Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

Integrating vehicle dimension features for vision-based traffic density prediction using YOLOv5-LSTM architecture

Filda Angellia[1], Nita Merlina[2], Agus Subekti[3], Rahmadya Trias Handayanto[4]
[1,2,3,4]Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia

Article Info

Article history:

Received May 17, 2026
Revised June 06, 2026
Accepted June 06, 2026

Keywords:

Traffic density prediction
Vehicle dimension features
YOLO v5
LSTM

ABSTRACT

Traffic congestion in urban areas requires intelligent technology-based solutions to support modern transportation systems. This study proposes a vision-based traffic congestion prediction framework that integrates YOLOv5 with a sequential deep learning model to improve forecasting accuracy. YOLOv5 is used for real-time vehicle detection, while the width and height of the bounding box are extracted as spatial occupancy features to provide additional information beyond conventional vehicle counting methods. Experiments are conducted using six urban traffic videos consisting of 90,012 frames collected under various traffic conditions. The extracted features are converted into sequential temporal records and subsequently used to train Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. Model performance is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Experimental results show that both models achieve competitive performance for traffic congestion forecasting. LSTM achieved the best performance with an MSE of 3.77, an RMSE of 1.94, and an MAE of 1.47, demonstrating its superior ability to capture long-term temporal dependencies in large-scale sequential traffic data. In contrast, GRU exhibited lower computational complexity and faster inference time due to its simpler architecture. These findings suggest that integrating vehicle dimensional features with sequential deep learning models provides a more effective approach to artificial intelligence.

Corresponding Author:

Filda Angellia,
Faculty of Information Technology,
Universitas Nusa Mandiri,
Margonda No. 545, Pondok Cina, Beji, Depok, Jawa Barat 16424, Indonesia.
Email: fildaibik57@gmail.com
https://doi.org/10.52465/joscex.v7i2.119

1. INTRODUCTION

One such major development challenge that is faced by urban planners and policy-makers in almost all developing countries today, even with the advancement in technology and transport systems to cope up with rapid population growths bringing about tool of from fuel consumption productivity efficiency mobility etc. Rising vehicle numbers with not large enough infrastructures pushed the growing in Intelligent Transportation System (ITS) to do monitoring and forecast traffic state at real time [1]. However, recent years have also witnessed significant advances in AI and Computer Vision technologies that enable automatic vehicle detection, traffic monitoring & predictive analysis for news ways of supporting adaptive traffic management [2][3].

Advances in deep learning have substantially improved the performance of vehicle detection systems based on digital imagery. Among various object detection approaches, YOLOv5 has become one of the most widely used models due to its high detection accuracy and real-time processing capability [4], [5], [6] successfully implemented YOLO-based approaches for vehicle detection and traffic monitoring in complex urban environments. Similarly, Hammoud et al. in 2024 demonstrated that YOLO architectures are effective for real-time traffic surveillance under varying lighting and environmental conditions [7]. However, most existing studies primarily focus on instant detection from single frames without fully utilizing sequential historical visual information for traffic density prediction.

In addition to object detection, sequential deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have demonstrated strong capability in modeling temporal dependencies from sequential data. Research by Garg et al. in 2025 showed that LSTM-based architectures perform effectively in learning temporal visual patterns from sequential imagery [8], while Kim et al. in 2025 reported that recurrent deep learning models are capable of improving prediction stability in dynamic traffic-related environments [9]. According to Pyo; in 2025 GRU provides computational performance with fewer parameters while achieving similar predictive accuracy when compared to recurrent models [10]. The results suggest that both LSTM and GRU are appropriate for the traffic density prediction tasks in which temporal dependency of traffic is necessary.

Unlike conventional traffic prediction methods that rely solely on vehicle counts, bounding-box width and height provide additional spatial information regarding road occupancy and congestion levels. Larger bounding boxes generally indicate either physically larger vehicles or vehicles located closer to the camera, both of which contribute differently to traffic flow dynamics. Previous studies in intelligent transportation systems have suggested that spatial occupancy features extracted from computer vision models can improve traffic state estimation by capturing information beyond simple object counts [11], [12], [13]. In fact, vehicle dimensions, including length and width, may provide additional information related to road occupancy and traffic density patterns. Moreover, limited studies have integrated vehicle dimension features extracted from real-time visual detection into sequential prediction models such as LSTM and GRU [14], [15]. Vehicle orientation and lane occupancy ratio often require additional geometric calibration and lane segmentation procedures, increasing computational complexity. In contrast, bounding-box width and height are directly generated by YOLOv5 without additional processing, making them computationally efficient for real-time deployment.

