Data Driven Machine Learning Model for Traffic Flow Forecasting using VANET

Praveen

Computer Science and Engineering, Gautam Buddha University, Greater Noida, India
Email: *praveengautam1997[at]gmail.com*

Abstract: ***This study focuses on the design and development of data-driven machine learning model using VANET sensing. The proposed model leverages the dynamic and decentralized nature of VANETs to gather extensive traffic-related data from various sensors embedded in vehicles. Advanced machine learning algorithms LSTM networks, CNNs, and hybrid models, are employed to analyze and predict traffic flow patterns. This model with VANET data aims to address the complexities and nonlinearities inherent in traffic dynamics. This study advances the subject of intelligent transportation systems by providing a scalable and helpful traffic management solution. The implementation of this model can lead to reduced traffic congestion, lower travel times, and enhanced road safety. Future work will explore the formation of additional data sources and the utilization of more intelligent machine learning techniques to further improve the robustness and accuracy of traffic flow forecasts.***

Keywords: Vehicular Ad-hoc Networks (VANETs), Machine learning models, LSTM networks, CNNs

1. Introduction

Many nations are experiencing extreme traffic congestion due to the sharp the number of cars on the road has increased, which has nearly overwhelmed the capacity of our present transportation systems. However, adding more road infrastructure is not a desirable choice due to its expensive and constrained location. In one like that, the construction of a HOV lane in the city of Los Angeles can cost up to $750,000 per lane and per mile. Specifically, the costs are exorbitant to build additional infrastructure to maintain traffic flow throughout the building phase and for ensuring the safety of construction workers. One useful tool is a traffic monitoring system substitute to reduce congestion in the traffic. It is an essential component of ITS, which are used to accumulate traffic data, such as the number, type, and speed of automobiles. Using the data gathered, it does traffic analysis to improve transportation safety, estimate future needs for transportation, and make better use of the road networks overall. Developing, implementing, and maintaining traffic monitoring systems costs enormous amounts of money for transportation agencies in many different countries. Accurately classifying cars into various groups is essential for efficient traffic management and transportation planning. For instance, the capacity of a highway segment and the scheduling of pavement maintenance activities can be estimated using the number of heavy trucks on the section.

2. Related Studies

Won, M. [1] reviews contemporary traffic monitoring systems, focusing on automobile classification as a crucial component of ITS, examining various classification methods enabled by advances in wireless communication, AI, and MEMS technologies, while analyzing their performance, architectural aspects, and implementation challenges. **Choudhary, A. et al. [2]** presents an efficient traffic controls system which illustrate improvements over the existing manual traffic management system.

Abbasi et al. [4] present a comprehensive review of deep learning applications in NTMA, highlighting how these models effectively handle the complex traffic patterns and massive data generated by modern networks (including cellular networks and IoT), particularly focusing on traffic prediction and categorization where traditional management techniques fall short. **Akhtar et al. [5]** provide a systematic review of AI and machine learning approaches for traffic congestion prediction, examining how these models utilize both historical and real-time data from stationary sensors and probe vehicles to assess various traffic factors, particularly focusing on short-term congestion forecasting.

Bhuvan et al. [6] proposed an IoT-based intelligent traffic management system that combines integrated and decentralized approaches, utilizing cameras, sensors, RFID technology, and AI-based algorithms to monitor traffic density, predict congestion, prioritize emergency vehicles, and detect hazardous situations like fires through smoke sensors. **Ali R. Abdellah et al. [11]** propose a LSTM deep learning model for predicting VANET network traffic, training it with collected traffic data and evaluating its prediction accuracy using RMSE and MAPE metrics to enable Intelligent Transport Systems (ITS) to proactively respond to traffic events.

***Berlotti et al.* [14]** present a machine learning-based system for predicting traffic flow patterns and providing real-time congestion mitigation solutions at busy intersections, aiming to enhance traffic efficiency and reduce economic and social costs through data-driven traffic management.

***Sunny et al.* [15]** develop a deep learning-based traffic flow prediction model that analyzes extensive traffic data to provide real-time insights into traffic patterns and density, offering an effective solution for traffic management and congestion reduction at critical intersections within urban transportation networks.

***Afzali, M. et al.* [16]** developed a hybrid traffic flow prediction model that combines artificial neural networks and genetic algorithms using VANET data to improve prediction accuracy and reduce traffic congestion.

3. Research Methodology

The data-driven machine learning model for traffic flow forecasting using a VANET involved several key steps. Initially, a comprehensive literature review was conducted to identify the state-of-the-art techniques in traffic flow forecasting and vehicular ad-hoc networks (VANETs).

