# Incremental Gate State Output Decomposition Model for Highway Traffic Forecasting Using Toll Collection Data

**Liang Yu[1], Ming Li[2], Kaifeng Liu[3,*], Xiangping Cheng[4]**
*[1]Jiangxi Traffic Monitoring Command Center, Nanchang 330036, China*
*[2]Jiangxi Transportation Institute Co., Ltd, Nanchang 330200, China*
*[3]Hangzhou iGreenTrans Co., Ltd., Hangzhou, Zhejiang 310051, China*
*[4]Jiangxi Academy of Sciences, Nanchang 330029, China*
*ylncu@jt.jiangxi.gov.cn; mingli@jxjtkyy.com; [*]kaifengliu@igreentrans.tech; chengxiangping@jxas.ac.cn*

***Abstract*—Traffic flow on long-distance highways, especially at sections with multi-interchanges and ramps, exhibits nonlinear trends affected by long-term and short-term spatiotemporal dependencies, resulting limited fitting capabilities for the major applied spatiotemporal forecasting models in use. This paper tackles this challenge by integrating an incremental gate state output decomposition (IGOD) mechanism into the recurrent neural network (RNN) model framework, accounting for the interdependencies of spatiotemporal traffic data. The proposed method improves the ability of the RNN model to estimate traffic data series by segmenting consecutive time intervals and accumulating incremental changes across these time intervals, allowing for more precise traffic predictions. This study also explores how threshold amplitudes affect prediction effectiveness. We applied it to real traffic data from segment k602+630 to k625+420 on the Changjiu Highway. The results demonstrate that the proposed model consistently exhibits robustness, with variations in threshold magnitude having little impact on its prediction accuracy.**

***Index Terms*—Highway traffic prediction; Incremental gate state output decomposition; Self-attention; Toll collection data.**

## I. INTRODUCTION

The electronic toll collection (ETC) gantry systems deployed on highway segments integrate the main functions of toll collection, traffic monitoring, and communication on highways, effectively improving the capabilities of traffic flow monitoring, event detection, safety alerts, and emergency response in the highway network [1]. The roadside units (RSUs) of the ETC gantry system read information from vehicle onboard units (OBUs) through radio frequency devices, enabling precise recording of traffic vehicles about their time of entry and exit and location at toll booths [2], [3]. This system facilitates timely and precise data collection for highway traffic monitoring and operations, including predicting peak hour traffic volume, congested segments, average travel time, etc., so that transportation authorities can develop strategies and manage traffic concerns efficiently.

As numerous traffic sensors have become more and more prevalent, a wide range of traffic prediction models leverage both historical and real-time data collected from traffic sensors. Accordingly, traffic management systems are increasingly adopting data-driven approaches to enhance their effectiveness [4]–[6]. The challenges of traffic flow prediction have been re-evaluated, with increasing attention given to deep architecture models that can effectively utilise extensive traffic data. The adoption of deep architecture models reflects the growing recognition of large datasets as valuable resources for improving the accuracy and reliability of traffic flow prediction. The advent of deep learning has led to development of a suitable traffic prediction model based on deep neural networks. Deep learning has recently gained significant attention in both academia and industry, demonstrating successful applications. This technique uses complex hierarchical structures, known as layered architectures or intricate frameworks, to uncover implicate characteristics from the data, transitioning from low-level to high-level representations [7]. Through this, deep learning models uncover complex patterns within the data.

In this study, we utilised field-collected ETC samples to investigate the spatiotemporal distribution of highway traffic. To address this, an enhanced RNN model, integrated with incremental gate state output decomposed (IGOD) factors, was purposefully designed, leveraging its inherent flexibility and dual data mining capabilities, thus significantly augmenting the precision of traffic flow predictions. This is an endeavour of utmost significance given the pivotal role these predictions play in the realm of efficient traffic management and planning. The contributions of this study

Manuscript received 7 August, 2024; accepted 15 December, 2024.
This research is supported by the Jiangxi Provincial Key R&D Program under Grant No. 20201BBE51015; Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB204066; Jiangxi Academy Institute Key R&D Program under Grant No. 2022YSBG21017.

can be summarised as follows.

– Our proposed model not only forecasts traffic flow values but also incorporates the directional aspect of traffic movement. This innovative approach accounts for both the quantity of traffic and the movement patterns, offering a more holistic perspective on traffic behaviour.

– Our study emphasises the versatility of the proposed mechanism, demonstrating its potential for easy adaptation to other models and prediction tasks. This scalability represents a significant contribution, not only in terms of its positive impact on traffic flow prediction, but also holds promise for broader applications across diverse domains where flexible output prediction and dual data mining capabilities are beneficial.

The remainder of this paper is organised as follows. Section II provides an extensive review of recent studies of traffic flow prediction. Section III offers a detailed implementation of the proposed IGOD mechanism, including training of the prediction network and data processing. In Section IV, we explore the discussion of data sources and present the experimental results. Finally, Section V presents conclusions.

