Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 25 - 39 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Vehicle Traffic Flow Forecasting on Caltrans PeMS Dataset Using Machine Learning Algorithms and LSTM Networks

Jiarui Chang, Rice University (Houston, USA)
Jingwen Du, Cornell University (Ithaca, USA)
Hojin Chung, Gyeonggi Suwon International School (Suwon, Korea)

ABSTRACT:

In Intelligent transportation systems, accurate traffic flow prediction is fundamental in transportation modeling and management.Previous studies have classified prediction approaches into three categories including a time series approach with ARIMA model for finding traffic flow patterns and using those patterns for prediction, a probabilistic approach for modeling and forecasting from a probabilistic perspective, and nonparametric approaches that can perform better by handling un-deterministic and complex time series traffic datasets.This paper analyzes historical timeseries traffic data from sensors using machine learning algorithms as baseline models and designs a deep learning LSTM model to train using the historical dataset to forecast traffic flow using the trained model. The paper also compares the performance of machine learning algorithms and the deep learning model. The results show the deep learning LSTM model to outperform machine learning models.

KEYWORDS: Deep learning model, Traffic flow prediction, Caltrans PeMs dataset, LSTM model

I. Introduction:

Traffic flow prediction is a major issue in intelligent traffic systems and for public as well as private sector. It helps road users make better travel decisions and enable reduced carbon emissions and improved traffic conditions. Accurate traffic flow prediction on real-time basis providesroad users with information to optimize travel decisions and reduce travel costs, also helping traffic authorities to better mitigate congestion.

However, accurate traffic prediction is a challenging problem. Traditional traffic prediction methods include models such as autoregressive integrated moving average(ARIMA),multi-variable linear regression, and support vector regression. However, these linear models do not consider the whole range of features in traffic flow and thus do not perform optimally [8].In addition, because of stochastic and nonlinear featuresof traffic flow, parametric approaches with linearity cannot provide high traffic flow prediction performance, motivating greater attention to nonparametric approaches.[5]

Without accurate traffic flow prediction, no intelligent transportation systems could optimally perform. Previous studies have addressed this problem and classified prediction approaches into three categories including time series approach with ARIMA model for finding patterns of traffic flow and using those patterns for prediction, probabilistic approach

for modeling and forecasting from probabilistic perspective, and nonparametric approaches that performed better due to their ability to handle undeterministic and complex time series traffic datasets.

Deep learning[8] is a nonparametric approach and a type of machine learning based on neural networks. Through dependency in high-dimensional sets of variables, clear discontinuities in traffic flow emerging in large-scale networks can be captured. Deeplearning is increasingly seen as an essential tool for artificial intelligence research in areas such as traffic flow prediction. Recent years have seen the use of deep learning in traffic prediction, which this paper focuses on.

The rest of this paper is organized as follows: Section II provides a literature review. Section III discusses the materials and methods, and Section IV explains the experiments. Section V addresses the results and discussion, and Section VI concludes.

II. Literature Review:

Dai et al. [1] consider temporal patterns in traffic flow and propose a deep learningmodel for traffic flow prediction by considering DeepTrend, a deep hierarchicalneural network for predicting traffic flow based on time-variant trends. They find that DeepTrend can improve prediction performance for some popular prediction models.

Chen et al. [2] propose a novel fuzzy deep-learning approach called FDCN to better predict city traffic flow based on fuzzy theory and a deep residual network model, introducing fuzzy representation to reduce data uncertainty and proposing pretraining and fine-tuning strategies for more efficient learning of FDCN parameters. They find the proposed approach to outperform existing approaches.

Manoranjitham et al. [3] demonstrate potential benefits of deeplearning in short-term traffic flow prediction by using traffic flow data to train adeep neural network to recognize traffic patterns and provide short-term forecasts. They highlight the potential of existing GPS-based systems in improving traffic prediction accuracy and efficiency.

Jia et al. [4] introduce the deep belief network (DBN) and long short-term memory (LSTM) to better predict urban traffic flow in rainfall conditions, finding the capability of rainfall-integrated DBN and LSTMto learn traffic features and showing deep learning predictors to have better accuracy than existing predictors.

Yang et al. [5] connect long time step sequences to currenttime steps by including high-impact traffic flow values using the attentionmechanism and smoothening data beyond normal rangesfor better prediction, demonstrating the proposed predictionmodel to be better for short-term traffic flow prediction.

Kenworthy-Groen [6] reviews traffic data from three metropolitan arterialroads in Perth and compares the traditional compound traffic growth rate model to the linear traffic growth rate model, highlighting sound long-term traffic data to ensure appropriate traffic growth rate models for optimal sustainable pavement projects.

Polson and Sokolov [7] develop a deep learning model for traffic flow prediction by combining a linear model fitted using 1 regularizationand a sequence of tanh layers, demonstrating deep learning architecture to capture nonlinear spatiotemporal effects and provide accurate short-term traffic flow predictions.

Du et al. [8] propose a hybrid multimodal deep learning method for short-term traffic flow forecasting, namely one-dimensional Convolutional Neural Networks (1D CNN) and Gated Recurrent Units (GRU) with attention mechanism. They incorporate representation features of modality traffic data and find the proposed model to accurately predict complex nonlinear urban traffic flow.

Lv et al. [9] proposea novel model called LC-RNN for traffic speed prediction by considering RNN andCNN models. They also propose a network-embedded convolution structure to better incorporate topology-aware features and consider periodicity and other context factorsfor better prediction accuracy, finding the proposed LC-RNNto outperform some popular methods.