Therefore, this study proposes a vision-based traffic density prediction framework integrating vehicle dimension features extracted using YOLOv5 with sequential deep learning architectures. The proposed framework detects vehicles in real time and extracts dimensional information from each detected object before transforming the results into sequential temporal data for traffic density prediction. The study evaluates and compares the performance of LSTM and GRU models using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The key contribution of this paper is the introduction of two vehicle dimension features and their embedding into a sequential traffic prediction framework that provides significant improvements in predicting accuracy computational effectiveness for Intelligent Transportation Systems (ITS).

2. METHOD

In this study, we present a vision-based traffic density prediction framework which enables vehicle dimension features extracted from traffic video imagery to be integrated with data-driven sequential deep learning models. It is an idea that unifies computer vision and temporal learning techniques to forecast traffic density patterns using past vehicle movements. The detailed research workflow is composed of several stages including data acquisition, preprocessing, vehicle detection, feature extraction of sequential data construction (traffic density prediction), training process and model evaluation. Figure 1 shows the entire research workflow.

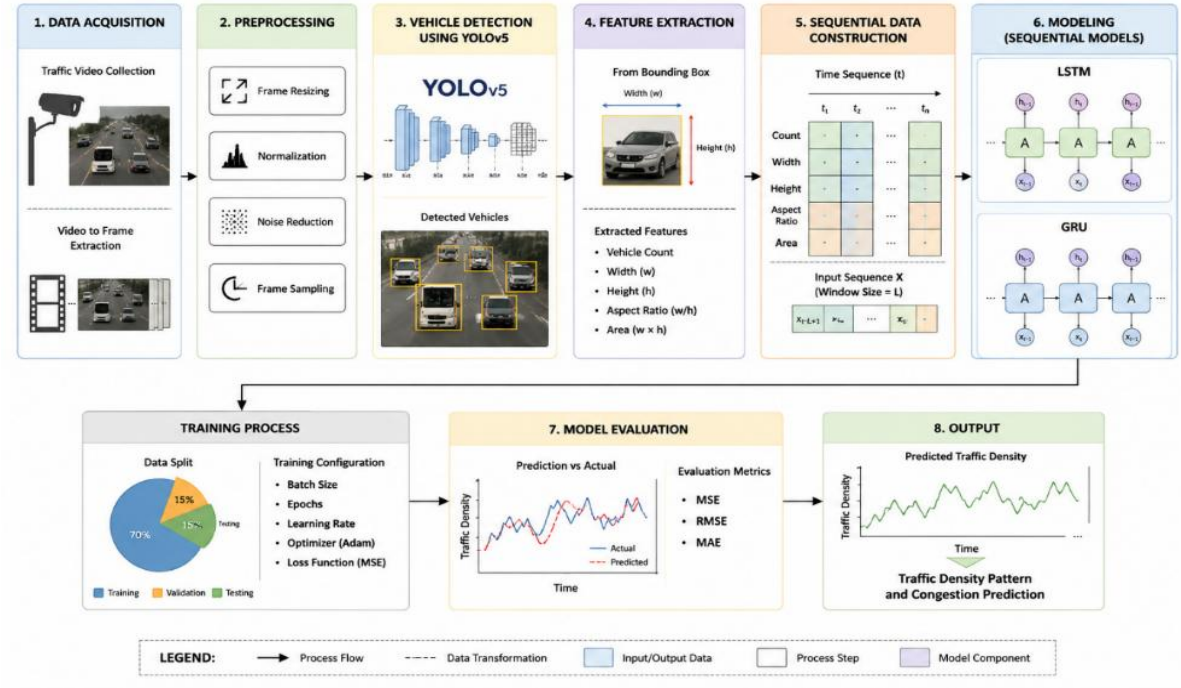


Figure 1. Research design of the proposed traffic density prediction system

2.1. Data Acquisition

The dataset used in this study was self-collected from urban traffic surveillance systems using CCTV cameras installed at major roads and intersections. The recorded videos represent diverse traffic conditions, including normal traffic flow, moderate congestion, and high-density traffic situations, thereby providing a comprehensive representation of urban traffic behavior [16]. The experimental dataset consists of 12 traffic videos, namely video_bagian_part1.mp4 to video_bagian_part12.mp4, with a total duration of 7,202.87 seconds. The complete dataset contains 162,060 image frames with an average frame rate of 22.5 frames per second (fps), while each video contains approximately 13,505 frames. The recorded traffic scenes include an average of 6.01 detected vehicles per frame, reflecting varying levels of road occupancy under real urban traffic conditions. The collected videos were subsequently converted into image frames to facilitate vehicle detection and sequential temporal analysis. During preprocessing, the extracted frames were organized chronologically and standardized before being processed by the YOLOv5 model. The use of multiple traffic scenarios with varying traffic densities enables a comprehensive evaluation of the robustness and adaptability of the proposed traffic density prediction framework [17].

