Traffic Volume Prediction with Automated Signal Performance Measures (ATSPM) Data

Leah Kazenmayer[†], Gabriela Ford[‡], Jiechao Zhang[‡], Rezaur Rahman[‡], Furkan Cimen[§], Damla Turgut[§], and Samiul Hasan[‡]
[†]Department of Computer Science, The College of New Jersey
[‡]Department of Civil, Environmental, and Construction Engineering, University of Central Florida
[§]Department of Computer Science, University of Central Florida
kazenml1@tcnj.edu, {gabrielaford1512, zlyjsl123, rezaur.rahman, furkan}@knights.ucf.edu
turgut@cs.ucf.edu, Samiul.Hasan@ucf.edu

***Abstract*—Predicting short-term traffic volume is essential to improve transportation systems management and operations (TSMO) and the overall efficiency of traffic networks. The real-time data, collected from Internet of Things (IoT) devices, can be used to predict traffic volume. More specifically, the Automated Traffic Signal Performance Measures (ATSPM) data contain high-fidelity traffic data at multiple intersections and can reveal the spatio-temporal patterns of traffic volume for each signal. In this study, we have developed a machine learning-based approach using the data collected from ATSPM sensors of a corridor in Orlando, FL to predict future hourly traffic. The hourly predictions are calculated based on the previous six hours volume seen at the selected intersections. Additional factors that play an important role in traffic fluctuations include peak hours, day of the week, holidays, among others. Multiple machine learning models are applied to the dataset to determine the model with the best performance. Random Forest, XGBoost, and LSTM models show the best performance in predicting hourly traffic volumes.**

***Index Terms*—IoT, Traffic Signal, Arterial, ATSPM, Machine Learning, Traffic State Prediction**

I. Introduction

Traffic congestion has been an increasing problem seen in many cities in the world. Florida cities, in particular, are facing high population and economic growth leading to challenges such as increased traffic congestion on small roads, traffic accidents, excessive fuel consumption, among others. To overcome these issues, reliable traffic prediction methods are needed. Traffic prediction has been beneficial to adapt city planning strategies and manage traffic. Knowing how traffic volume changes over time can help to avoid congestion, delays in a particular corridor or intersection [1], [2]. In order to achieve future traffic prediction, traffic behavior data is collected and analyzed. The analysis of real-world traffic data has been beneficial to increase the accuracy in models that help improve traffic networks, save commuters' time, and decrease environmental pollution due to increased car emissions.

Automated Traffic Signal Performance Measures (ATSPM) [3] is a high-resolution data-logging capability added to existing traffic signal infrastructure, Internet of Things (IoT) devices, and data analysis techniques. It provides transportation agencies the information needed to identify and correct deficiencies in traffic signals [4]. ATSPM is an enabling technology that leverages data collection and analysis for proactive traffic signal system management. It reports information about signals such as traffic lights changes, pedestrian walk signals, vehicle passing, and so on. The analysis of the volume of cars dataset, collected by these devices, will be conducted by multiple machine learning models that will learn how the non-linear patterns of traffic behave.

This study is based on the ATSPM dataset, acquired from Seminole County, Florida. The SR-426 corridor, including nine signals, was chosen to analyze traffic patterns (Figure 1). While predicting hourly traffic, in addition to taking the last 6 hours of traffic for a particular signal, other factors taken into consideration included the day of the week, time of the day, holidays, hurricanes, and precipitation. We used multiple machine learning models such as multiple linear regression, KNN, Decision Tree, Random Forest, XGBoost, and LSTM in our evaluation study.

The contributions of this study include: (i) it analyzes real-world ATSPM data to predict intersection-level traffic volume in short term; no study has previously analyzed such IoT based signal performance data for predictive purposes; (ii) it develops a data-driven approach for prediction through a rigorous testing (i.e., the trade-off between interpretability and accuracy) of multiple machine learning and deep learning models; and (iii) it provides valuable insights on the performance of different models for intersection level short-term volume prediction using ATSPM data.

II. Related Work

Over the past decade, we have seen advancements in different components of intelligent transportation systems such as adaptive traffic signal control, automated ramp metering, etc. Such advanced technologies largely depend on real-time monitoring and prediction of traffic for a short-term period [5]. That is why short-term traffic prediction has been a growing necessity and researchers are exploring different

978-1-6654-9792-3/22/$31.00 ©2022 IEEE

approaches to improve the accuracy of such prediction models [6]. Existing methods to solve short-term traffic prediction problems can be broadly classified into three groups: mathematical, simulation-based, and data-driven methods [1], [2]. Although several early studies have shown the potential of mathematical models and simulation-based approaches to produce solutions of traffic prediction problems [7], [8], such approaches rely on many assumptions to model traffic flow behaviors. Consequently, with increasing complexity and computation time for traffic prediction, these approaches become less suitable for a real-time application.