Wang et al. [10] evaluate a path-based deep learning framework for traffic speedprediction by dividing a road network into core paths and modeling each path based on the bidirectional long short-term memory neural network (Bi-LSTM NN). They find the proposed model to outperform various benchmark methods.

Zhang et al. [11] propose a method based on the cascaded artificial neuralnetwork (CANN) to predict traffic flow by incorporating actualroad network distance into the model and using real-world data from video surveillance cameras in Xiamen, China.

Xiao and Yin [12] propose a hybrid Long Short–Term Memory (LSTM) neural networkbased on the LSTM model and optimize it for various traffic environments. The prediction error of the hybrid LSTM model is lower than others but requires a slightly longer running time.

Tian et al. [13] consider a novel approach based on Long Short-TermMemory (LSTM) and multiscale temporal smoothing, demonstrating its higher accuracy in traffic flow prediction.

III. Materials & Methods:

a) Dataset:

This study uses the California department of transportation (Caltrans) dataset.In Caltrans Performance Measurement System (PeMS) dataset, data are collected on a real-time basis from individualsensors along the freeway system across all major metropolitan areas of California.

PeMS provides real-time data from over 39,000 sensors and is an Archived Data User Service (ADUS), which provides more than a decade of historical data.The dataset had 7776 instances for training.

The study focuses on analyzing historical traffic data in PeMs dataset and considers the traffic prediction model at various time interval using Long short-term memory (LSTM) networks to predict traffic flow during peak and nonpeak hours for a given city. The study also compares the LSTM model with various other baseline models.

b) Metrics:

Two popular performance indices used as metrics include the mean absolute percentage error(MAPE) and the RMS error (RMSE). MAPE measures prediction accuracy of a forecasting method typically shown in as a percentage, and RMSE is the standard deviation of residuals (prediction errors). This study's LSTM model uses Keras deep learning library, which calculates and provides a suite of standard metrics in training deep learning models. In addition, Keras defines and gives custom metrics in training deep learning models, which is useful when tracking performance measures. The Results and Discussion section provides a comparison of various algorithms with respect to metrics for a period of 5 minutes.

c) Methodology:

For experiments, a process flow is followed where the initial phase gathers required datasets from repositories (Fig.1). After relevant datasets are obtained, the data are cleaned in

preprocessing stage, followed by feature selection to identify important features that are correlated. After feature selection, training is conducted using training data, and the model will be saved if performance metrics are satisfactory. Otherwise, training is repeated using additional training data.

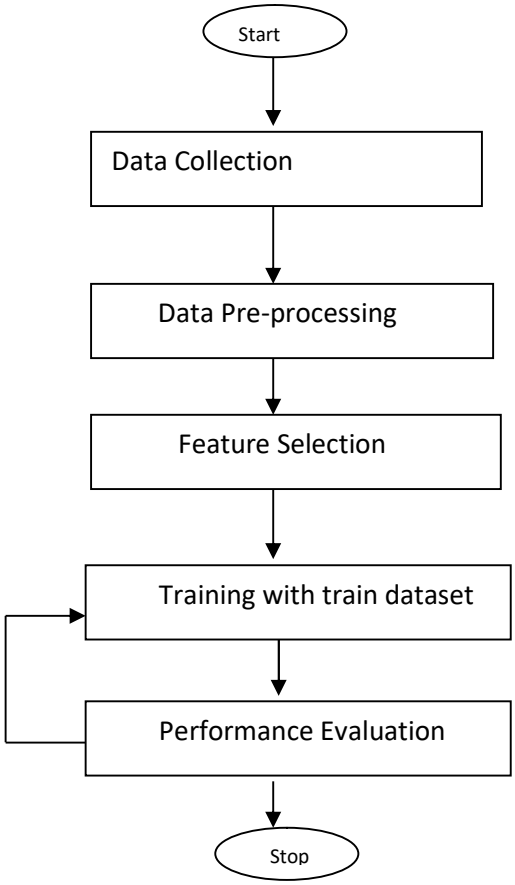


Fig.1. Methodology

IV. Experiments

The experiments are conducted in both machine learning and deep learning models. For machine learning models, WEKA tools are used to train the model with the training dataset, and performance metrics are measured using test data.Weka 3.8 is used fordata preprocessing,and regression models are trained using popular models such as Linear Regression, Multi-Layer Perceptron (MLP), RBF Network, RBF Regressor, and SMO Reg algorithms.

The experimentsuse a dataset with 7776 instances for training obtained from PeMS. As discussed in Section 4.1, the experiments are conducted using machine learning models. Section 4.2 discusses experiments using the deep learning model. For LSTM models, 5

different cases with increasing numbers of epochs are considered with reduced errors and increased performance.

a) Machine Learning Models

(i) Linear Regression

Linear regression is used to determine the linear relationship between the target and one or more predictors. Simple linear regression is useful for determining relationships between two continuous variables. One is the predictor or independent variable, and the other is the response or dependent variable. Here statistical, not deterministic, relationships are focused on. Relationships between two variables are considered deterministic if one variable is accurately expressed by the other. A statistical relationship is not accurate in determining the relationship between two variables. Here the core idea is to obtain the best-filling line, that is, the one for which the total prediction error (all data points) is the smallest. Error is the distance between a point to the regression line

(ii) Multi-Layer Perceptron

The multilayer perceptron (MLP) is a class of feedforward artificial neural networks. The MLP consists of at least three layers of nodes: input, hidden, and output. Except for input nodes, each node is a neuron using a nonlinear activation function. The MLP uses a supervised learning technique called back propagation for training, and its multiple layers and nonlinear activation distinguish the MLP from the linear perceptron. It can distinguish data that are not linearly separable.