2.2. Image Processing

After data collection, preprocessing techniques were applied to improve image quality and standardize the input data before the detection process. The preprocessing stage included frame resizing, normalization, frame extraction, and noise reduction [18].

Frame resizing was performed to ensure that all input images had consistent dimension compatible with the YOLOv5 model. Normalization was applied to stabilize pixel intensity values and improve model convergence during training. In addition, noise reduction techniques were used to minimize unwanted visual disturbances that could affect the vehicle detection performance. The preprocessing stage also ensured that the generated image frames were temporally organized and suitable for sequential learning processes.

2.3. Vehicle Detection Using YOLOv5

Vehicle detection was performed using the YOLOv5 architecture due to its high detection accuracy and real-time processing capability [19]. YOLOv5 is a single-stage object detection model capable of simultaneously performing object localization and classification in a single forward propagation process.

The model detects vehicle objects from each traffic frame and generates bounding box coordinates for every detected vehicle. The generated bounding boxes are subsequently used to extract vehicle-related spatial information [20].

The bounding box representation can be formulated as Eq. (1).

$$B = (x, y, w, h) \tag{1}$$

Where x *and* y denote the center coordinates of the detected object, w represents the width of the bounding box, and h represents the height of the bounding box.
YOLOv5 enables efficient multi-object vehicle detection under varying environmental and traffic conditions.

2.4. Vehicle Feature Extraction

After the vehicle detection stage, important traffic-related features were extracted from the generated bounding box information. The extracted features included vehicle count, vehicle width, vehicle height.
Vehicle count represents the number of detected vehicles within a specific observation interval, while vehicle width and height represent spatial occupancy characteristics of vehicles on the road.
This allows some additional spatial information against traditional vehicle counting approaches, and thus improves traffic density prediction performance.

2.5. Sequential Data Construction

The vehicle information extracted was turn into sequential temporal datasets which has been arranged according to the ordering of timestamps. Each sequence consists of vehicle count and vehicle dimension features, collected in consecutive observation intervals.

Sequential input windows were constructed using a sliding window technique for use as predictors in deep learning prediction models. This involved learning from visual traffic data sequences that weakly supervised the system on historical traffic patterns and temporal dependencies. Different models for LSTM and GRU prediction can be trained with the generated sequential datasets as input.

2.6. Traffic Density Prediction Models

Long Short-Term Memory (LSTM) a type of recurrent neural network architecture uses long-term dependencies modeling for the sequence data [21]. LSTM is equipped with memory cells and different gating mechanisms to retain the temporally important information while escaping the vanishing gradient issue of regular RNNs. The hidden state of the LSTM model is defined in Eq. (2).

$$h_t = O_t \,.\tanh(C_t) \tag{2}$$

Here h_t refers to hidden state, O_t is the output gate, and C_t is the memory cell state. LSTM is applied in this study to investigate the analysis of historic vehicle movement sequences and predict the future traffic density conditions.

Gated Recurrent Unit (GRU) is a type of recurrent neural networks, with simpler structure than LSTM and designed to model sequential data [22]. There are update and reset gates for controlling the information flow while still learning over time. We also define the hidden state of GRU as Eq. (3).

$$h_t = (1 - z_t) \odot h_t + z_t \odot \bar{h_t} \tag{3}$$

Where h_t denotes the hidden state, z_t is the update gate, $\bar{h_t}$ denotes the candidate hidden state. GRU was implemented to evaluate the effectiveness of lightweight sequential architectures for traffic density prediction tasks.

2.7. Training Process

At the training stage, the sequential dataset was divided into training, validation, and testing subsets using a ratio of 70:15:15. This partitioning strategy ensures that sufficient data are available for model training while maintaining independent subsets for hyperparameter tuning and unbiased performance evaluation. To provide

a fair comparison, both LSTM and GRU models were trained using identical experimental settings. The Hyperparameter configuration of the proposed models can be shown in Table 1.

Table 1. Hyperparameter configuration of the proposed LSTM and GRU models

Parameter	LSTM	GRU
Recurrent Layers	1	1
Hidden Units	64	64
Dropout Rate	0.20	0.20
Sequence Length	15	15
Batch Size	16	16
Number of Epochs	200	200
Optimizer	Adam	Adam
Learning Rate	0.001	0.001
Loss Function	Mean Squared Error (MSE)	Mean Squared Error (MSE)
Dataset Split	70:15:15	70:15:15

The sequential samples were generated using a sliding window approach with a sequence length of 15 observations. The Adam optimizer was employed to minimize the Mean Squared Error (MSE) loss function during the training process. Both recurrent neural network architectures were implemented using the same hyperparameter configuration, allowing the observed performance differences to be primarily attributed to their intrinsic architectural characteristics rather than differences in model settings.

