity planning, and congestion management. Moreover, it has implications for intelligent transport systems, which can align traffic management strategies with these predictable usage cycles to enhance overall efficiency and user experience.



(a)



(b)

Figure 2: The traffic time series. (a) The first week of the traffic dataset; (b) The last week of the traffic dataset.

The deep learning models are implemented with Python and PyTorch. The mean squared error (MSE) loss between true and predicted values is used as the loss function. Adam is used as the optimizer with the default setting in PyTorch. The batch size is set to 128 and the learning rate is set to 0.0001, with 1000 warm-up steps and a cosine decay of 0.8. The early stopping threshold is set to 100 epochs. For the proposed Transformer-based prediction model, the hidden size is set to 256. The number of the Encoder layers is set to 3. The number of the Normalization layer is set to 1. The patch

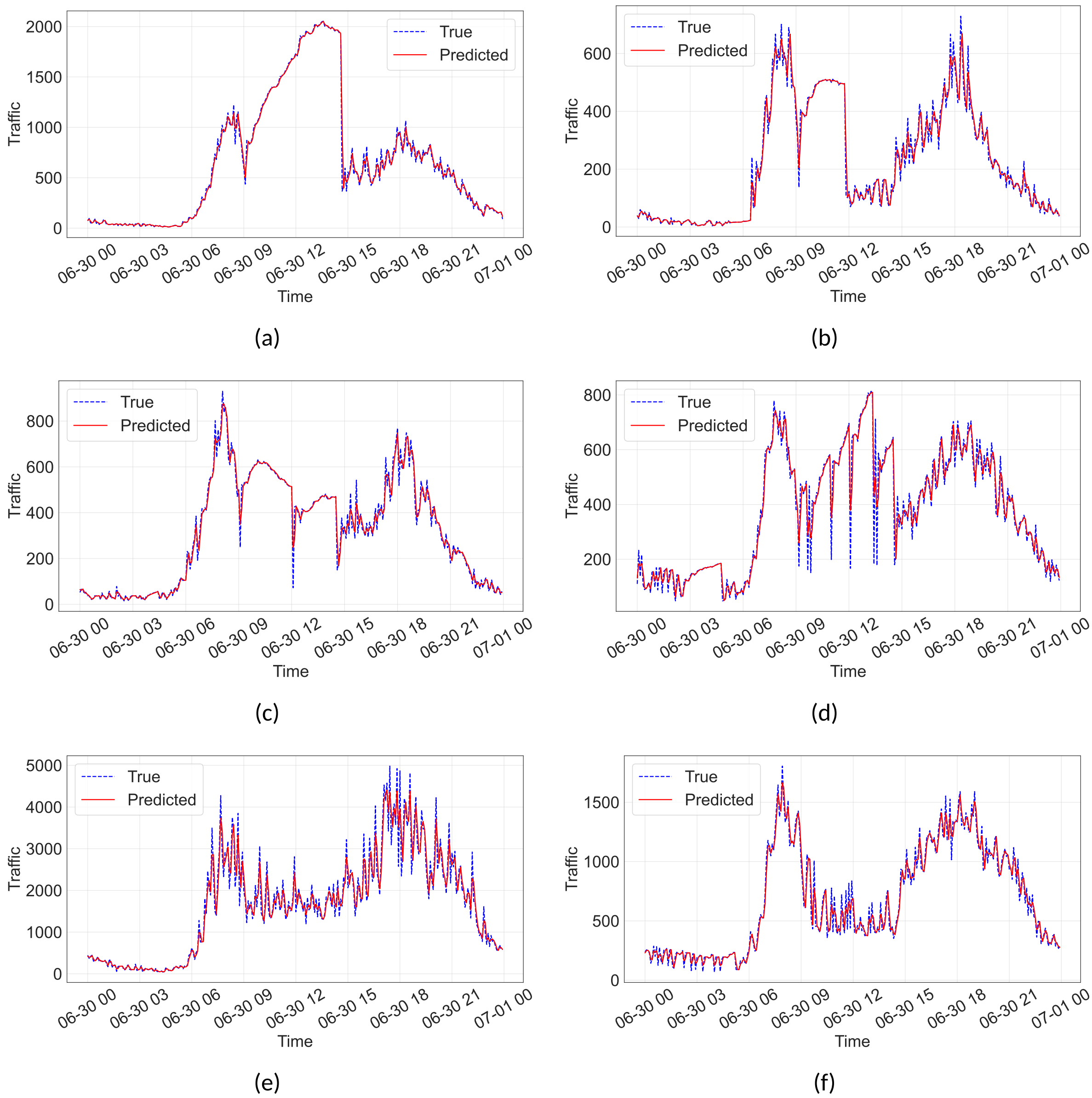


Figure 3: The true vs. predicted values in different intersections. (a) Intersection 1; (b) Intersection 2; (c) Intersection 3; (d) Intersection 4; (e) Intersection 5; (f) Intersection 6.

size is 16 and the stride is set to 8. The dropout rate is set to 0.2.

We divide the whole dataset into a training subset and a test subset with a split ratio of 5:1. The input lookback window size is 24 hours, i.e., 12×5=60 time points. The prediction horizon is 1 hour, e.g., the next 12 time points. A rolling horizon approach is used to validate the model in the test subset. Root mean square error (RMSE) and mean absolute error (MAE) are chosen as the evaluation metrics for prediction accuracy in this study.

For visualization, the true vs. predicted values at different intersections for a single day are presented in Figure 3. In most instances, the proposed Transformer-based prediction model generates forecasts that are closely aligned with the true values. However, for extreme values, the model's predictions exhibit a delay and fail to capture sudden changes accurately. The distribution of prediction errors across various intersections is further depicted in Figure 4. Consistent with previous studies, the error distribution follows a normal distribution, which is expected since, in most cases, the prediction