In recent years, data-driven approaches have emerged as an alternative solution to overcome the limitations of traditional models for real-time traffic prediction applications [2]. Data-driven approaches can be classified into different ways, such as parametric and non-parametric models, time series, neural networks (NNs), parametric regression (ARIMA, Kalman filter), non-parametric regression, and so on [9]. Some commonly used data-driven approaches include support vector machine (SVM) [10], k-nearest neighbor (KNN) [11], artificial neural network (ANN) [12], random forest (RF) [13], and ARIMA [14]. Although these data-driven models perform reasonably well for most of the traffic prediction problems (i.e., speed, volume, travel time), their performances deteriorate with the increase in non-linearity in traffic patterns from unexpected events such as crashes or other incidents leading to a sudden drop in traffic flows [15]. To overcome these challenges, neural computation based deep learning models [16], [17] were introduced in traffic prediction.

However, one of the issues with such complex models is that for a simple traffic prediction problem, they tend to overfit the data. Moreover, these models use a high number of parameters, resulting in models being less interpretable and more difficult to select appropriate hyper-parameters. Thus one of the main challenges in developing a data-driven model is to select an appropriate model for a specific task. Existing applications of these data-driven models mostly involve predicting traffic stats such as speed, volume, and travel time using data from roadway detectors (i.e., freeways and arterials) such as microwave radar detectors, loop detectors, and so on [17], [18]. Few studies have explored adopting a data-driven approach for solving intersection-level short-term traffic volume prediction problem [19], [20].

III. Methodology

A. Data Collection

High-fidelity data collection at intersections has created an opportunity to deal with more complex problems in transportation. The proper analysis of this data can reveal important information about real-world traffic flow and allow for possible traffic forecasting in networks at intersection levels. There are multiple sources from which traffic movement data can be collected, stored, and accessed. According to the selected study site of Seminole County, most of the signalized intersections are equipped with advanced traffic

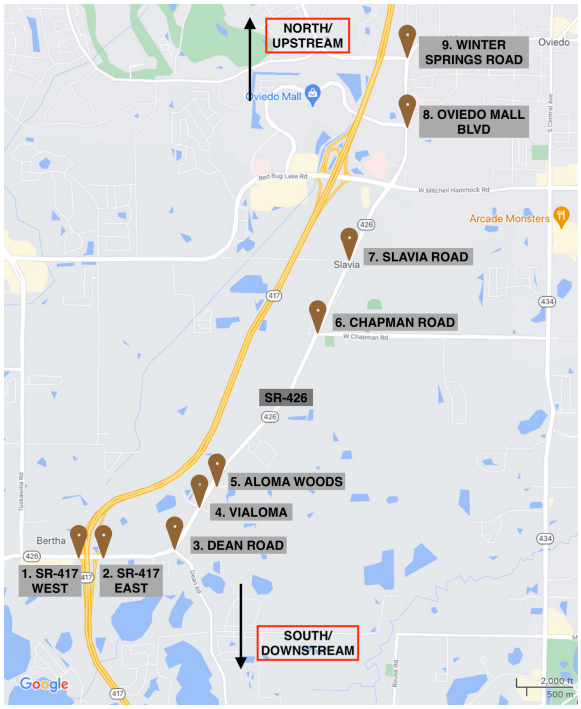

Fig. 1. Selected Study Corridor.

signal controllers on the arterials, and each signal provides Automated Traffic Signal Performance Measures (ATSPM). ATSPM is an enabling technology that contains real-time and historical performance at signalized intersections. This technology encodes events that occur on traffic signal controllers with high-resolution data loggers; the time resolution is to the nearest 100 milliseconds [21]. The ATSPM data was collected for a selected corridor in Seminole County, nine signals in State Road 426, for 2016 and 2019 (Fig. 1). The data collected would allow us to calculate traffic volume at each intersection with the straight/through movement and create a machine/deep learning model to forecast future traffic volume at selected intersections based on previous hours. All events related to left-turn and right-turn were not considered for traffic volume at an intersection due to all signals not having data available for all movement types and directions in the ATSPM dataset. Straight/through traffic movements are present for all the signals in the selected corridor.

B. Data Processing

To calculate traffic volume at each intersection within the selected corridor before creating the models, the data was first filtered to remove any irregularities seen, such as detector malfunctioning, false encoding during storing the data into the server, duplicate entries, bad weather conditions, etc. To validate the quality of ATSPM data from selected signalized intersections, we compared the ATSPM data to the one-day Turning Movement Counts (TMCs) available from the Florida Department of Transportation (FDOT). Since the TMCs from the SR426 & Dean Rd are not available, we compared the

other 8 signalized intersections. The TMCs were manually counted at different intersections at peak hours, which can be the ground truth of traffic volume. Compared to TMCs, we can collect ATSPM data in a more efficient way and on a larger scale. To evaluate the performance, we applied the GEH statistic as an evaluation metric. The formula of GEH statistic is given below:

$$GEH = \sqrt{\frac{2(M-C)^2}{M+C}} \tag{1}$$

where M is hourly traffic volume estimated from ATSPM data and C represents the Turning Movement Count. For evaluating traffic models, compared to the "baseline" scenario, a GEH of less than 5.0 is considered a good match between the modeled and observed volumes. The GEH value was less than 5 for 85% of the data points, which matches standard values recommended [22]. From the GEH statistic, all scores above 10 were removed for each volume of traffic because it is counted as nonreliable data points.