2.8. Model Evaluation

Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [23]. were used to assess the performance of the predicted models. These metrics are common for regression-based forecasting problems to quantify the accuracy of each prediction.

The MSE metric is calculated using Eq. (4):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

The RMSE metric is calculated using Eq. (5):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

The MAE metric is calculated using Eq. (6):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

In the formula, y_i is the actual traffic density value, and $\hat{y}_i$ is the predicted traffic density value, n is a total number of observations. Lower values of MSE, RMSE, and MAE indicate better prediction performance. To identify the most effective sequential deep learning architecture for traffic density prediction, a comparative evaluation between the LSTM and GRU models was conducted.

3. RESULTS AND DISCUSSIONS

3.1. Vehicle Detection Results Using YOLOv5

We evaluated the proposed framework on large scale traffic video datasets generated with urban surveillance systems. Within this investigation, both the two experimental scenario was chosen as the main assessment scenario, because it has shown stable temporal learning efficiency in sequential traffic density prediction. This dataset contains six traffic videos of about 90,012 frames in total and the observation is approximate to 3600 seconds. The videos consisted of different traffic situations with an average of 7.36 vehicles detected per frame.

The real-time vehicle detection model YOLOv5 then was trained on traffic imagery and used to get raw vehicle dimension features. Under various traffic density conditions, the detection process successfully detected vehicle objects such as cars, motorcycles, buses, and trucks. The model also produced bounding box information that was used to extract vehicle width and height features for sequential traffic analysis apart from object localization.

Vehicle detection results using YOLOv5 on test data for the fourth experimental scenario are shown in Figure 2. The detection outputs show that the model was able to detect several vehicle objects simultaneously under normal and heavy urban traffic conditions.

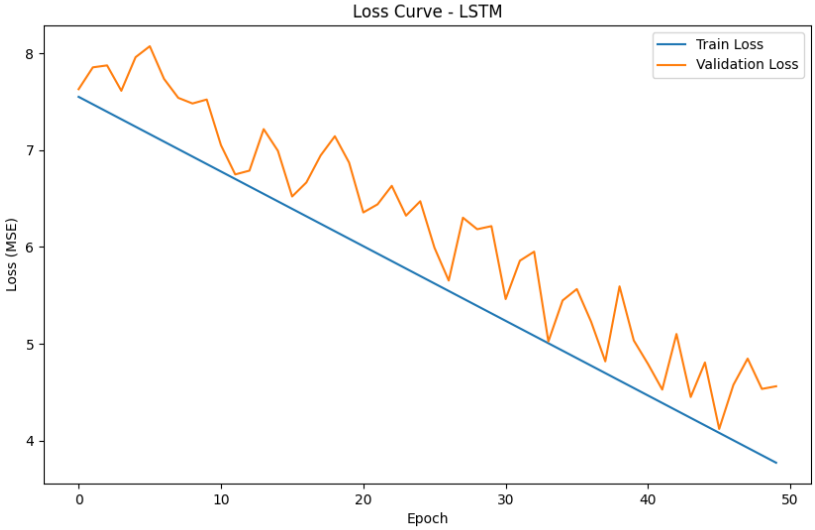
Figure 2. The result of vehicle detection using YOLOv5

The converted detection outputs include vehicle number and dimension information which later changed to serial temporal datasets for traffic density prediction. The results demonstrate that YOLOv5 held stable detection under different vehicular scenarios and successfully created spatial traffic features essential for sequential deep learning analysis.

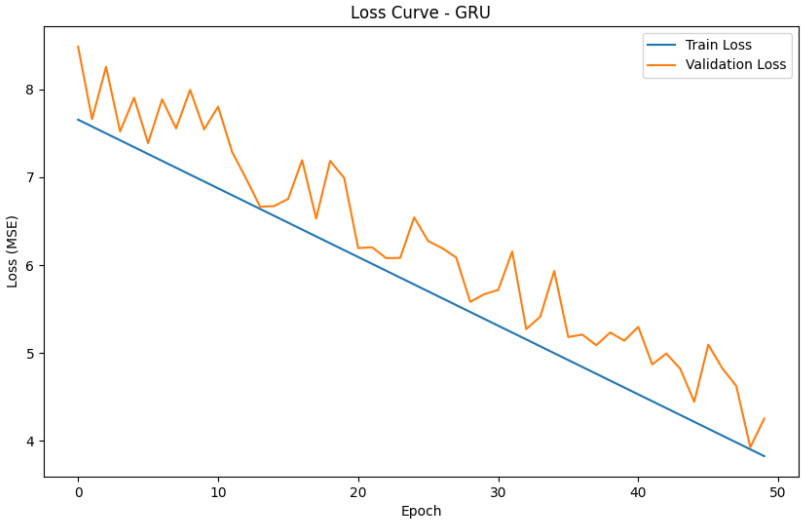
3.2. Sequential Traffic Density Prediction Results

Then as the output of these stages which is sequential datasets generated were fed to LSTM and GRU models in order to predict future traffic density patterns. The prediction performance was assessed based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and prediction time.