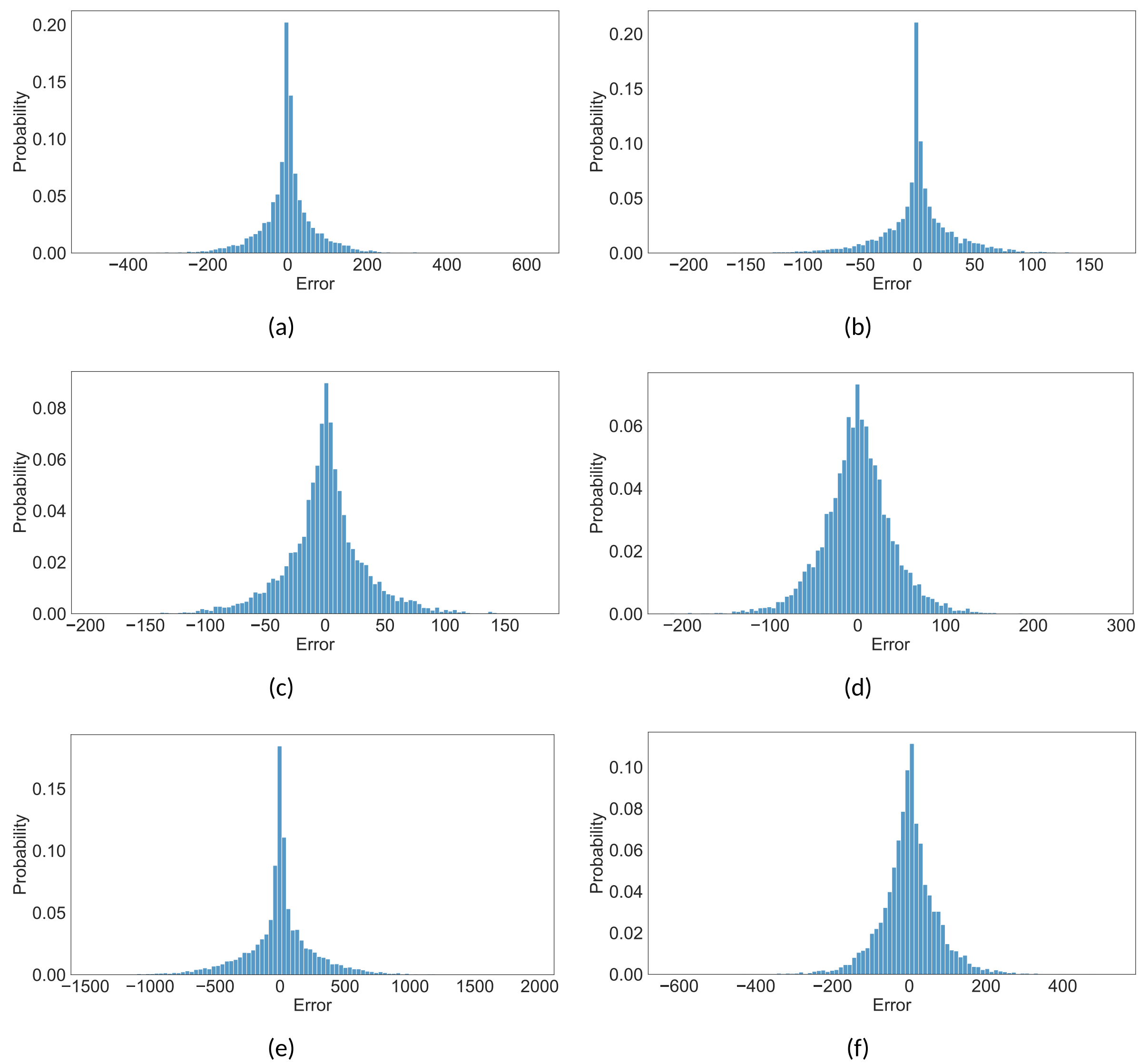


Figure 4: The prediction error distribution in different intersections. (a) Intersection 1; (b) Intersection 2; (c) Intersection 3; (d) Intersection 4; (e) Intersection 5; (f) Intersection 6.

errors are small. Instances of larger errors are less frequent, as reflected in the error distribution.

To validate the proposed Transformer-based prediction model, we also compare it with some recent time series prediction baselines, including Fedformer [60], Autoformer [61], Informer [62], and Pyraformer [63]. For the baselines, the default hyperparameters are used from their original settings.

- Pyraformer [63]: Pyraformer is a time series forecasting model that leverages a pyramidal attention mechanism to capture both short-term and long-term dependencies efficiently. It achieves this by constructing a multi-resolution representation of the time series and exchanging information across different scales, resulting in a model with linear time and space complexity relative to sequence length.

- Informer [62]: Informer is an efficient transformer-based model designed for long sequence time-series forecasting. Informer addresses the challenges of quadratic time complexity and high memory usage in traditional Transformers by introducing a ProbSparse self-attention mechanism and a self-attention distilling operation, which reduce the time and memory complexity.

Additionally, Informer employs a generative style decoder that predicts long sequences in a single forward operation, significantly improving inference speed.

- Autoformer [61]: Autoformer is a decomposition architecture designed for long-term time series forecasting, which incorporates an auto-correlation mechanism to efficiently discover dependencies and aggregate information at the series level.

- Fedformer [60]: Fedformer is a Transformer-based model designed for long-term time series forecasting that combines the strengths of seasonal-trend decomposition and frequency domain analysis. It enhances predictive performance by leveraging the sparse representation of time series in Fourier and Wavelet domains, leading to improved accuracy and linear computational complexity.

The RMSE and MAE results for different prediction models are shown in Table 1 and Table 2, respectively. The results are calculated for each intersection first. An average error is shown in the last line for a fair comparison. It is observed that the proposed model outperforms all baselines in terms of RMSE and MAE metrics.

Table 1: The RMSE results for different prediction models.