## II. Related Works

Traffic forecasting has played a pivotal role in modern intelligent transportation systems and constitutes an indispensable component of traffic control and planning. Despite recent significant advances in data collection and modelling methodologies, there is still a pressing need to enhance the precision and reliability of traffic forecasting. The inherent complexity and variability of traffic patterns pose challenges in this domain. Many factors, such as weather conditions, accidents, roadworks, and special events, can exert influence on traffic flow [8]. In aim to achieve dependable traffic forecasting, some early researches are focussed on parametric models, but the research direction has now shifted to include nonparametric and hybrid models. In the pursuit of dependable traffic forecasting, early research primarily focussed on parametric models, but the research direction has since shifted to include nonparametric and hybrid models [9]. Typical parametric models, such as moving average (MA) model, autoregressive moving average (ARMA) model, autoregressive integrated moving average (ARIMA) model, and their evolution have been demonstrated in this regard [10]–[12]. However, these models heavily depend on the assumption of stationarity and face challenges in capturing the complex and uncertain nature of traffic flow due to its nonlinear characteristics. Furthermore, some researchers have used nonparametric models to analyse the complex dynamics of spatiotemporal traffic patterns [13]–[15]. However, large-scale networks can pose computational challenges, e.g., making support vector machine (SVM) models less efficient for traffic flow forecasting; also, simplifying the architecture of an artificial neural network (ANN) may struggle to capture the spatial dependencies of traffic networks.

In comparison, deep neural network models, i.e., the recurrent neural network (RNN), convolutional neural network (CNN), graph convolution network (GCN), and derivative models, offer greater accuracy in expressing the complex patterns of traffic data [16], [17]. Graph neural networks (GNNs) have been explored to address complex spatiotemporal forecasting tasks by leveraging the graph structure of data to model dependencies across spatial and temporal dimensions [18]. Recent studies have highlighted the effectiveness of spatiotemporal GNN architectures in capturing interseries dependencies within periodic traffic data, enabling the representation of nonlinear features in traffic series data [19]. A typical modelling method employs 1-D convolution to extract features in the time dimension and then uses GCN to capture the spatial dependencies of each node [20], [21]. However, these models are limited in the receptive field size of the convolution operation, resulting in poorer performance for long-term predictions. To address this issue, Wang, Ren, and Li [22] introduced dilated convolution, which can effectively handle long time series. Another typical model combines RNN-based structures with graph convolution to simultaneously capture temporal and spatial features [23], [24]. These models exhibit excellent results in both short-term and long-term predictions. Nevertheless, a primary drawback is their high model complexity. Regarding the trade-off between training time and effectiveness, the choice between using convolutional or recurrent networks has been extensively studied by some scholars, who concluded that the effectiveness of Conv-Net and RNN-type structures increases and decreases with increasing network depth, respectively [25], [26]. Although GNNs demonstrate strong performance in modelling complex spatiotemporal forecasting problems, they often require substantial computational resources, particularly when processing large graphs [27]. This limitation results in high computational costs and significant memory usage, posing challenges for real-time applications involving large-scale traffic networks with numerous nodes and edges.

In summary, while many studies have achieved high accuracy in predicting highway traffic flow, existing models have issues overlooking prediction delay. This occurs when the predicted traffic flow curve exhibits a time shift relative to the ground truth values. Although these predictions may have small errors in conventional metrics, they are essentially inaccurate predictions. On the contrary, a model capable of capturing the complexity of the prediction task and having an adequate amount of high-quality training data will exhibit a low but still possible frequency of such time shifts occurring, especially at peak points of data variations and long-term discontinuity points within the same trend. This is common in time series forecasting because future points can only be predicted based on past observed values. If the past ground truth values have consistently followed a stable trend, the model is more likely to continue following that trend during prediction, missing the discontinuity until it is reflected in the series data. It can be anticipated that during peak periods of traffic flow, where data features are rapidly changing, ensuring the effectiveness of predictions becomes challenging to guarantee.

## III. Methodology

In this section, we explore the prediction of traffic on long-distance highway segments, where the sparsity of electronic toll collection (ETC) data often shows delays in the prediction results. To address this problem, we introduced an incremental state-gate output decomposition (IGOD)

mechanism, integrated within a recurrent neural network (RNN) framework, to enhance both the accuracy and timeliness of our traffic forecasts. The proposed IGOD model has been validated with real-world GNSS probe data for low latency urban traffic predictions [28]. We apply this enhanced model to forecast highway traffic, where IGOD decomposes inputs and outputs into base values and residuals, and allows the residuals to probabilistically determine the direction of value changes. This approach offers the advantage of not providing a fixed predicted output, but instead allowing for a flexible range of amplitude variations, thus improving prediction robustness.

The overall framework of the traffic flow prediction algorithm is depicted in Fig. 1. The process starts with the thorough preprocessing of traffic data obtained from the highway ETC gantry. Then, we segregate the data set into distinct subsets, i.e., the testing set and the training set. The segmented series, denoted as $\{x_1, x_2, \ldots, x_t\}$, serves as the input for our model. Based on the IGOD-RNN model, we employ a harmonious interplay between the encoder and decoder mechanisms to generate precise short-term traffic predictions on the targeted road segment. This procedure involves the application of a two-head self-attention layer and gated recurrent units (GRUs) to ensure reliable outcomes. During the training phase, continuous adjustments are made to enhance the functionality of the model. The feedback from the predictions is analysed and incorporated into the model, enabling gradual improvements in accuracy and performance over time. Specifically, the output of the model consists of predicted traffic flow results, which are used to inform further iterations of the process, ensuring a continuous cycle of refinement and optimisation. To enhance the stability and accuracy of traffic flow prediction, the proposed RNN mechanism can be readily extended to other models and prediction tasks. The following steps outline the specific process of the model.