Fig. 3: The training loss curves for LSTM (a) and GRU (b). Both models converged stably during training, showing that the sequential architectures learned temporal traffic patterns from historical visual traffic data well.



(a)



(b)

Figure 3. The result of loss curve graph (a) LSTM (b) GRU

Actual and predicted traffic density values generated by the LSTM and the GRU models are compared in Figures 4 and 5 respectively. The prediction curves demonstrate that these two models could successfully capture temporal traffic density progression using sequential traffic data.

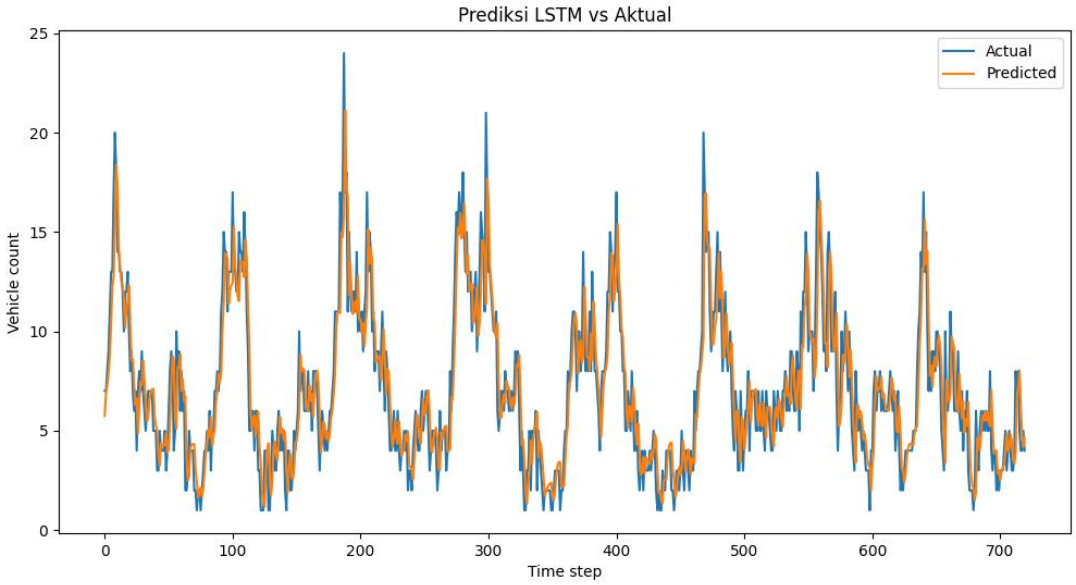


Figure 4. Traffic density prediction result using LSTM

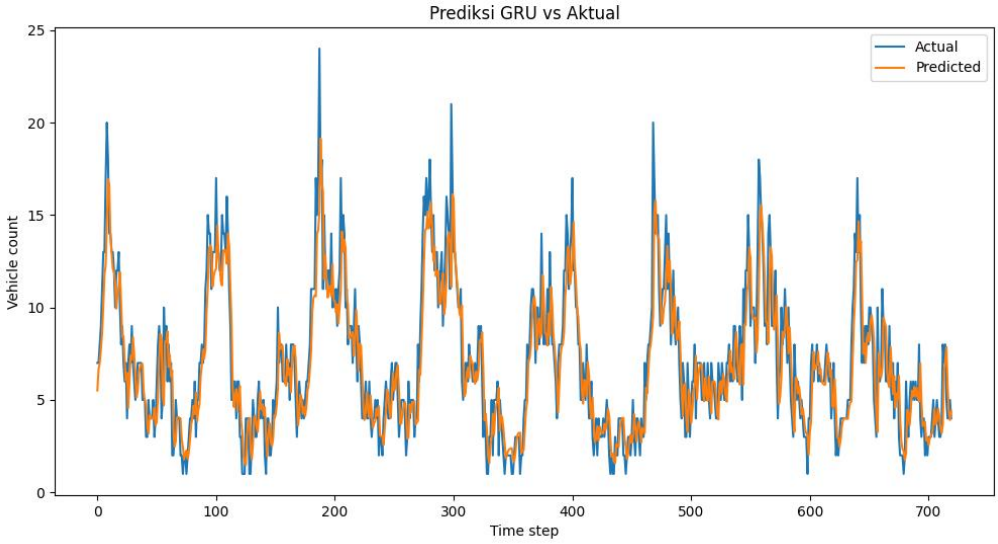


Figure 5. Traffic density prediction result using GRU

Furthermore, the Figure 6 presents the comparative prediction performance between the LSTM and GRU models.