Inter.	Pya. [63]	Inf. [62]	Auto. [61]	Fed. [60]	Proposed
1	107.692	102.365	92.140	92.470	**79.674**
2	47.695	45.292	42.616	**40.226**	41.040
3	49.774	49.102	42.948	43.293	**38.967**
4	60.028	57.127	55.132	51.147	**47.136**
5	407.375	386.048	377.98	359.843	**317.326**
6	111.774	111.081	97.376	**94.425**	94.830
Avg.	130.723	125.169	118.032	113.567	**103.162**

Table 2: The MAE results for different prediction models.

Inter.	Pya. [63]	Inf. [62]	Auto. [61]	Fed. [60]	Proposed
1	67.387	61.831	57.210	55.590	**50.693**
2	28.831	30.214	28.355	26.966	**23.274**
3	34.325	31.124	30.590	29.691	**27.803**
4	41.476	42.826	40.965	38.066	**34.349**
5	253.222	249.637	242.959	211.956	**204.565**
6	81.317	75.902	66.916	69.204	**64.358**
Avg.	84.426	81.922	77.832	71.912	**67.507**

To further evaluate the functionality of each component in the proposed GRAformer-based traffic prediction model, an ablation study is conducted by comparing the proposed model with and without GRAU and CE. The RMSE and MAE results for different variants are shown in Table 3 and Table 4, respectively. It is observed that the proposed model with both GRAU and CE components outperform the remaining variants, in terms of both RMSE and MAE metrics.

The experimental results demonstrate the superiority of the proposed GRAformer-based prediction model, which can be attributed to several factors. The integration of GRAU enhances the model's ability to capture complex temporal patterns by combining attention mechanisms with gating for better parameter efficiency and predictive accuracy. The residual attention in GRAU allows for better flow

Table 3: The RMSE results for different variants.

Intersection	Without GRAU&CE	Without GRAU	Without CE	Proposed
1	83.041	82.134	80.150	79.674
2	44.310	43.304	42.238	41.040
3	41.992	40.020	40.215	38.967
4	51.384	49.428	49.289	47.136
5	343.847	328.108	325.959	317.326
6	101.525	99.543	99.245	94.830
Avg.	111.017	107.090	106.183	103.162

Table 4: The MAE results for different variants.

Intersection	Without GRAU&CE	Without GRAU	Without CE	Proposed
1	52.623	55.012	51.740	50.693
2	24.515	25.004	25.266	23.274
3	28.204	29.007	29.645	27.803
4	34.976	35.053	34.731	34.349
5	216.681	205.164	208.299	204.565
6	65.404	69.663	70.438	64.358
Avg.	70.401	69.817	70.020	67.507

and representation of temporal information, while the element-wise multiplication with the residual attention in the feedforward network ensures efficient computation. The channel embedding technique further boosts performance by allowing the model to distinguish and adapt to the unique characteristics of each variable in multivariate time series data. However, the model's limitations are also evident. For extreme values, the model's predictions exhibit a delay and fail to capture sudden changes accurately. This suggests that while the GRAformer excels in capturing general traffic patterns, it may struggle with highly volatile or anomalous traffic conditions, indicating areas for potential improvement in future research.

5. Conclusion

The rapid growth of the IoVT has underscored the critical need for accurate mobile traffic prediction in smart transportation systems. This study introduces a Transformer-based mobile traffic prediction model that leverages the gated residual attention transformer (GRAformer) to address the challenges of long-term dependencies and computational inefficiency in traditional models. The proposed model integrates a GRAU and CE technique, enhancing predictive accuracy and computational efficiency while maintaining recurrent dynamics.

Through extensive experiments on a real-world mobile traffic dataset, our model demonstrated superior performance over several state-of-the-art baselines, including Fedformer, Autoformer, Informer, and Pyraformer, in terms of both RMSE and MAE. The results validate the effectiveness of the GRAformer-based prediction model in capturing complex temporal patterns and its robustness in various traffic conditions. An ablation study further confirmed the contributions of the GRAU and CE components to the overall performance, highlighting their significance in improving the model's predictive capabilities. The proposed model's ability to accurately forecast mobile traffic not only aids in optimizing network resource allocation but also contributes to the development of intelligent transport systems, aligning with the growing demands of smart cities for efficient mobility solutions.