(iii) RBF Network

The RBF network is an artificial neural network with input, hidden, and output layers. The hidden layer includes hidden neurons whose activation functionis a Gaussian function. The hidden layer generates a signal corresponding to an input vector in the input layer, and the network generates a response corresponding to the signal.

(iv) RBF Regressor

The RBF network trains hidden layers in an unsupervised manner, and RBFRegressor and RBFClassifier are fully supervised. RBFNetwork implements a normalized Gaussian radial basis function network and uses the k-means clustering algorithm for basis functions, also learning either logistic regression (discrete class problems) or linear regression (numeric class problems). Symmetric multivariate Gaussians are fit to data from each cluster, and if the

class is nominal, it uses a given number of clusters per class. RBFRegressor implements Gaussian radial basis function networks for regression, trained fully supervised using WEKA Optimization class by minimizing squared error using BFGS. Here it is possible to use conjugate gradient descent instead of BFGS updates, which is faster with many parameters, and normalized basis functions instead of unnormalized ones. RBFClassifier is the equivalent of RBFRegressor in classification problems.

(v) SMO Regressor

The sequential minimal optimization algorithm (SMO) is effective in training support vector machines (SVMs) for classification defined on sparse datasets. SMO differs from many other SVM algorithms in that it requires no quadratic programming solver. SMOreg implements the support vector machine for regression, and parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer.

b) Deep Learning Model

LSTM Networks

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) in deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections and can not only process single data points but also whole data sequences. LSTM networks are optimal for data collected over a time period at regular intervals, namely timeseries data.

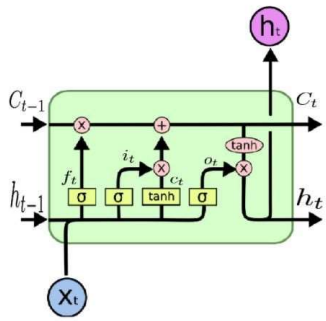


Fig.2. LSTM cell

A common LSTM unit is composed of a cell and input, output, and forget gates. The cell remembers values over arbitrary time intervals, and these three gates regulate information flow across the cell.Fig. 2 shows a simple LSTM cell.

LSTM networks are optimal for classifying, processing and making predictions using time series data since there may be lags of unknown duration between important events in time series. LSTMs can address exploding and vanishing gradient problems that can be encountered when training traditional RNNs. Relative lack of sensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov, and other sequence learning methods.

V. Results & Discussion

a) Linear Regression

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.6093	8.5389	9.6568	10.7303	11.7833	12.8264	13.7602	14.7962	15.8355	16.7475
RAE	90.444	101.481	104.5887	105.7303	105.3203	105.5449	104.8005	106.0831	107.0088	105.768
RMSE	10.3206	11.6506	13.1329	14.5647	15.9731	17.3717	18.6284	19.9162	21.1754	22.3
MSE	106.5142	135.7361	172.4725	212.1391	255.1391	301.7745	347.0184	396.6546	448.3987	497.2889