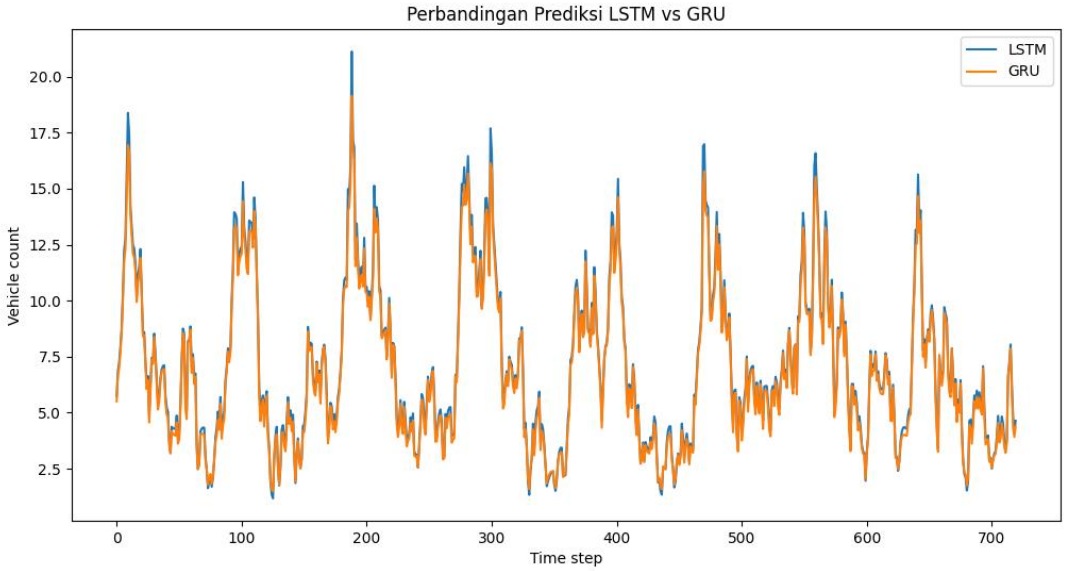


Figure 6. Comparative Prediction Performance Between LSTM and GRU

A summary of the quantitative evaluation results is provided in Table1.

Table 2. The Performance of LSTM and GRU Models Compared in All Experimental Scenarios

Experiment	Model	MSE	RMSE	MAE	Prediction Time (s)
Test 1	LSTM	4.07	2.01	1.54	0.32
	GRU	4.50	2.12	1.70	**0.26**
Test 2	LSTM	**3.77**	**1.94**	**1.47**	0.33
	GRU	3.82	1.95	**1.47**	**0.25**

The experimental results show that both LSTM and GRU have achieved competitive prediction performance on large-scale traffic datasets compared with other models. Nonetheless, in both experimental scenarios the use of LSTM resulted in lower prediction error. Test 1: LSTM MSE value is 4.07; RMSE is 2.01; MAE is 1.54 GRU performed worse than LSTM for all prediction accuracy metrics. In a similar case, Test 2 again proved LSTM prediction performance with an MSE value of about 3.77 and for RMSE of this test it was 1.94, better than GRU but very closely.

In the end, LSTM performs a little bit better in prediction accuracy and GRU outperform in prediction time. While the difference in performance between these two models was small, both model architectures are viable depending on application needs. LSTM is more suited in use cases where one can afford maximum training times seeking accurate predictions with temporal stability, while GRU is faster prediction with low computational complexity trade-off. In terms of quality of prediction, both the models provide relatively low-value error for the scale of traffic dataset used in this work. Nevertheless, significant improvements can still be made especially for the real-time traffic monitoring applications that require reducing RMSE value to as close to 1.5 or less in order to provide more accurate key performance indicators for operational analysis and decision making.

3.3. Discussion

According to the experimental results, a traffic density prediction method is proposed that combines YOLOv5-based vehicle detection with sequential deep learning architectures through the global average pooling layer. The obtained vehicle dimension features (i.e., vehicle width and height) supplied more spatial information, which made the traffic occupancy condition representation more precise than the traditional vehicle-count-based methods. It should also be recognized that the YOLOv5 model proposed in this paper can still maintain a good performance at different traffic conditions, and it can produce the bounding box information needed for

vehicle feature extraction. These synthetic sequential datasets allowed LSTM and GRU models to learn temporal traffic patterns through visual traffic data accumulated through past time intervals.