The proposed Transformer-based mobile traffic prediction model has broad potential applications in the field of intelligent transportation systems. Beyond the IoVT scenario explored in this study, it can be adapted to other transportation-related prediction tasks such as traffic flow prediction on highways, transit schedule optimization, and even infrastructure planning for public transportation systems. The model's ability to capture complex temporal patterns makes it suitable for predicting passenger demand for ridesharing services or optimizing delivery routes for logistics companies. In terms of future research directions, one avenue is to further optimize the model by incorporating additional data sources such as weather conditions, special events, or social media trends that can impact traffic patterns. Another direction is to enhance the model's real-time processing capabilities to enable more immediate predictions and responses. Additionally, exploring hybrid models that combine the strengths of GRAformer with other advanced architectures like graph neural networks could further improve prediction accuracy for more complex transportation networks. Finally, extending the model to handle spatiotemporal data more effectively could open up new possibilities for urban traffic management and smart city development.

References

[1] Singh, P. R., Singh, V. K., Yadav, R., & Chaurasia, S. N. (2023). 6g networks for artificial intelligence-enabled smart cities applications: A scoping review. *Telematics and Informatics Reports*, *9*, 100044. https://doi.org/10.1016/j.teler.2023.100044

[2] Xu, Z., Tang, N., Xu, C., & Cheng, X. (2021). Data science: Connotation, methods, technologies, and development. *Data Science and Management*, *1*(1), 32–37. https://doi.org/10.1016/j.dsm.2021.02.002

[3] Gill, S. S., Wu, H., Patros, P., Ottaviani, C., Arora, P., Pujol, V. C., Haunschild, D., Parlikad, A. K., Cetinkaya, O., Lutfiyya, H., et al. (2024). Modern computing: Vision and challenges. *Telematics and Informatics Reports*, 100116. https://doi.org/10.1016/j.teler.2024.100116

[4] Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2023.06.001

[5] Lu, Y., Wang, W., Bai, R., Zhou, S., Garg, L., Bashir, A. K., Jiang, W., & Hu, X. (2025). Hyper-relational interaction modeling in multi-modal trajectory prediction for intelligent connected vehicles in smart cites. *Information Fusion*, 102682. https://doi.org/10.1016/j.inffus.2024.102682

[6] Yang, B., Zhou, J., Zhang, S., Xing, Y., Jiang, W., & Xu, L. (2024). Lightweight knowledge distillation and feature compression model for user click-through rates prediction in edge computing scenarios. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2024.3446640

[7] Jiang, W. (2022a). Graph-based deep learning for communication networks: A survey. *Computer Communications*, *185*, 40–54. https://doi.org/10.1016/j.comcom.2021.12.015

[8] Jiang, W., Han, H., He, M., & Gu, W. (2024). Ml-based pre-deployment sdn performance prediction with neural network boosting regression. *Expert Systems with Applications*, *241*, 122774. https://doi.org/10.1016/j.eswa.2023.122774

[9] Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., & Feng, Q. (2024). Modality fusion vision transformer for hyperspectral and lidar data collaborative classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. https://doi.org/10.1109/JSTARS.2024.3415729

[10] Hong, S., Yue, T., You, Y., Lv, Z., Tang, X., Hu, J., & Yin, H. (2025). A resilience recovery method for complex traffic network security based on trend forecasting. *International Journal of Intelligent Systems*, *2025*(1), 3715086.

[11] Jiang, W., He, M., & Gu, W. (2022). Internet traffic prediction with distributed multi-agent learning. *Applied System Innovation*, *5*(6), 121. https://doi.org/10.3390/asi5060121

[12] Ali, A., Ullah, I., Singh, S. K., Jiang, W., Alturise, F., & Bai, X. (2025). Attention-driven graph convolutional networks for deadline-constrained virtual machine task allocation in edge computing. *IEEE Transactions on Consumer Electronics*.