First, certain redundant, missing, and illogical values were eliminated from the raw dataset based on the results. After that, we performed basic pre-processing by normalizing our data using Standard Scaler normalization, which scaled all feature values within the range of [0, 1]. The definition of the formula is:

$$Z_{scaled} = \frac{(X - \mu)}{\sigma} \quad (1)$$

Where σ is the standard deviation, μ is the mean, and X is the input variable. The key elements that will be covered in Section II were chosen after normalization. Assuming variable X with (i) the number of selected features as the input data, the study finally divided the dataset into training and testing sets. The target variable (y), which is the traffic prediction, was thus labeled and studied in two classes: traffic (1) and no traffic (0).

Table 1: Comparative Study of IR and ML in Traffic Flow Forecasting

Aspect	*Information Retrieval (IR)*	*Machine Learning (ML)*
Data Collection	Uses traditional methods and sensors	Utilizes advanced sensors (GPS, LIDAR)
Data Processing	Basic preprocessing and normalization	Advanced preprocessing, normalization, and feature selection
Prediction Techniques	Rule-based systems, manual analysis	Regression models, neural networks, ensemble methods
Real-time Analysis	Limited capabilities	High capabilities with real-time data analysis
Efficiency	Lower due to manual intervention	Higher due to automated learning and predictions
Applications	Historical data analysis, basic traffic patterns	Predicting traffic flow, congestion forecasting

a) Overview of the stk-ebm model architecture

Data from V2I and V2V in VANETs was gathered, combined, and pre-processed in the first section. Using LightGBM and Boruta techniques, the study examined the data's most informative features. We separated our dataset into training and testing at the end of the first part so that the ML model could use it to make predictions. The model's components—the base learner element selection, the meta learner, and the stacking heterogeneous ensemble model structure—are covered in the next subsections. stacking group education. two types of learners: base and meta. The stacking optimised heterogeneous ensemble model was developed as a solution to the network traffic forecast problem (STK–EBM).

b) Units

The most widely used classification evaluation measures in research were the confusion matrix, classification report, and CPU time. Further, took into account the Area Under Curve (AUC) to comprehend the stability of the model and the Receiver Operating Characteristic (ROC) curve, which is a well-known tool for evaluating the performance of binary classifiers.

The link between the actual and projected classes is shown in Table 1. The accuracy of a favorable forecast is shown by precision. Precision and recall have an impact on the F1 score, with 1.0 being the optimal value.

The classification algorithms' prediction quality is evaluated using the Classification Report. The precision of a model is defined as the proportion of predicted positives that are actual positives.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (2)$$

The number of Actual Positives that the model accurately detects is calculated by recall.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (3)$$

The F1 Score aims to strike a balance between recall and precision.

$$\text{F 1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

AUC, or area under the curve an unbiased estimator of a predictor f's AUC: determines if positives are given a better ranking than negatives.

$$AUC_{(F)} = \frac{\sum t_0 \in D^0 \sum t_1 \in D^1 1[f(t_0) < f(t_1)]}{|D^0| . |D^1|} \quad (5)$$

The set of negative examples is D 0, whereas the set of positive examples is D 1. The indicator function 1[f (t0) < f (t1)] returns 1 if f (t0) < f (t1); otherwise, it returns 0.

4. Implementation

a) Data Set

The dataset gives an extensive overview of the data used in the design data-driven machine learning model for traffic flow forecasting using a VANET-enabled vehicular sensing framework. This chapter details the sources, characteristics, and preprocessing steps of the dataset, which forms the foundation of the machine learning model. The dataset comprises extensive traffic data collected from multiple sources. Subsequently, the chapter explores the exploratory data analysis (EDA) performed to gain insights into the dataset. The actual and predicted traffic volumes during various hours and days of the week, together with the prediction errors, are displayed using heatmaps. The relationships between the various variables in the dataset can be enhanced with the use of correlation and scatter matrices.

b) *Data Cleaning*

When creating a data-driven machine learning model for traffic flow forecasting with a VANET-enabled vehicular sensing framework, cleaning the data is an essential step.

Achieving accurate and trustworthy predictions requires maintaining the dataset's quality and integrity. Numerical feature scaling and normalization are critical processes for standardizing the dataset and enabling effective model training. In order to ensure that every characteristic contributes suitably to the prediction models, this chapter describes the methods used to convert the data into an organized way. Furthermore, appropriate methods are used to encode categorical variables so that machine-learning algorithms can process them in numerical formats. The ultimate objective of these cleaning processes is to generate a dependable, high-quality dataset that faithfully captures actual traffic situations, improving the machine learning models' efficiency.