Table 1. Linear Regression metrics

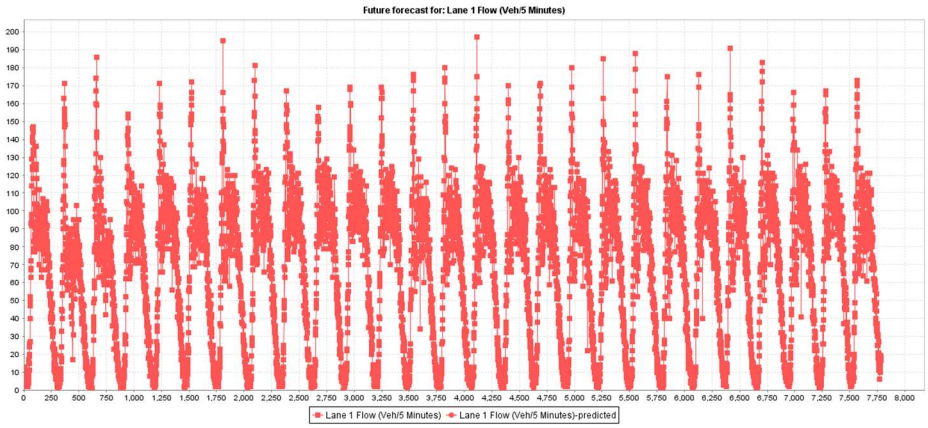


Fig.3. Future forecast for target using linear regression

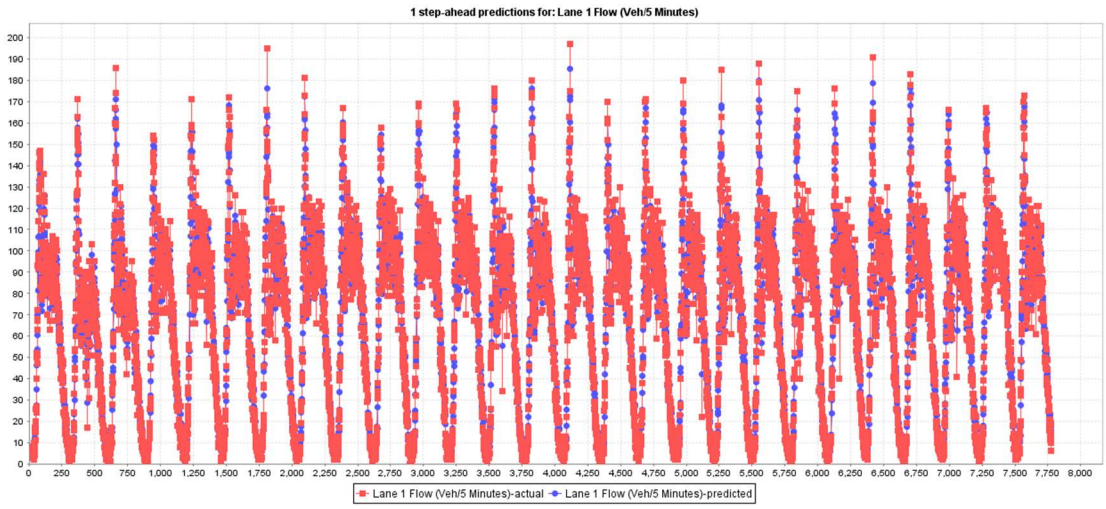


Fig.4. 1-step ahead predictions for target using linear regression

b) Multi-Layer Perceptron

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	11.2635	13.7831	16.3976	18.8599	21.0923	23.4731	25.8266	28.3938	31.0433	33.5108
RAE	133.863	163.8068	177.5945	184.5593	108.5255	193.1534	196.7012	203.5733	209.7759	211.6359
RMSE	13.6397	16.3185	19.0884	21.7114	24.1751	26.8254	29.4279	32.2339	35.1494	37.9459
MSE	186.0425	266.2947	364.3683	471.3865	584.4355	719.6047	866.0039	1039.0264	1235.4822	1439.8896

Table 2. Multi-Layer Perceptron Metrics

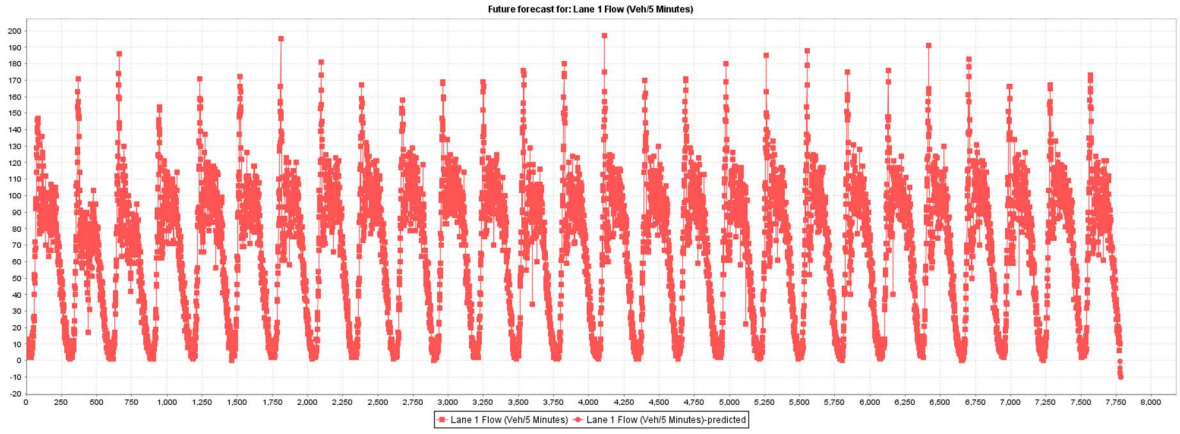


Fig. 5. Future forecast for target using multi-mayer perceptron

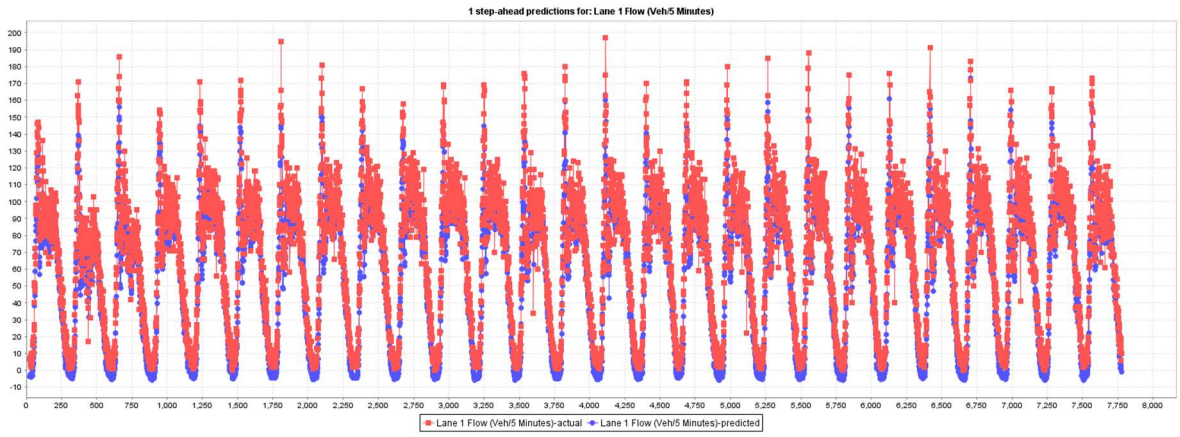


Fig. 6.Step-ahead predictions for target using multi-layer perceptron

c) RBF Network

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	20.5881	20.858	21.1335	21.4303	21.7275	22.0565	22.3605	22.6948	23.0372	23.3828
RAE	244.6782	247.8889	228.8866	209.7128	194.2027	101.4967	170.3026	162.7135	144.6749	147.673
RMSE	27.1056	27.5907	28.1361	28.7435	29.3172	29.9695	30.5983	31.2313	31.9144	32.5965
MSE	734.7131	761.2447	791.6411	826.1909	859.4997	858.1735	936.2555	975.2555	1018.5315	1062.5295