[13] He, M., Jiang, W., & Gu, W. (2023). Making wines smarter: Evidence from an interpretable learning paradigm. *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 60–65. https://doi.org/10.1109/SECON58729.2023.10287467

[14] Zheng, Y., Jiang, W., et al. (2022). Evaluation of vision transformers for traffic sign classification. *Wireless Communications and Mobile Computing*, *2022*. https://doi.org/10.1155/2022/3041117

[15] Zhou, Z., Bao, Z., Jiang, W., Huang, Y., Peng, Y., Shankar, A., Maple, C., & Selvarajan, S. (2024). Latent vector optimization-based generative image steganography for consumer electronic applications. *IEEE Transactions on Consumer Electronics*. https://doi.org/10.1109/TCE.2024.3354824

[16] XIA, Z., LIU, Y., WANG, X., ZHANG, F., CHEN, R., & JIANG, W. (2024). Infrared and visible image fusion via hybrid variational model. *IEICE TRANSACTIONS on Information and Systems*, *107*(4), 569–573. https://doi.org/10.1587/transinf.2023EDL8027

[17] Chen, X., Li, H., Li, C., Jiang, W., & Zhou, H. (2023). Single image dehazing based on sky area segmentation and image fusion. *IEICE TRANSACTIONS on Information and Systems*, *106*(7), 1249–1253. https://doi.org/10.1587/transinf.2023EDL8010

[18] Li, D., Hou, J., & Gao, W. (2022). Instrument reading recognition by deep learning of capsules network model for digitalization in industrial internet of things. *Engineering Reports*, e12547. https://doi.org/10.1002/eng2.12547

[19] Haghshenas, S. S., Astarita, V., Guido, G., Seraji, M. H. M., Gonzalez, P. A. A., Haghdadi, A., & Haghshenas, S. S. (2022). Assessment of machine learning techniques and traffic flow: A qualitative and quantitative analysis. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE32021062

[20] Zhao, M., & Zhang, Y. (2022). Gan-based deep neural networks for graph representation learning. *Engineering Reports*, e12517. https://doi.org/10.1002/eng2.12517

[21] Lim, L. W. K. (2022). Implementation of artificial intelligence in aquaculture and fisheries: Deep learning, machine vision, big data, internet of things, robots and beyond. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE3202803

[22] Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, *184*, 115537. https://doi.org/10.1016/j.eswa.2021.115537

[23] Jiang, W. (2022b). Deep learning based short-term load forecasting incorporating calendar and weather information. *Internet Technology Letters*, *5*(4), e383. https://doi.org/10.1002/itl2.383

[24] Jiang, W. (2022c). Bike sharing usage prediction with deep learning: A survey. *Neural Computing and Applications*, *34*(18), 15369–15385. https://doi.org/10.1007/s00521-022-07380-5

[25] Pedrycz, W. (2023). Autonomous and sustainable machine learning: Pursuing new horizons of intelligent systems. *AI Autonomous Syst.*, *2023*(1), 0002. https://doi.org/10.55092/aias20230002

[26] Yang, C., Wang, Y., Yang, B., & Chen, J. (2024). Graformer: A gated residual attention transformer for multivariate time series forecasting. *Neurocomputing*, *581*, 127466. https://doi.org/10.1016/j.neucom.2024.127466

[27] Jiang, W. (2022d). Cellular traffic prediction with machine learning: A survey. *Expert Systems with Applications*, *201*, 117163. https://doi.org/10.1016/j.eswa.2022.117163

[28] Hou, C., Wu, J., Cao, B., & Fan, J. (2021). A deep-learning prediction model for imbalanced time series data forecasting. *Big Data Mining and Analytics*, *4*(4), 266–278. https://doi.org/10.26599/BDMA.2021.9020011

[29] Jiang, W., Zhang, Y., Han, H., Huang, Z., Li, Q., & Mu, J. (2024). Mobile traffic prediction in consumer applications: A multimodal deep learning approach. *IEEE Transactions on Consumer Electronics*. https://doi.org/10.1109/TCE.2024.3361037