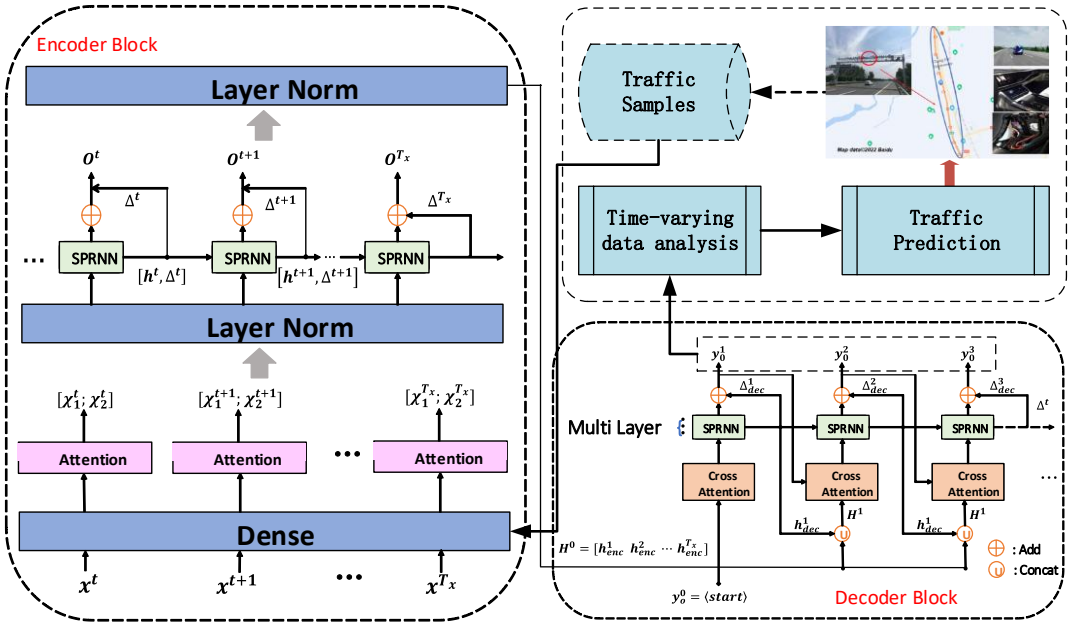


Fig. 1. General framework for traffic flow forecasting. The encoder employs dense layers to refine attention mechanisms for extracting temporal and spatial features, while the decoder utilises cross-attention mechanisms to integrate multilayer spatial and temporal information, enhancing prediction accuracy.

To commence, we initialise a two-head self-attention layer within the encoder. This architectural choice empowers the model to concurrently address two interconnected tasks, as opposed to exclusively concentrating on a single task. The preprocessed highway gantry data are fed into the encoder, and this operation leads to the creation of two distinct delta substates, denoted as $s_1^{(t)}$ and $s_2^{(t)}$. These substates can be formally represented as follows

$$s_1^{(t)}, s_2^{(t)} = MultiHeadAtt\left(X^{(t)} \middle| X^{(1)}, \ldots, X^{(t-1)}, X^{(t+1)}, \ldots, X^{(T)}\right), \quad (1)$$

where $X^{(i)} \in \mathbb{R}^{N \times D}$, $i = 1, 2, \ldots, T$, $s_1^{(t)}, s_2^{(t)} \in \mathbb{R}^{n \times 1}$, and $n$ denotes the number of nodes in each graph. It is worth noting that we do not perform the final multi-head concatenation operation to obtain two substates. Considering that real-world traffic flow tends to evolve gradually in small increments between two sampling points, we introduce the concept of inverse probability to capture this incremental change, which can be expressed as either a positive or negative value with a certain probability. Leveraging the two previously mentioned substates, we compute the inverse probability as follows

$$\bar{p}_{inv}^{(t)} = 1 - p_{inv}^{(t)} = 1 - \sigma\left(f\left(s_2^{(t)}\right)\right), \quad (2)$$

where $p_{inv}$ is defined as the inverse probability, $\bar{p}_{inv}$ represents the probability that the previous direction will be maintained, f (.) is a mapping function to be defined later, and $\sigma$ (.) is the sigmoid function. We further describe the inverse direction as the sign of the difference between these two probabilities.

$$1. \quad \text{direction} = \text{sign}\left(\bar{p} - p\right) = \text{sign}\left(1 - 2p\right) \in \mathbb{R}^{N \times 1}. \quad (3)$$

To quantify the uncertainty increment in traffic flow prediction, we utilise the size of substate $s_1^{(t)}$ as the increment size, and substate $s_2^{(t)}$ as the base value for prediction. This allows us to capture both the magnitude and the direction of the change in traffic flow. The uncertainty increment is calculated as follows

$$\delta^{(t)} = \left| s_2^{(t)} \right| \in \mathbb{R}^{N\times 1}. \quad (4)$$