Table 3. RBF metrics

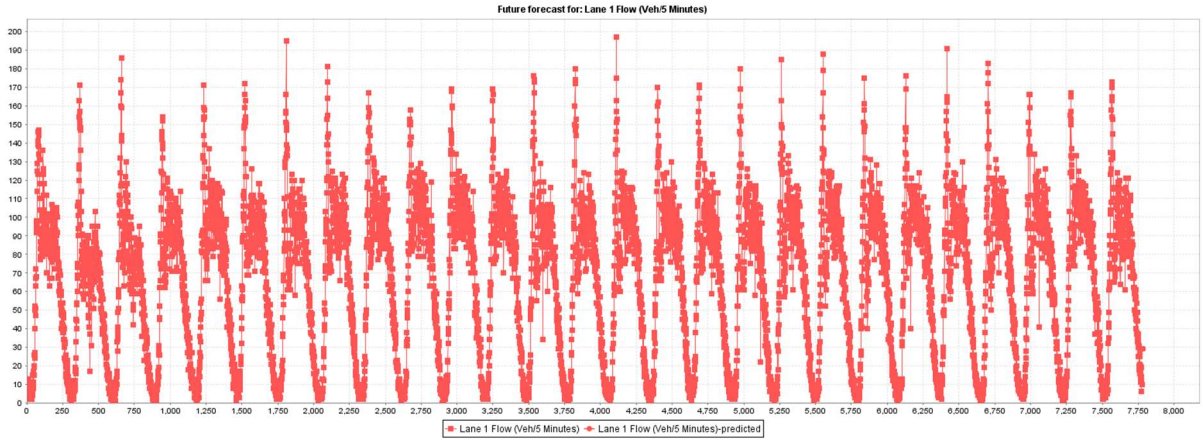


Fig. 7. Future forecast for target using RBF network

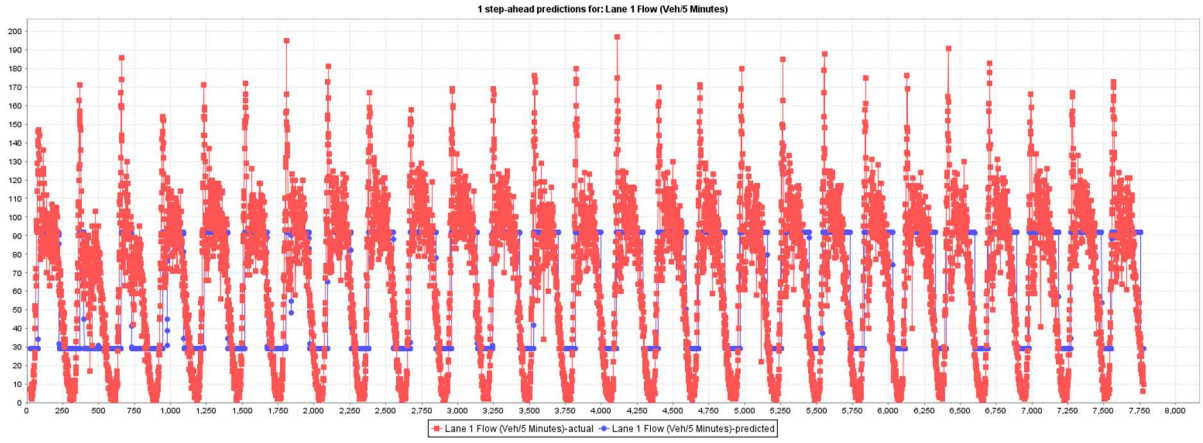


Fig.8. 1-step-ahead predictions for target using RBF network

d) RBF Regressor

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.2196	7.8062	8.4618	9.0304	10.0404	10.0404	10.4784	10.9455	11.4588	11.9374
RAE	85.8086	92.7736	91.6459	88.3695	82.2457	82.6199	79.8059	78.4751	77.4334	75.3904
RMSE	9.7249	10.5025	11.3025	11.9471	12.5349	13.0998	13.5589	14.6372	14.6372	15.2118
MSE	94.5731	110.3029	127.7474	142.7326	157.1247	171.6049	183.8441	197.988	214.248	231.1992