5. Result

Solution that is adaptable to different urban environments and can be scaled. Top of Form

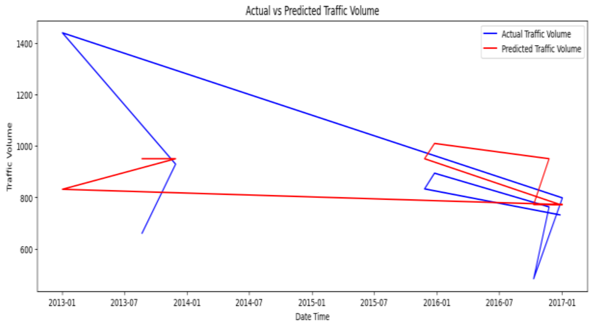

Figure 1: Actual vs Predicated Traffic Volume

This graph shows "Actual vs Expected Traffic Volume" compares the actual and expected traffic volumes from January 2013 to January 2017. Time is represented by the x-axis, and traffic volume is shown by the y-axis. The real traffic volume is represented by the blue line, which has notable variations during the period due to its strong peaks and declines. The red line, which represents the predicted traffic volume, on the other hand, shows a pattern that is less variable and more constant. There are notable differences between the actual predicted volumes in early 2013 and mid-2016, with the real numbers displaying rapid expands that the model is unable to correctly forecast. These variances suggest that the prediction model has to be further refined to improve its response to unexpected changes in traffic volume, even while it captures broad trends but struggles with rapid changes.

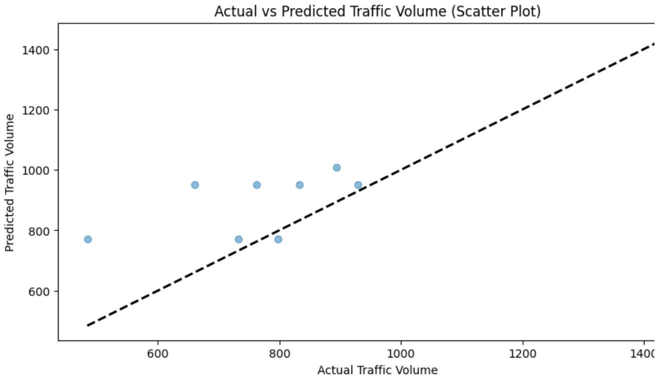

Figure 2: Actual vs predicated traffic volume (scatter plot)

This scatter plot "Actual vs Predicted Traffic Volume" displays the relationship between the actual traffic volumes (x-axis) and the predicted traffic volumes (y-axis). A data pair of actual and expected values is represented by each point on the plot. The ideal situation, when predictions exactly match actual values (i.e., y=x), is represented by the diagonal dashed line. Perfect forecasts are shown by points that fall on this line. The observed data points, on the other hand, diverge from the line, indicating differences between the anticipated and actual traffic levels and pointing out potential areas for improvement in the predictions.

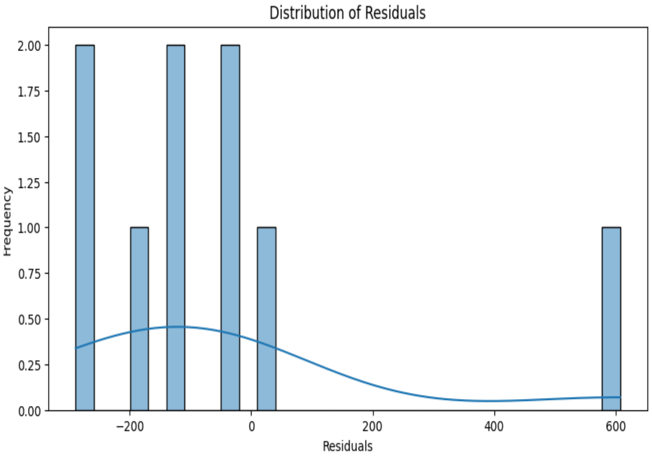

Figure 3: Distribution of Residuals

The frequency distribution of residuals, or the variations between actual and expected traffic volumes, is shown in the "Distribution of Residuals" graph. The x-axis represents the residual values, while the y-axis represents their frequency. The superimposed curve gives a smoothed estimate of the distribution, and the blue bars indicate the frequency of each residual value. A number of important details regarding the residual distribution are shown by the plot. First, the residuals show a slightly symmetric distribution around zero, indicating that both underestimations and overestimations are included in the model's predictions, which are generally balanced. Additionally, a significant number of residuals are clustered around zero, indicating that many of the model's predictions are close to the actual values. However, there are notable outliers, particularly around -200 and 600, indicating instances where the model significantly underestimated or overestimated the traffic volume. Further analysis and model improvements are needed to address these discrepancies and enhance overall prediction accuracy.