To further assess the effectiveness of the proposed framework, its performance was compared with several representative studies in the field of vision-based traffic density estimation and prediction. The comparison is presented in Table 3.

Table 3. Comparison of the proposed study with existing studies

Study	Method	Performance
Hu et al. (2022) [24]	CNN + Transformer (Weakly Supervised)	MSE = 5.80 MAE = 3.90
Hu et al. (2023) [25]	YOLOv5 + Camera Calibration	MAE = 7.03
Arif et al. (2025) [26]	YOLOv8 + LSTM	MSE = 7.22 MAE = 1.8 RMSE = 2.28
Proposed study	YOLOv5-LSTM	MSE = 3.77 RMSE = 1.94 MAE = 1.47

The proposed YOLOv5-LSTM framework achieved lower prediction errors than several existing approaches reported in the literature. Compared with the weakly supervised CNN-Transformer model proposed by Hu et al. (2022) [24], the proposed method reduced the MSE from 5.80 to 3.77 and the MAE from 3.90 to 1.47. Likewise, the proposed framework outperformed the YOLOv8-LSTM approach developed by Arif et al. (2025) [26], which reported an MSE of 7.22, an RMSE of 2.28, and an MAE of 1.80.

The improved performance can be attributed to the integration of vehicle dimension features with sequential deep learning models. Unlike conventional approaches that primarily rely on vehicle counts or camera calibration procedures, the proposed framework incorporates bounding-box width and height extracted directly from YOLOv5 as spatial occupancy features. These additional features provide richer information about road utilization while avoiding the computational overhead associated with geometric calibration methods. Furthermore, the LSTM architecture effectively models long-term temporal dependencies, resulting in more accurate traffic density prediction.

It should be noted that the studies presented in Table 3 employ different datasets and experimental settings. Therefore, the comparison serves as a qualitative benchmark to position the proposed framework with respect to existing state-of-the-art methods. Nevertheless, the consistently lower prediction errors demonstrate that the integration of vehicle dimension features and sequential deep learning provides a promising approach for vision-based traffic density prediction.

4. CONCLUSION

In this paper, we present a framework for vision-based traffic density prediction that combines YOLOv5-based vehicle detection with a sequential deep learning architecture. In this architecture, YOLOv5 is used to perform real-time vehicle detection and then extracts several vehicle dimensional features including vehicle width and height from a traffic video dataset. The collected data is converted into a temporal data sequence, and LSTM and GRU models are used to predict traffic density. Experimental results show that both models can effectively learn temporal traffic patterns. Predictive performance shows LSTM results outperform those of large-scale sequential traffic datasets. In Test 2, LSTM achieved the best results of MSE = 3.77, RMSE = 1.94, and MAE = 1.47, confirming their greater ability to model long-term temporal dependencies and retain historical traffic information. This study shows that, by incorporating vehicle dimensional features in a traffic density projection system, the representation of real-time traffic occupancy conditions such as traffic density is significantly improved. Although GRU feature learning has faster computational performance, LSTM provides more stable and accurate prediction results in complex traffic density prediction tasks. In summary, the proposed framework shows potential for Intelligent Transportation Systems (ITS) applications in real-time traffic monitoring and adaptive traffic management. Future work can be done to incorporate more traffic-related characteristics and appropriate hybrid deep learning structures to improve prediction accuracy and model robustness.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors who supported me during the whole course of the doctoral program. They provided many insights and comments were of great help in determining the scope and direction of this paper.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT (10 PT)

The main research activities were performed by **Author 1**, who had a major role in conceptualization, data collection, system implementation and experimentation, analysis and manuscript preparation. **Authors 2,3 and 4** acted as research supervisors by providing academic input, methodological guidance, technical assessment and critical review throughout the course of the work leading to this manuscript.

REFERENCES

[1] A. Agustan, T. S. Soeparyanto, T. Azikin, L. Welendo, U. Mangidi, and I. Isnawaty, "A Systematic Mapping Study On Multi-Algorithm Methods For Optimizing Transportation Systems," *J. INFOTEL*, vol. 17, no. 3, pp. 516–536, 2025.

[2] H. Khan *et al.*, "Machine learning driven intelligent and self adaptive system for traffic management in smart cities," *computing*, vol. 104, no. 5, pp. 1203–1217, 2022.

[3] P. Samuel and P. K. Sharma, "Intelligent Traffic Management System using Artificial Intelligence and Computer Vision," *Int. J. Res. Technol.*, vol. 13, no. 4, pp. 68–80, 2025.