Table 4. RBF regressor metrics

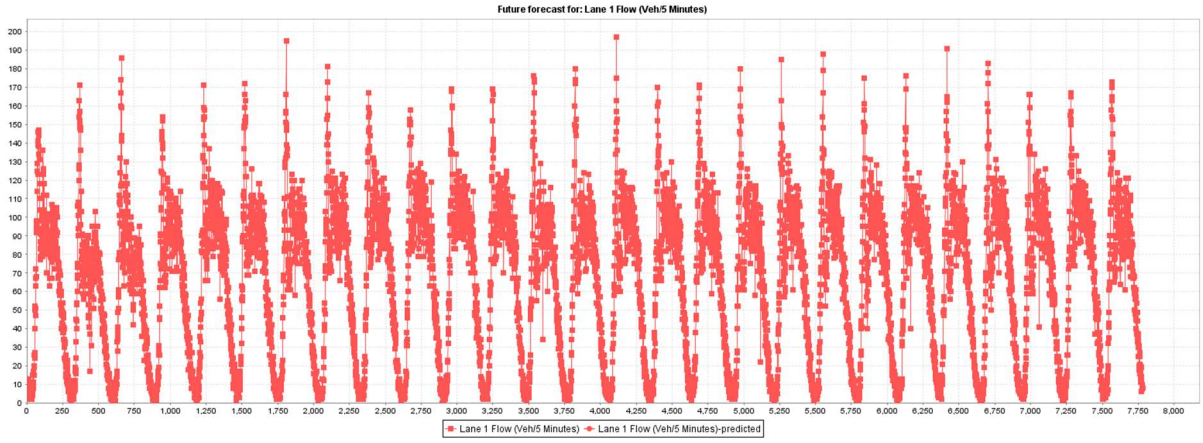


Fig.9. Future forecast for target using RBF regressor

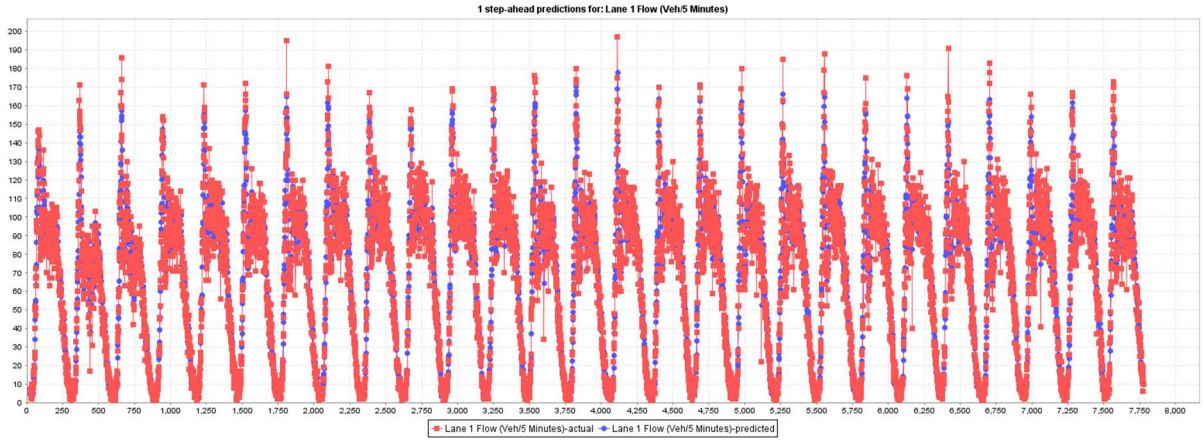


Fig.10. 1-step-ahead predictions for target using RBF regressor

e) SMO Reg

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.5506	8.4396	9.472	10.4654	11.4144	12.3323	13.1224	14.0219	14.9454	15.7413
RAE	89.7457	100.3014	102.5869	102.4128	102.0229	101.4788	99.9429	100.9941	100.9941	99.4133
RMSE	10.3603	11.7133	13.2317	14.6998	16.155	17.6047	18.9205	21.6009	21.6009	22.7999
MSE	107.3354	137.2019	174.079	216.0831	260.9834	309.9269	357.9841	466.9046	466.5988	519.8339