Paper ID: SR241207130435 DOI: https://dx.doi.org/10.21275/SR241207130435

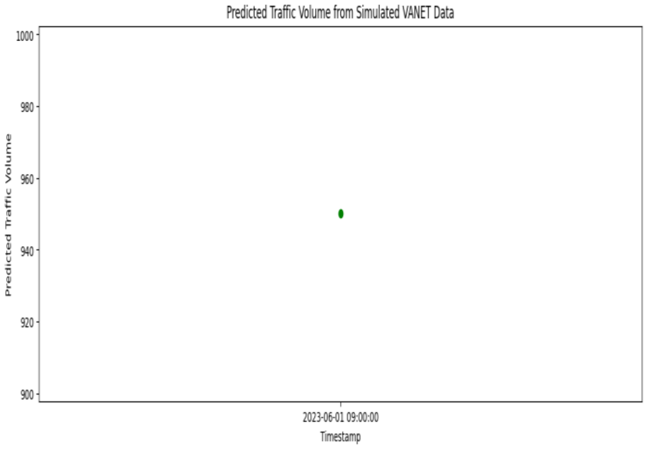

Figure 4: Predicted Traffic Volume from simulated VANET Data

This figure shows the expected traffic volume from simulated VANET data for a specific timestamp, depicted on the horizontal axis as "2023-06-01 09:00:00." The vertical axis represents the predicted traffic volume, which is approximately 950. The plot includes a single data point, marked by a green dot, indicating that the prediction at this timestamp falls within the range of 900 to 1000. The centralized nature of the data point suggests a focused prediction without any surrounding variability, implying either a snapshot prediction or a single instance of traffic volume prediction from the simulation. This visualization provides a concise view of the traffic volume at the given timestamp, useful for understanding traffic flow predictions in a vehicular network scenario. The limited data points highlight the need for further simulation data to analyze trends or changes over time.

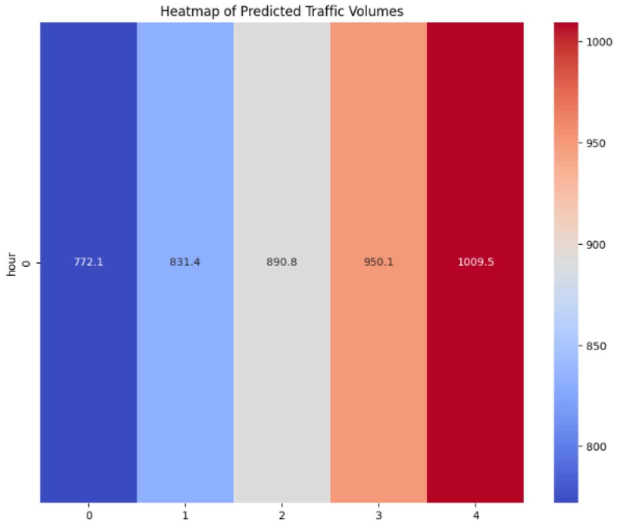

Figure 8: Heatmap of the predicted traffic volumes

This heatmap depicts the expected traffic volumes over a sequence of five hours. The color gradient ranges from blue to red, with blue standing for lower traffic volumes and red showing higher volumes. At hour 0, the expected traffic volume is the smallest at 772.1, as indicated by the dark blue color. As the hours go by, the traffic volume rises. At hour 1, the volume rises to 831.4, shown by a lighter blue shade. By hour 2, the traffic volume reaches 890.8, indicated by a neutral gray color. The quantity keeps going to climb to 950.1 at hour 3, represented by a light orange shade. Lastly, at hour 4, the traffic volume peaks at 1009.5, pointed out by a deep red color.

6. Conclusion

The analysis of traffic volume projections and actual data over the period from January 2013 to January 2017 indicates several key insights into the model's functionality and potential improvements. The analogy of actual and projection traffic volumes indicates that while the model captures general patterns, it struggles with quickly fluctuations, especially notable in early 2013 and mid-2016. The distribution of residuals demonstrates a somewhat symmetrical pattern around zero, with notable outliers suggesting areas for model modification. A specific timestamp expected from simulated VANET data demonstrates a focused expected but underscores the need for more data to examine trends over time. Advanced machine learning methods that can better handle unplanned variations in traffic patterns. For example, applying deep learning models that are intended to capture temporal dependencies, like recurrent neural networks (RNNs) or LSTM networks, could greatly enhance the model's capacity to forecast rapid shifts in traffic flow.