[4] S. Feng, H. Qian, H. Wang, and W. Wang, "Real-time object detection method based on YOLOv5 and efficient mobile network," *J. Real-Time Image Process.*, vol. 21, no. 2, p. 56, 2024.

[5] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-time vehicle detection based on improved yolo v5," *Sustainability*, vol. 14, no. 19, p. 12274, 2022.

[6] J. Wang, Y. Chen, Z. Dong, and M. Gao, "Improved YOLOv5 network for real-time multi-scale traffic sign detection," *Neural Comput. Appl.*, vol. 35, no. 10, pp. 7853–7865, 2023.

[7] M. S. Hammoud and S. Lupin, "Optimization of road detection using semantic segmentation and deep learning in self-driving cars," *Ann. Emerg. Technol. Comput.*, vol. 8, no. 3, 2024.

[8] R. Garg, S. Makhmudov, T. Eshchanov, A. D. Mansurovich, M. Garg, and A. Karn, "Enhanced Stock Market Movements towards Economic Sustainability using LSTM Hybrid Algorithm," in *2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE)*, IEEE, 2025, pp. 792–796.

[9] S. Kim, H. Kim, and Y. Hwang, "Data-driven dynamic response forecasting and anomaly detection in long-span bridges," *J. Civ. Struct. Heal. Monit.*, vol. 15, no. 7, pp. 3045–3062, 2025.

[10] D.-J. Pyo, "Enhancing GDP Growth Forecasting with LSTM, GRU, and Hybrid Model: Evidence from South Korea," *SAGE Open*, vol. 15, no. 3, p. 21582440251359828, 2025.

[11] T. Azfar, J. Li, H. Yu, R. L. Cheu, Y. Lv, and R. Ke, "Deep learning-based computer vision methods for complex traffic environments perception: A review," *Data Sci. Transp.*, vol. 6, no. 1, p. 1, 2024.

[12] W. Zhou *et al.*, "Vision technologies with applications in traffic surveillance systems: A holistic survey," *ACM Comput. Surv.*, vol. 58, no. 3, pp. 1–47, 2025.

[13] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, 2013.

[14] C. Yanmin, A. Sarkar, J. M. Zain, A. Bhar, A. Noorwali, and K. M. Othman, "Leveraging LSTM and GRU-based deep neural coordination in intelligent transportation to strengthen security in the Internet of Vehicles," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 4, pp. 2431–2467, 2025.

[15] X. Fei, F. Long, F. Li, and Q. Ling, "Multi-component fusion temporal networks to predict vehicle exhaust based on remote monitoring data," *Ieee Access*, vol. 9, pp. 42358–42369, 2021.

[16] H. Zou, K. Cao, and C. Jiang, "Spatio-temporal visual analysis for urban traffic characters based on video surveillance camera data," *ISPRS Int. J. Geo-Information*, vol. 10, no. 3, p. 177, 2021.

[17] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura, "Understanding traffic density from large-scale web camera data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5898–5907.

[18] M. T. Shahriar and H. Li, "A study of image pre-processing for faster object recognition," *arXiv Prepr. arXiv2011.06928*, 2020.

[19] M. H. Hamzenejadi and H. Mohseni, "Fine-tuned YOLOv5 for real-time vehicle detection in UAV imagery: Architectural improvements and performance boost," *Expert Syst. Appl.*, vol. 231, p. 120845, 2023.

[20] Y. Yao *et al.*, "On improving bounding box representations for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2022.
[21] A. Graves, "Long short-term memory," *Supervised Seq. Label. with Recurr. neural networks*, pp. 37–45, 2012.
[22] F. M. Salem, "Gated RNN: the gated recurrent unit (GRU) RNN," in *Recurrent neural networks: from simple to gated architectures*, Springer, 2021, pp. 85–100.
[23] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," in *IOP conference series: materials science and engineering*, IOP Publishing, 2018, p. 12049.
[24] Y. Hu, R. Jia, Y. Liu, Y. Li, and H. Sun, "Knowledge-Based Systems WSNet : A local – global consistent traffic density estimation method based on weakly supervised learning," *Knowledge-Based Syst.*, vol. 255, p. 109727, 2022, doi: 10.1016/j.knosys.2022.109727.
[25] Z. Hu, W. H. K. Lam, S. C. W. Andy, and H. F. C. Wei, "Turning traffic surveillance cameras into intelligent sensors for traffic density estimation," *Complex Intell. Syst.*, vol. 9, no. 6, pp. 7171–7195, 2023, doi: 10.1007/s40747-023-01117-0.
[26] T. Arif, A. Salyh, M. Mussa, M. Rahman, and M. Alam, "Vision-based real-time traffic flow monitoring system for road intersections in Dhaka city," 2025.