Table 5. SMO reg metrics

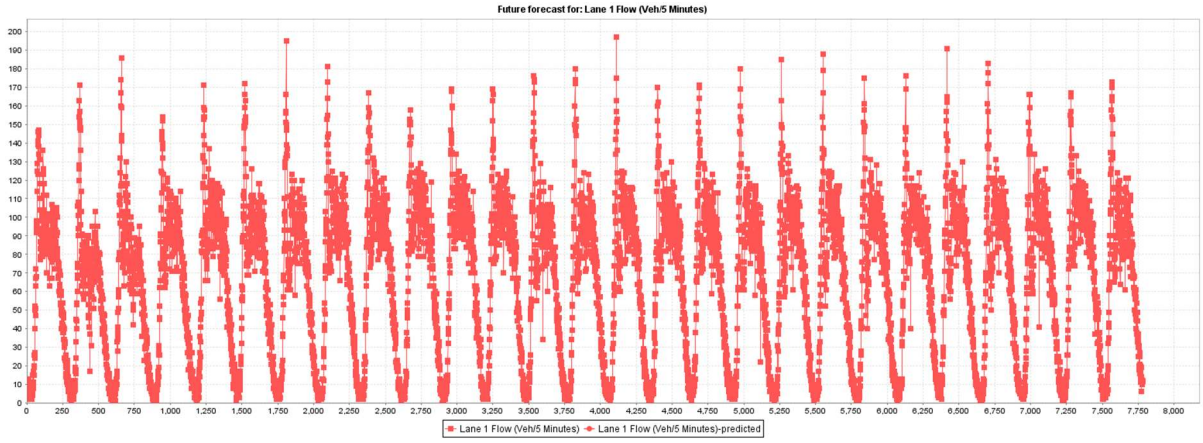


Fig.11. Future forecast for target using SMO reg

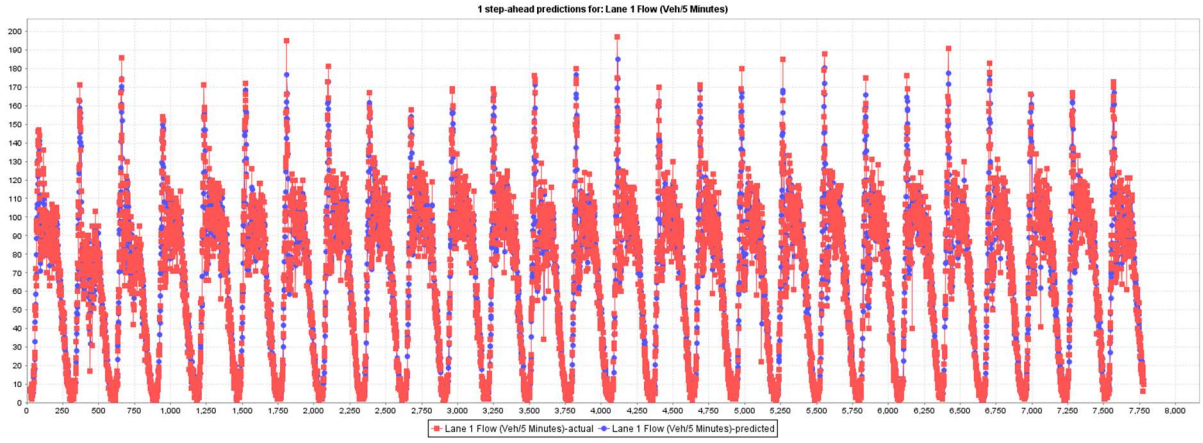
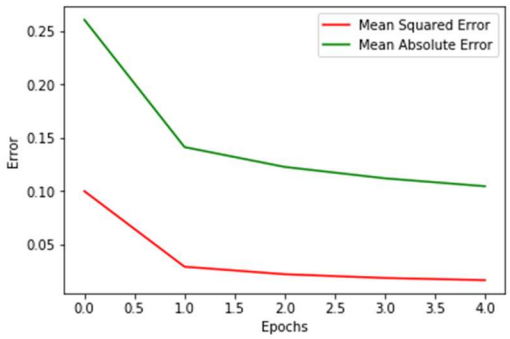


Fig.12. 1-step-ahead predictions for target using SMO reg

f) Deep Learning Model (LSTM)