2. $$\gamma(t) = s_1^{(t)} \in \mathbb{R}^{N\times 1}. \quad (5)$$

Finally, integrating (2)–(5) yields the full expressions for the increment and the predicted output:

$$\Delta^{(t)} = \lambda \times \text{sign}\left(1-2p^{(t)}\right) \times \delta^{(t)},\ \lambda > 0, \quad (6)$$

$$O^{(t)} = \left(W \times h^{(t)} + b\right) + \Delta^{(t)}. \quad (7)$$

where $h^{(t)}$is the hidden state of gated RNN units at time t. W and b denote the parameters of the fully connected layer. Learnable parameter $\lambda$ is the scale factor which is hoped to fine-tune the size of increment. To couple with the decoder, it is necessary to define a long-term cross-attention mechanism to play a connection role, as shown in Table I. As a lightweight variant of RNN, the gated recurrent unit incorporates long short-term memory functionality. Building upon its structure, we propose a novel computing unit and provide the overall expression as follows:

$$\gamma^{(t)} = s_1^{(t)}, \quad (8)$$

$$\Delta^{(t)} = \lambda \times \text{sign}\left(1-2p^{(t)}\right) \times \delta^{(t)}, \quad (9)$$

$$\delta^{(t)} = \left| s_2^{(t)} \right|, \quad (10)$$

$$p_{inv}^{(t)} = \sigma\left( -\varphi + \sum_{i=1}^{t-1} \Delta^{(i)} \right), \quad (11)$$

$$\left[ r^{(t)}; u^{(t)} \right] = \sigma\left( \Theta_r \times \mathcal{G}\left[ \gamma^{(t)}; h^{(t-1)} \right] \right), \quad (12)$$

$$C^{(t)} = Tanh\left( \Theta_c \times \mathcal{G}\left[ \gamma^{(t)}; \left( r^{(t)} \odot h^{(t-1)} \right) \right] \right), \quad (13)$$

$$h^{(t)} = u^{(t)} \odot h^{(t-1)} + \left(1-u^{(t)}\right) \odot C^{(t)}, \quad (14)$$

$$O^{(t)} = \left(W \times h^{(t)} + b\right) + \Delta^{(t)}. \quad (15)$$

where equations (8)-(11) are used to compute the complete uncertainty increment, while equations (12)-(15) are the calculation process in the recurrent unit, where $\mathcal{G}$ denotes graph convolution operator, $[\cdot;\cdot]$ denotes the operation of concatenate, $\odot$ represents element-wise product, $h^{(t)}$ is new hidden state, $O^{(t)}$ is the output prediction (as well as $y^{(t)}$ in the decoding side), $\Theta_r$ and $\Theta_c$ are corresponding graph convolution kernels.

The system employs a training process where the predicted values are compared with the actual values, allowing for the backpropagation of errors to update the network weights iteratively. Once the model achieves predefined criteria for prediction accuracy, the predicted traffic flow values are output. These predicted values are then obtained by applying an inverse normalisation technique.

TABLE I. DECODER CALCULATION STEPS.

| Algorithm 1: Decoding using long term cross attention | |
|---|---|
| 1 | Function<br>Input:<br>$y_o^{(0)}$, prediction length $k$, hidden state sequence from encoder $H = \{h_{enc}^{(t-p)}, h_{enc}^{(t-p+1)} \dots\ h_{enc}^{(t-1)}\}$ |
| 2 | Output: |
| 3 | Concatenate all hidden states $H$ of encoder, noted as $H^{(0)}$; |
| 4 | Calculate the $MultiHeadAtt\left(y_o^{(0)} \middle| H^{(0)}\right)$; # by (1) |
| 5 | Calculate $y_o^{(1)}$、$h_{dec}^1$、$\Delta_{dec}^1$; # by (7), (8) |
| 6 | for $i = 1$, to $k-1$ do:<br>$H^{(i)} = concat\left(h_{dec}^{(i)}, H^{(i-1)}\right)$; |
| 7 | Calculate $MultiHeadAtt\left(y_o^{(i)} \middle| H^{(i)}\right)$; # by (1) |
| 8 | Calculate $y_o^{(i+1)}$、$h_{dec}^{i+1}$、$\Delta_{dec}^{i+1}$; # by (7), (8)<br>break; |
| 9 | end |
| 10 | |

## IV. EXPERIMENT

### A. Experiment and Data

Toll collection data provide exact vehicle counts and classifications at specific points, ensuring high accuracy in traffic volume measurement. The historical records of toll collection offer valuable long-term insights into traffic trends and seasonal patterns, making it ideal for trend analysis. Additionally, consistent and systematic collection of toll data allows monitoring of traffic at strategic highway points, such as entry and exit points, providing comprehensive and reliable data sets. Economic insights derived from toll revenues further enhance the value of these data for financial forecasting and understanding travel behaviour. On the contrary, while GNSS data offers dynamic and real-time information on vehicle location, speed, and travel time, it is often hampered by device accuracy and signal reception issues, and requires significant processing and storage capacity due to the large volumes of data generated. Therefore, despite the widespread use of GNSS data, toll collection data are often considered more in highway scenarios for their accuracy, reliability, and long-term analytical benefits. We therefore turn to toll collection data, which offers greater reliability and precision.

In this section, we utilise data from the ETC gantries along the Changjiu Highway in China, with a particular focus on the segment spanning K602+630 to K625+420 as the primary subject of our analysis. It records information for 45,429 vehicles over a 24-hour period, obtained from two adjacent ETC gantries between September 20–22, in 2021. These adjacent gantries on the highway log entry and exit times for each vehicle passing through this section. Importantly, the vehicle count of the section can be determined by subtracting the number of vehicles exiting from those entering, as these

adjacent gantries on the highway segment can track both the entry and exit counts. The density is then derived by dividing the gantry distance by the count of vehicles. Multiplying the traffic density by the speed of travel yields the traffic volume, reflecting the number of vehicles passing through a specific point in this section.