[30] Liu, J., Jiang, W., Han, H., He, M., & Gu, W. (2023). Drought level prediction based on meteorological data and deep learning. *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 74–75. https://doi.org/10.1109/SECON58729.2023.10287478

[31] Jiang, W., & Luo, J. (2022a). Big data for traffic estimation and prediction: A survey of data and tools. *Applied System Innovation*, *5*(1), 23. https://doi.org/10.3390/asi5010023

[32] Zhang, J., & Xu, Q. (2021). Attention-aware heterogeneous graph neural network. *Big Data Mining and Analytics*, *4*(4), 233–241. https://doi.org/10.26599/BDMA.2021.9020008

[33] Jiang, W., Luo, J., He, M., & Gu, W. (2023). Graph neural network for traffic forecasting: The research progress. *ISPRS International Journal of Geo-Information*, *12*(3), 100. https://doi.org/10.3390/ijgi12030100

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

[35] Sun, X., He, Y., Wu, D., & Huang, J. Z. (2023). Survey of distributed computing frameworks for supporting big data analysis. *Big Data Mining and Analytics*, *6*(2), 154–169. https://doi.org/10.26599/BDMA.2022.9020014

[36] Qu, X., Guan, C., Xie, G., Tian, Z., Sood, K., Sun, C., & Cui, L. (2023). Personalized federated learning for heterogeneous residential load forecasting. *Big Data Mining and Analytics*, *6*(4), 421–432. https://doi.org/10.26599/BDMA.2022.9020043

[37] Jain, R., Dhingra, S., Joshi, K., & Grover, A. (2024). An adaptive model of optimal traffic flow prediction using adaptive wildfire optimization and spatial pattern super learning. *Wireless Networks*, 1–9. https://doi.org/10.1007/s11276-023-03609-w

[38] Li, P., Guo, C., Xing, Y., Shi, Y., Feng, L., & Zhou, F. (2024). Core network traffic prediction based on vertical federated learning and split learning. *Scientific Reports*, *14*(1), 4663. https://doi.org/10.1038/s41598-024-53193-y

[39] Portela, A. L., Ribeiro, S. E., Menezes, R. A., de Araujo, T., & Gomes, R. L. (2024). T-for: An adaptable forecasting model for throughput performance. *IEEE Transactions on Network and Service Management*. https://doi.org/10.1109/TNSM.2024.3349701

[40] Zhang, Y., Han, N., Zhu, T., Zhang, J., Ye, M., Dou, S., & Guo, Z. (2023). Prophet: Traffic engineering-centric traffic matrix prediction. *IEEE/ACM Transactions on Networking*. https://doi.org/10.1109/TNET.2023.3293098

[41] Wang, L., Zhang, J., Zhang, Z., & Zhang, J. (2023). Analytic network traffic prediction based on user behavior modeling. *IEEE Networking Letters*. https://doi.org/10.1109/LNET.2023.3278498

[42] Jiang, W. (2022e). Internet traffic matrix prediction with convolutional lstm neural network. *Internet Technology Letters*, *5*(2), e322. https://doi.org/10.1002/itl2.322

[43] Jiang, W. (2022f). Internet traffic prediction with deep neural networks. *Internet Technology Letters*, *5*(2), e314. https://doi.org/10.1002/itl2.314

[44] Li, Y., Hao, M., Sun, X., & Zhang, H. (2023). Modeling super-lightweight cellular traffic prediction via multiservice and multimodal feature fusion network. *IEEE Networking Letters*. https://doi.org/10.1109/LNET.2023.3329744

[45] Park, J., Mwasinga, L. J., Yang, H., Raza, S. M., Le, D.-T., Kim, M., Chung, M. Y., & Choo, H. (2024). Regional correlation aided mobile traffic prediction with spatiotemporal deep learning. *2024*