Case 1: 5 Epochs



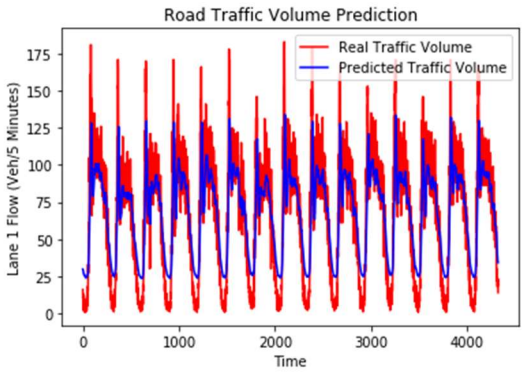


Fig.13. MSE vs MAE with 5 Epochs

Fig.14. Real Traffic Vs Predicted Traffic with 5 Epochs

Case 2: 10 Epochs

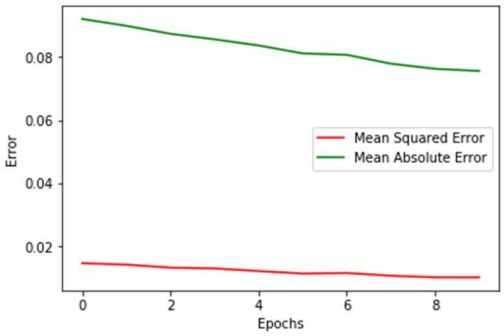


Fig.15. MSE vs MAE with 10 Epochs

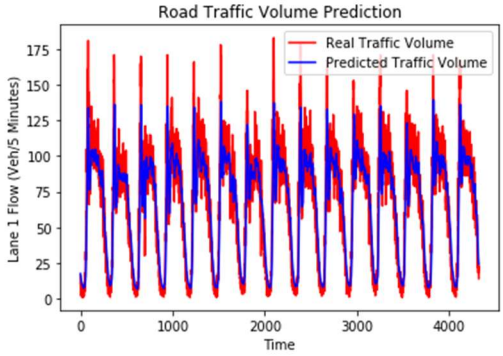


Fig.16. Real Traffic Vs Predicted Traffic with 10 Epochs

Case 3: 25 Epochs

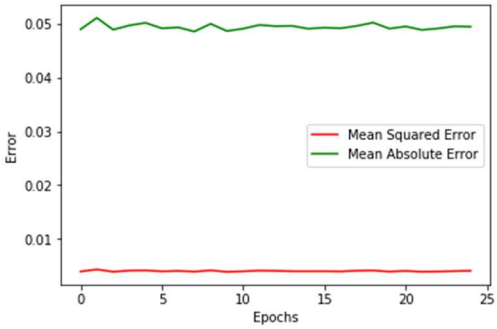


Fig. 17. MSE vs MAE with 25 Epochs

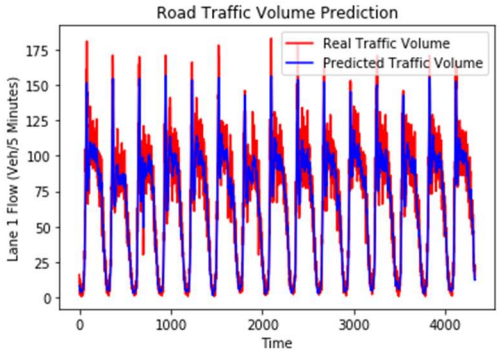


Fig.18. Real Traffic Vs Predicted Traffic with 25 Epochs

Case 4: 50 Epochs

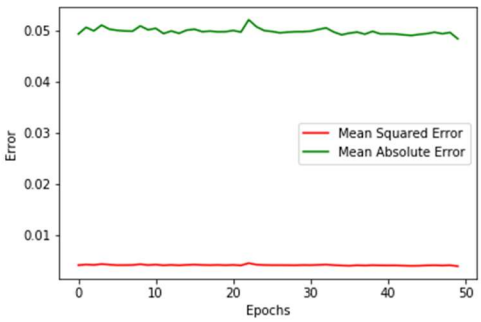


Fig. 19. MSE vs MAE with 50 Epochs

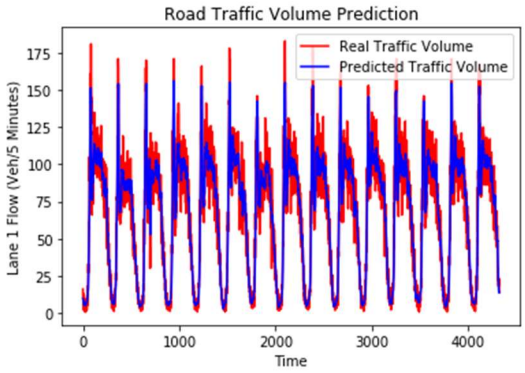
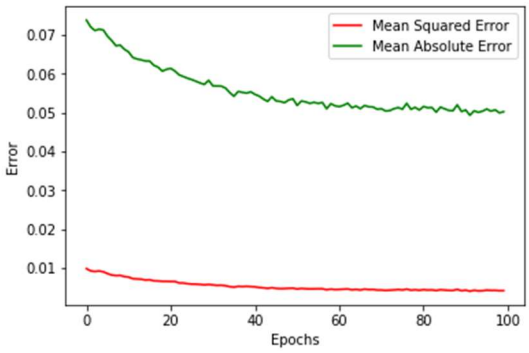


Fig. 20. Real Traffic Vs Predicted Traffic with 50 Epochs

Case 5: 100 Epochs



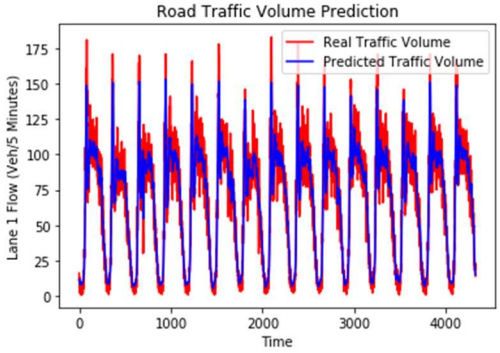


Fig. 21. MSE vs MAE with 100 Epochs

Fig. 22. Real Traffic Vs Predicted Traffic with 100 Epochs

VI. Conclusions

This study conducts experiments in WEKA using a dataset from the Caltrans Performance Measurement System (PeMS) containing vehicle traffic volume across all major metropolitan areas of California to analyze historical traffic volume and forecast traffic volume for a given day. This is a regression problem, so experiments are conducted using 5 popular regression algorithms including Linear Regression, Multi-Layer Perceptron, RBF Network, RBF Regressor, and SMO Reg algorithms. For these 5 algorithms, regression metrics are tabulated in Tables 1 to 5. Figs. 3 to 12 forecast the target variable and the future forecast of 1-step ahead data for all 5 regression algorithms. The results suggest that RBF Regressor provides the best results over linear regression, followed by SMO Regression, Linear Regression, RBF Network and MLP Network. The MLP network is not suitable. Experiments with LSTM deep learning model are conducted using the same dataset to compare LSTM with RBF Regressor. In LSTM model, results for 5 different cases are obtained for increasing numbers of epochs. Figs. 13 to 22 plotmetrics and predict real versus predicted traffic for various epochs.The results show that performance metrics improve with increasesin the number of epochs. MSE and MAE for the LSTM model is far superior to regression models.

Acknowledgements

The authors acknowledge Weiqi Wang of St. Francis Prep and Yuchen Ji of Bodwell High Schoolfor participating in this research.

REFERENCES

[1] Dai, X., Fu, R., Lin, Y., Li, L., & Wang, F. Y. (2017). DeepTrend: A deep hierarchical neural network for traffic flow prediction. *arXiv preprint arXiv:1707.03213*.

[2] Chen, W., An, J., Li, R., Fu, L., Xie, G., Bhuiyan, M. Z. A., & Li, K. (2018). A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features. *Future Generation Computer Systems*, *89*, 78-88.

[3] Manornjitham, Raj. P, Lal, H.K (2018). A Survey of Road Traffic Prediction with deep learning. *International Journal of Pure and Applied Mathematics,* 2065-2073.

[4] Jia, Y., Wu, J., & Xu, M. (2017). Traffic flow prediction with rainfall impact using a deep learning method. *Journal of advanced transportation*, *2017*.

[5] Yang, B., Sun, S., Li, J., Lin, X., & Tian, Y. (2019). Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing*, *332*, 320-327.

[6] S.K.Groen. (2012), A Review of Traffic growth rate calculations. *Shaping the future: Linking policy, research and outcomes, 25th ARRB Conference, Perth, Australia,* 1-15.

[7] Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1-17.

[8] Du, S., Li, T., Gong, X., Yu, Z., Huang, Y., &Horng, S. J. (2018). A hybrid method for traffic flow forecasting using multimodal deep learning. *arXiv preprint arXiv:1803.02099*.

[9]Lv, Z., Xu, J., Zheng, K., Yin, H., Zhao, P., & Zhou, X. (2018, January). LC-RNN: A Deep Learning Model for Traffic Speed Prediction. In *IJCAI* (pp. 3470-3476).

[10] Wang, J., Chen, R., & He, Z. (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies*, *100*, 372-385.

[11] Zhang, S., Kang, Z., Hong, Z., Zhang, Z., Wang, C., & Li, J. (2018, July). Traffic flow prediction based on cascaded artificial neural network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*(pp. 7232-7235). IEEE.

[12] Xiao, Y., & Yin, Y. (2019). Hybrid LSTM Neural Network for Short-Term Traffic Flow Prediction. *Information*, *10*(3), 105.

[13] Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B. (2018). LSTM-based traffic flow prediction with missing data. *Neurocomputing*, *318*, 297-305.