During the data collection process, addressing the presence of erroneous and missing data is crucial, often resulting from unexpected external factors. Since speed and travel time exhibit a proportional relationship, focussing on vehicle speed alone suffices to identify erroneous data. Erroneous data mainly arise from malfunctioning equipment and abnormal vehicle behaviour. Vehicles passing through the section during nonpeak hours or experiencing significant changes in driving conditions can produce speeds that fall outside the reasonable range for highway travel. To identify erroneous data, this study utilises a threshold judgment method. Using empirical equations, this method establishes a reasonable range for traffic data and flags any data points that fall outside this range as erroneous. The design speed of the Changjiu highway and the characteristics of the data source guide the determination of abnormal data. Specifically, vehicle speeds below 30 km/h are considered anomalous. Therefore, this threshold serves as a criterion to identify erroneous data, allowing its subsequent removal or appropriate treatment. We set the threshold as $30 \leq v \leq v_1 f_v$, where $v$ is the vehicle speed, $v_1$ is the design speed of the highway, which is 120 km/h, and $f_v$ is the correction coefficient, which is usually taken between [1.3, 1.5] based on empirical values. The gaps left after deletion of erroneous data are treated as data loss and filled based on spatiotemporal correlation taking into account the time-varying characteristics and spatial changing characteristics of traffic flow at the same time. In this work, we adopt the data repair method based on adjacent periods using the average data of the first 5 adjacent periods, like $\boldsymbol{x_k} = \frac{x_{k-5}+x_{k-4}+\cdots+x_{k-1}}{5}$ to repair the lost data. This method takes the average value of the data from the preceding N adjacent time periods to better ensure the precision and accuracy of data processing.

Temporal characteristics play a crucial role in understanding traffic patterns on a particular road segment. To investigate these temporal variations, data were collected at 10-minute intervals, providing a detailed overview of the traffic status, as shown in Fig. 2, which offers insight into the temporal patterns and fluctuations in traffic conditions. The traffic status on this road segment shows regularity in its daily variations, indicating that the characteristics of traffic flow, average speed, and average density tend to be similar on different dates. This analysis reveals similarities in traffic distribution patterns while also shedding light on specific factors that may have influenced traffic volume, average speed, and average density on different days. To further verify the temporal correlation of traffic flow, average speed, and average density over the three days, the Pearson correlation coefficient $\rho$ is introduced to indicate the degree of correlation between two variables, and the Pearson coefficients of flow, average velocity, and average density between pairs within three days are shown in Table II. In general, it is considered that when $\rho_{X,Y} \geq 0.8$, the degree of correlation is relatively high.

To delve into temporal variations of traffic flow and understand traffic patterns within a specific road segment, data were collected at 10-minute intervals, providing comprehensive information on temporal patterns and fluctuations in traffic conditions, as depicted in Fig. 2. Traffic status on this road segment demonstrates a consistent daily rhythm, suggesting that attributes of traffic status tend to exhibit similarity across different dates. This analysis reveals commonalities in traffic distribution patterns, while also shedding light on specific factors that may have influenced traffic volume, average speed, and average density on different days. To further validate the temporal correlation between traffic flow, average speed, and density over the course of three days, we use the Pearson correlation coefficient $\rho$ as a measure to quantify the level of correlation between these variables, as shown in Table II, where a correlation coefficient $\rho_{X,Y} \geq 0.8$ equal to or greater than 0.8 is considered indicative of a relatively high degree of correlation between the variables.

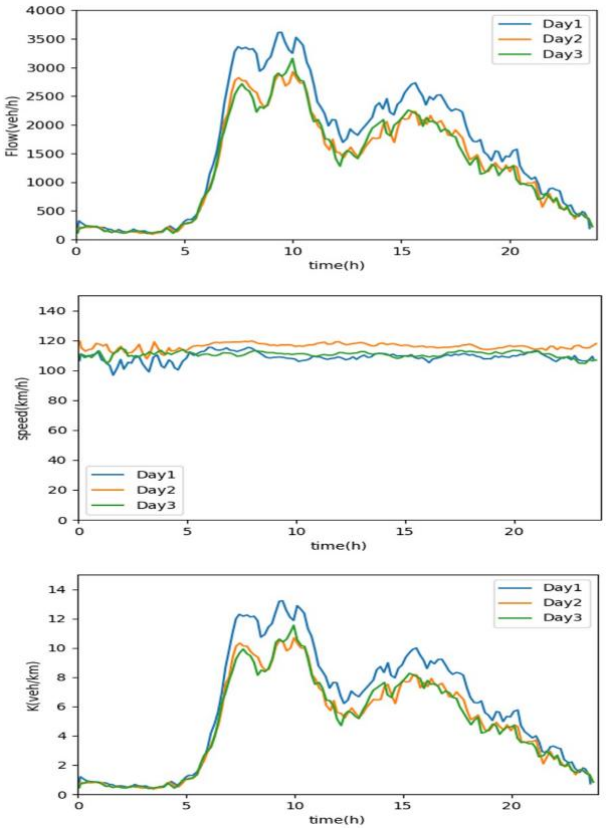


Fig. 2. Traffic data comparison.