IEEE 21st Consumer Communications & Networking Conference (CCNC), 566–569. https://doi.org/10.1109/CCNC51664.2024.10454764
[46] Peng, Y., Guo, Y., Hao, R., & Xu, C. (2024). Network traffic prediction with attention-based spatial-temporal graph network. *Computer Networks*, 110296. https://doi.org/10.1016/j.comnet.2024.110296
[47] Pandey, C., Tiwari, V., Rodrigues, J. J., & Roy, D. S. (2024). 5gt-gan-net: Internet traffic data forecasting with supervised loss based synthetic data over 5 g. *IEEE Transactions on Mobile Computing*. https://doi.org/10.1109/TMC.2024.3364655
[48] Hu, X., Liu, W., & Huo, H. (2024). An intelligent network traffic prediction method based on butterworth filter and cnn–lstm. *Computer Networks*, *240*, 110172. https://doi.org/10.1016/j.comnet.2024.110172
[49] Liu, J., Tang, X., Zhu, G., Cheng, X., Zhao, L., & Li, H. (2023). Multi-feature traffic prediction based on signaling information for cellular network. *IEEE Transactions on Vehicular Technology*. https://doi.org/10.1109/TVT.2023.3319351
[50] Jiang, W., Han, H., Zhang, Y., & Mu, J. (2024). Federated split learning for sequential data in satellite–terrestrial integrated networks. *Information Fusion*, *103*, 102141. https://doi.org/10.1016/j.inffus.2023.102141
[51] Park, H., & Yoon, S. H. (2024). Deep reinforcement learning for base station switching scheme with federated lstm-based traffic predictions. *ETRI Journal*. https://doi.org/10.4218/etrij.2023-0065
[52] Saha, S., Haque, A., & Sidebottom, G. (2024). Multi-step internet traffic forecasting models with variable forecast horizons for proactive network management. *Sensors*, *24*(6), 1871. https://doi.org/10.3390/s24061871
[53] Saqib, N. U., Song, S., Xie, H., Cao, Z., Hahm, G.-J., Cheon, K.-Y., Kwon, H., Park, S., Jeon, S.-W., & Jin, H. (2024). Digital twin enabled cellular network management and prediction. *ICT Express*. https://doi.org/10.1016/j.icte.2024.02.011
[54] Wang, X., Lyu, B., Guo, C., Xu, J., & Zukerman, M. (2023). A base station sleeping strategy in heterogeneous cellular networks based on user traffic prediction. *IEEE Transactions on Green Communications and Networking*. https://doi.org/10.1109/TGCN.2023.3324486
[55] Cai, Z., Tan, C., Zhang, J., Zhu, L., & Feng, Y. (2024). Dbstgnn-att: Dual branch spatio-temporal graph neural network with an attention mechanism for cellular network traffic prediction. *Applied Sciences*, *14*(5), 2173. https://doi.org/10.3390/app14052173
[56] Chakour, I., Daoui, C., Baslam, M., Sainz-de-Abajo, B., & Garcia-Zapirain, B. (2024). Strategic bandwidth allocation for qos in iot gateway: Predicting future needs based on iot device habits. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3351111
[57] Chang, B.-J., & Lin, Y.-T. (2023). Adaptive flowing traffic prediction in contention random access for optimizing virtual/physical resource in b5g/5g new radio and core network. *IEEE Transactions on Network Science and Engineering*. https://doi.org/10.1109/TNSE.2023.3334744
[58] Chen, Z., Ma, M., Li, T., Wang, H., & Li, C. (2023). Long sequence time-series forecasting with deep learning: A survey. *Information Fusion*, *97*, 101819. https://doi.org/10.1016/j.inffus.2023.101819
[59] Lin, C.-Y., Su, H.-T., Tung, S.-L., & Hsu, W. H. (2021). Multivariate and propagation graph attention network for spatial-temporal prediction with outdoor cellular traffic. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3248–3252. https://doi.org/10.1145/3459637.3482152
[60] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. *International conference on machine learning*, 27268–27286. https://doi.org/10.48550/arXiv.2201.12740

[61] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, *34*, 22419–22430. https://doi.org/10.48550/arXiv.2106.13008
[62] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI conference on artificial intelligence*, *35*, 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325
[63] Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2022). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. *The Tenth International Conference on Learning Representations (ICLR 2022)*, 1–20.