Essential focus for future research is enhancing the classification model to reduce incorrect classifications, particularly between closely related bins. Misclassifications, as highlighted in the confusion matrix examination, indicate that the model struggles to differentiate between certain traffic volume ranges. Addressing this issue could involve studying advanced classification methods or refining the existing model's parameters. Future research could investigate advanced outlier detection and handling techniques to ensure that these anomalies do not particularly affect the model's forecasts. Techniques such as robust statistical techniques, anomaly detection algorithms, and regularization method can help reduce the impact of outliers and improve overall model stability. Continued validation and testing with real-world data are essential to make sure the model remains robust and accurate across different situations. In dynamic environments like VANETs, where traffic conditions can change quickly, ongoing validation is crucial. Creating a framework for continuous model observing and updating will help maintain high predictive accuracy over time. This could involve setting up computerized pipelines for data gathering model training, and evaluation, allowing the model to quickly modify to new trends and changes in traffic behavior.

References

[1] Won, Myounggyu. "Intelligent traffic monitoring systems for vehicle classification: A survey." IEEE Access 8 (2020): 73340-73358.
[2] Choudhary, Abhijeet, et al. "Artificial Intelligence Based Smart Traffic Management System Using Video Processing." Artificial Intelligence 5.03 (2018).
[3] Lee, Sangmin, et al. "Intelligent traffic control for autonomous vehicle systems based on machine learning." Expert Systems with Applications 144 (2020): 113074.

[4] Abbasi, Mahmoud, Amin Shahraki, and Amir Taherkordi. "Deep learning for network traffic monitoring and analysis (NTMA): A survey." Computer Communications 170 (2021): 19-41.

[5] Akhtar, Mahmuda, and Sara Moridpour. "A review of traffic congestion prediction using artificial intelligence." Journal of Advanced Transportation 2021 (2021): 1-18.

[6] Bhuvan, S.T., et al. "Smart Traffic Management System: A Literature Review." IJIREEICE 10 (2022): 10201

[7] Saleem, Muhammad, et al. "Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques." Egyptian Informatics Journal 23.3 (2022): 417-426.

[8] Ahmed, Adel A., et al. "Smart traffic shaping based on distributed reinforcement learning for multimedia streaming over 5G-VANET communication technology." Mathematics 11.3 (2023): 700.

[9] Zhu, Ruijie, et al. "Multi-agent broad reinforcement learning for intelligent traffic light control." Information Sciences 619 (2023): 509-525.

[10] Aroba, Oluwasegun Julius, et al. "Adoption of Smart Traffic System to Reduce Traffic Congestion in a Smart City." International Conference on Digital Technologies and Applications. Cham: Springer Nature Switzerland, 2023.

[11] Abdellah, A. R., & Koucheryavy, A. (2020). Vanet traffic prediction using lstm with deep neural network learning. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12525 LNCS (Issue December 2020). Springer International Publishing. https://doi.org/10.1007/978-3-030-65726-0_25

[12] Khatri, S., Vachhani, H., Shah, S., Bhatia, J., Chaturvedi, M., Tanwar, S., & Kumar, N. (2021). Machine learning models and techniques for VANET based traffic management: Implementation issues and challenges. Peer-to-Peer Networking and Applications, 14(3), 1778–1805. https://doi.org/10.1007/s12083-020-00993-4

[13] Christalin, N. S., Mandal, T. K., & Prakash, G. L. (2022). A Novel Optimized LSTM Networks for Traffic Prediction in VANET. Journal of System and Management Sciences, 12(1), 461–479. https://doi.org/10.33168/JSMS.2022.0130.

[14] Mariaelena, Berlotti., Sarah, Di, Grande., Salvatore, Cavalieri. (2024). Proposal of a Machine Learning Approach for Traffic Flow Prediction. doi: 10.3390/s24072348

[15] Mehdi, Afzali. (2023). Traffic flow prediction based on vanet data by combining artificial neural network and genetic algorithm. Azerbaijan journal of high performance computing, doi: 10.32010/26166127.2023.6.1.91.112

[16] M.J.M., Sunny. (2023). Vehicular Traffic Flow Prediction Model using Deep Learning. International Journal For Science Technology And Engineering, 11(7):275-279. doi: 10.22214/ijraset.2023.54605

[17] K., Suganyadevi., V., Swathi., T., Santhiya., S.Siva, Sankari., V., Swathi. (2023). Machine Learning Algorithm based VANET Traffic Management System. 747-752. doi: 10.1109/iceca58529.2023.10395426

[18] Daniel, Lane., Subhradeep, Roy. (2024). Validating a data-driven framework for vehicular traffic modeling. doi: 10.1088/2632-072x/ad3ed6