Table II reveals a clear correlation between the Pearson coefficients of flow rate and density on the days observed, with coefficients exceeding 0.8. This points to a significant positive linear relationship between flow rate and average density during this period. The high correlation coefficient suggests that as traffic flow increases, the density also tends to increase, signalling a denser concentration of vehicles on the road segment. Although the Pearson coefficients of average velocity for the three days do not exceed 0.8, they still exhibit a moderate positive correlation, with coefficients exceeding 0.5. This implies that there is indeed a certain relationship between average velocity and flow rate or average density, albeit not as robust as the correlation

between flow rate and average density. This positive correlation indicates that as the flow rate or average density increases, the average velocity tends to decrease to some extent. Overall, these correlation coefficients furnish quantitative evidence of the relationships between flow rate, average velocity, and density. The strong correlation between flow rate and density underscores a close connection between these two variables, whereas the moderate correlation between average velocity and the other variables suggests a related but somewhat less pronounced relationship.

TABLE II. TRAFFIC FLOW SIMILARITY ANALYSIS RESULTS.

| **Traffic flow distribution Pearson coefficient** | | | |
|---|---|---|---|
| | Day1 | Day2 | Day3 |
| Day1 | | 0.932 | 0.901 |
| Day2 | 0.932 | | 0.687 |
| Day3 | 0.901 | 0.687 | |
| **Traffic speed distribution Pearson coefficient** | | | |
| | Day1 | Day2 | Day3 |
| Day1 | | 0.599 | 0.527 |
| Day2 | 0.599 | | 0.763 |
| Day3 | 0.527 | 0.763 | |
| **Traffic density distribution Pearson coefficient** | | | |
| | Day1 | Day2 | Day3 |
| Day1 | | 0.996 | 0.991 |
| Day2 | 0.996 | | 0.993 |
| Day3 | 0.991 | 0.993 | |

### *B. Result Analysis*

The experiments were conducted using PyTorch, with all models trained on a single NVIDIA GeForce GTX 1660 SUPER with 6 GB of memory. Initially, we set the learning rate to 0.001 using the Adam optimiser and subsequently reduce it by a factor of 0.5 after 20, 30, 50, 70, and 90 epochs, respectively. The maximum number of epochs is set at 100, with the first five epochs used for warm-up. We employ a batch size of 64 and apply an early stopping strategy. Regarding network parameters, the input linear layer has 12 neurons, and the output linear layer has one neuron. The hidden neurons in the GRU cells number 64, while the attention mechanism employs 32 neurons on the encoding side and 12 neurons on the decoding side. Both the encoder and decoder consist of two identical layers. We determine the threshold for the cumulative increment value as 5.0 based on the performance evaluation in the validation set. Additionally, we incorporate dropout with a rate of $p = 0.1$ in both the input linear layer and the attention layer to mitigate overfitting. We evaluated model performance based on mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). To visualise the experimental results, we created a comparison plot displaying the predicted outcomes, which effectively illustrates the performance of the proposed model in comparison to the ARIMA and LSTM models.

Upon a thorough comparative analysis of both Fig. 3 and Table III, it becomes apparent that both the LSTM and the incremental gate state output decomposition (IGOD) model surpass the ARIMA model in terms of predictive accuracy. This is evident from the lower error metrics. These findings underscore the superiority of deep learning methods over traditional modelling approaches in the realm of time series forecasting. Remarkably, the IGOD model demonstrates a remarkably strong alignment between predicted values and actual observations, with significantly reduced RMSE and MAE values when compared with both the LSTM and ARIMA models. Furthermore, the MAPE metric is 5–10 times smaller compared to the other two forecasting methods. These results provide compelling evidence that our proposed models outperform all other methods and consistently deliver exceptional predictive accuracy. In our subsequent investigation, we delve into the influence of the threshold amplitude on the effectiveness of the model. We performed experiments to obtain MAE values for a prediction horizon of 15 minutes at various threshold values, specifically 0.0, 5.0, 7.5, 10.0, 15.0, and 20.0.

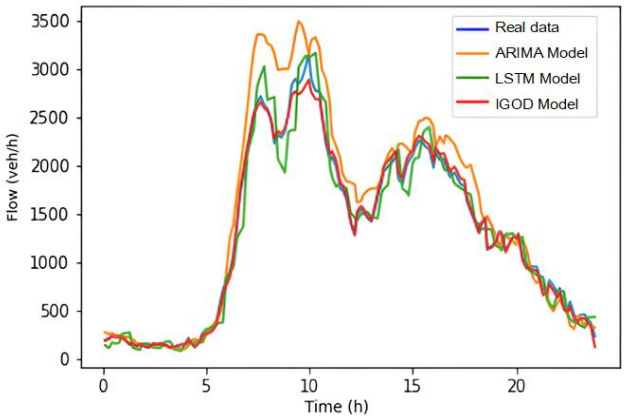


Fig. 3. Traffic flow prediction using RNN models of autoregressive integrated moving average (ARIMA), long short-term memory (LSTM), and incremental gate state output decomposition (IGOD) compared to real data. Time is plotted on the x-axis (hours) and vehicle flow rate on the y-axis (veh/h). The ARIMA model (orange) shows exaggerated peaks during peak traffic times (approximately 7–9 AM and 4–6 PM), while the LSTM (green) and IGOD (red) models more accurately track the actual flow (blue), especially noticeable during rush hours.

Figure 4 reveals that the prediction performance remains relatively stable even when the threshold size undergoes changes.

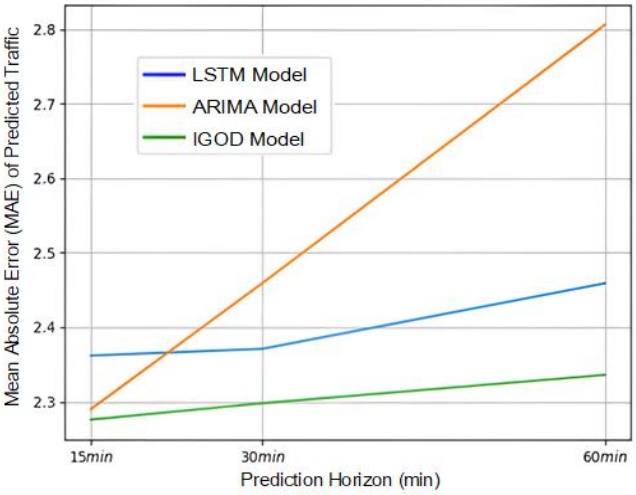


Fig. 4. MAE of traffic flow predictions for the ARIMA model (orange), LSTM model (blue), and the IGOD model (green) at prediction horizons of 15, 30, and 60 minutes. The IGOD model consistently achieves the lowest MAE across all time horizons. On the contrary, the prediction error of the ARIMA model increases significantly as the prediction horizon lengthens, while the LSTM model performs moderately, with a gradual increase in error.

This observation underscores the robustness of the

proposed IGOD model. It should be noted that among the six trials conducted, the delay phenomenon is most pronounced when the threshold value is set to 0.0. This is attributed to the fact that at this specific threshold value, the effect of the increments resembles random compensation, making precise control a challenging endeavour.

TABLE III. PREDICTION PERFORMANCE EVALUATION INDEX COMPARISON.

| | Evaluation index | Traffic forecast |
|---|---|---|
| **ARIMA** | RMSE | 450.215 |
| | MAE | 307.563 |
| | MAPE | 33.868 % |
| **LSTM** | RMSE | 168.656 |
| | MAE | 124.109 |
| | MAPE | 14.947 % |
| **Incremental Gate State Decomposition module** | RMSE | 50.9964 |
| | MAE | 3.3643 |
| | MAPE | 4.062 % |

## V. CONCLUSIONS

The accurate prediction of highway traffic flow is essential for effective highway management, congestion mitigation, and an enhanced travel experience. The objective of this study is to provide reliable predictions capable of facilitating traffic diversion, accident prevention, enhancing road safety, and optimising overall transportation efficiency. To realise this objective, we propose an incremental gate state output decomposition model, which accounts for the spatiotemporal complexity of road networks by analysing time series data through segmenting time intervals and aggregating traffic flow information for prediction. We validated the model using data from the K602+630 to K625+420 segment of the Changjiu Highway. The results of the experiment conclusively demonstrate that the applied IGOD model surpasses other benchmark models in terms of prediction accuracy, with the threshold amplitude having no significant effect on performance.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

## REFERENCES


[1] F. Zou *et al.*, "Expressway speed prediction based on electronic toll collection data", *Electronics*, vol. 11, no. 10, p. 1613, 2022. DOI: 10.3390/electronics11101613.

[2] Y. Zhao, W. Lu, Y. Rui, and B. Ran, "Classification of the traffic status subcategory with ETC gantry data: An improved support tensor machine approach", *Journal of Advanced Transportation*, vol. 2023, art. ID 765937, pp. 1–21, 2023. DOI: 10.1155/2023/2765937.

[3] F. Guo *et al.*, "Positioning method of expressway ETC gantry by multi-source traffic data", *IET Intelligent Transport Systems*, vol. 18, no. 3, pp. 540–554, 2024. DOI: 10.1049/itr2.12280.

[4] R. Zhang *et al*, "Intelligent path planning by an improved RRT algorithm with dual grid map", *Alexandria Engineering Journal*, vol. 88, pp. 91–104, 2024. DOI: 10.1016/j.aej.2023.12.044.

[5] G. Cantisani, G. Del Serrone, R. Mauro, P. Peluso, and A. Pompigna, "Traffic stream analysis by radar sensors on two-lane roads for free-moving and constrained vehicles identification", *Sensors*, vol. 23, no. 15, p. 6922, 2023. DOI: 10.3390/s23156922.

[6] W. Ma and S. Qian, "High-resolution traffic sensing with probe autonomous vehicles: A data-driven approach", *Sensors*, vol. 21, no. 2, p. 464, 2021. DOI: 10.3390/s21020464.

[7] F. J. Braz *et al.*, "Road traffic forecast based on meteorological information through deep learning methods", *Sensors*, vol. 22, no. 12, p. 4485, 2022. DOI: 10.3390/s22124485.

[8] D. Pavlyuk, "Feature selection and extraction in spatiotemporal traffic forecasting: A systematic literature review", *European Transport Research Review*, vol. 11, p. 1–19, 2019. DOI: 10.1186/s12544-019-0345-9.

[9] P. Duan, G. Mao, W. Liang, and D. Zhang, "A unified spatio-temporal model for short-term traffic flow prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3212–3223, 2019. DOI: 10.1109/TITS.2018.2873137.

[10] H. Zhang, X. Wang, J. Cao, M. Tang, and Y. Guo, "A hybrid short-term traffic flow forecasting model based on time series multifractal characteristics", *Applied Intelligence*, vol. 48, pp. 2429–2440, 2018. DOI: 10.1007/s10489-017-1095-9.

[11] S. Shahriari, M. Ghasri, S. A. Sisson, and T. Rashidi, "Ensemble of ARIMA: Combining parametric and bootstrapping technique for traffic flow prediction", *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 1552–1573, 2020. DOI: 10.1080/23249935.2020.1764662.

[12] X. Feng, X. Ling, H. Zheng, Z. Chen, and Y. Xu, "Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction", *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2001–2013, 2019. DOI: 10.1109/TITS.2018.2854913.

[13] R. B. Sharmila, N. R. Velaga, and A. Kumar, "SVM-based hybrid approach for corridor-level travel-time estimation", *IET Intelligent Transport Systems*, vol. 13, no. 9, pp. 1429–1439, 2019. DOI: 10.1049/iet-its.2018.5069.

[14] S. O. Mousavizadeh Kashi and M. Akbarzadeh, "A framework for short-term traffic flow forecasting using the combination of wavelet transformation and artificial neural networks", *Journal of Intelligent Transportation Systems*, vol. 23, no. 1, pp. 60–71, 2019. DOI: 10.1080/15472450.2018.1493929.

[15] W. Fang, W. Zhuo, J. Yan, Y. Song, D. Jiang, and T. Zhou, "Attention meets long short-term memory: A deep learning network for traffic flow forecasting", *Physica A: Statistical Mechanics and its Applications*, vol. 587, art. 126485, 2022. DOI: 10.1016/j.physa.2021.126485.

[16] R. Shi and L. Du, "Multi-section traffic flow prediction based on MLR-LSTM neural network", *Sensors*, vol. 22, no. 19, p. 7517, 2022. DOI: 10.3390/s22197517.

[17] Z. Islam, M. Abdel-Aty, and N. Mahmoud, "Using CNN-LSTM to predict signal phasing and timing aided by High-Resolution detector data", *Transportation Research Part C: Emerging Technologies*, vol. 141, art. 103742, 2022. DOI: 10.1016/j.trc.2022.103742.

[18] F. Shen *et al.*, "Long-term multivariate time series forecasting in data centers based on multi-factor separation evolutionary spatial-temporal graph neural networks", *Knowledge-Based Systems*, vol. 280, p. 110997, 2023. DOI: 10.1016/j.knosys.2023.110997.

[19] S. F. Ahmed *et al.*, "Enhancement of traffic forecasting through graph neural network-based information fusion techniques", *Information Fusion*, vol. 110, art. 102466, 2024. DOI: 10.1016/j.inffus.2024.102466.

[20] T. Mallick, P. Balaprakash, E. Rask, and J. Macfarlane, "Graph-partitioning-based diffusion convolutional recurrent neural network for large-scale traffic forecasting", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2674, no. 9, pp. 473–488, 2020. DOI: 10.1177/0361198120930010.

[21] J. J. Q. Yu and J. Gu, "Real-time traffic speed estimation with graph convolutional generative autoencoder", *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3940–3951, 2019. DOI: 10.1109/TITS.2019.2910560.

[22] Y. Wang, Q. Ren, and J. Li, "Spatial–temporal multi-feature fusion network for long short-term traffic prediction", *Expert Systems with Applications*, vol. 224, art. 119959, 2023. DOI: 10.1016/j.eswa.2023.119959.

[23] H. Zhu *et al.*, "A novel traffic flow forecasting method based on RNN-GCN and BRB", *Journal of Advanced Transportation*, vol. 2020, art. ID 7586154, pp. 1–11, 2020. DOI: 10.1155/2020/7586154.

[24] K. Yu, X. Qin, Z. Jia, Y. Du, and M. Lin, "Cross-attention fusion based spatial-temporal multi-graph convolutional network for traffic flow prediction", *Sensors*, vol. 21, no. 24, p. 8468, 2021. DOI: 10.3390/s21248468.

[25] S. Reza, M. C. Ferreira, J. J. M. Machado, and J. M. R. Tavares, "A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks", *Expert Systems with Applications*, vol. 202, art. 117275, 2022. DOI: 10.1016/j.eswa.2022.117275.

[26] G. Huo, Y. Zhang, B. Wang, J. Gao, Y. Hu, and B. Yin, "Hierarchical spatio–temporal graph convolutional networks and transformer network for traffic flow forecasting", *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3855–3867, 2023. DOI: 10.1109/TITS.2023.3234512.

[27] A. Cini, I. Marisca, F. M. Bianchi, and C. Alippi, "Scalable spatiotemporal graph neural networks", in *Proc. of the AAAI Conference on Artificial Intelligence*, 2023, pp. 7218–7226. DOI: 10.1609/aaai.v37i6.25880.

[28] Y. Lu, X. Meng, L. Peng, S. Xu, and E. Chen, "IODRNN - Incremental output decomposition for a valid traffic flow prediction with GNSS data", *Engineering Applications of Artificial Intelligence*, vol. 128, art. 107520, 2024. DOI: 10.1016/j.engappai.2023.107520.