After cleaning the dataset, the hourly volume was calculated for each selected intersection by filtering out all other movements included in the dataset. In this study, only the through northbound or upstream movement in the corridor was considered. It is important to highlight that non-motorized traffic was not taken into account and that only selected signals have ATSPM data being collected.

C. Factor Selection

From a preliminary analysis of the traffic patterns seen in the hourly traffic for the intersections, it became evident that multiple factors play roles in the fluctuation of traffic flows. A spatio-temporal analysis was made for potential factors that affected the volume per hour in the selected location. The factors seen to play an important role in the drastic fluctuation of traffic volumes per hour included: day of the week, the hour of the day, holidays, and occurrences of hurricanes and/or precipitation. The definite selection was made for when a clear difference was seen in the volume of cars per hour for each factor in each signal (Fig. 2).

- Day of the week: weekday or weekend
- Peak hours: 5:00 AM to 10:00 AM and 3:00 PM to 7:00 PM
- Holidays: Easter, Memorial Day, Independence Day, Labor Day, Halloween, Veteran's Day, Thanksgiving, Black Friday, Christmas, New Year's Eve, and New Year's Day
- Hurricanes: Hurricane Matthew and Harmine (2016) and hurricane Dorian (2019)
- Precipitation: collected from the closest weather station from the corridor (Sanford International Airport). Precipitation events higher than 30 mm/hr is considered as major precipitation event for the analysis

D. Model Exploration

Once all the hourly volumes were aggregated for the signals selected for the years 2016 and 2019 and their respective

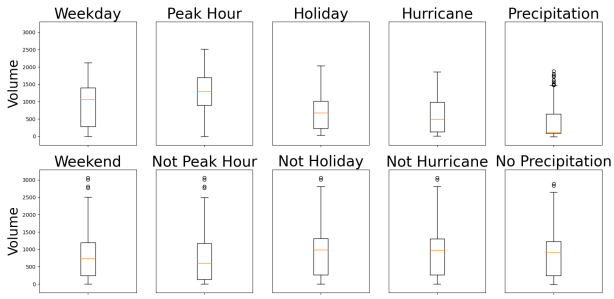

Fig. 2. Factors affecting traffic - signal 1.

factors for each hour, a predictive model was built to predict the hourly volume for each signal based on the previous 6 hours of traffic volume and factors. Multiple algorithms were used to predict the hourly traffic. Since traffic contains nonlinear characteristics, multiple types of models were tested to see which one demonstrated the best fit to the dataset. First, simple models were applied: Multiple Linear Regression and Decision Tree. From these, more complete and complex models were applied to the dataset: Random Forest, KNN, and Extreme Gradient Boost (XGBoost). Finally, because of the constant growth of deep learning models in traffic applications due to their success in learning patterns, the Long Short Term Memory (LSTM), a highly effective deep learning model in predicting time series was also applied.

For model training for volume prediction of each signal selected in the corridor, the data was split randomly into training and testing sets such that; the training data consisted of randomly selected 75% of the dataset and the testing data consisted of the remaining 25%. After the data is itemized as 6 hours of consecutive time series, they are shuffled for better training. Each model attempted to learn and recognize traffic patterns from the past 6 hours of traffic volume and given factors and predicted the next hour's traffic volume. The testing data was used to evaluate the prediction performance of trained models by comparing them with the actualized values of traffic volume.

E. Hyperparameter Tuning

For all the machine learning models, the default embedded hyper-parameters in each model were used and some parameters were changed to find the optimal conditions in which the hourly volume of traffic was predicted the best. The best performing parameters that were changed for predicting the hourly volume for each model can be found in Table I. The grid search method is applied to determine the optimal parameters for Multiple Linear Regression, Decision Tree, Random Forest, KNN, and XGBoost. Exhaustive search is applied for the most relevant parameters for each algorithm and 5-fold cross-validation across training data is incorporated to prevent over-fitting of the parameters. For the LSTM model, experimentation and intuition are used to select parameters, as the search space of hyper-parameters for the deep learning models is long.

TABLE I
OPTIMAL HYPERPARAMETERS DISCOVERED FOR MACHINE AND DEEP LEARNING MODELS

Model	**Optimal Parameters**
Linear Regression	normalize = True
Decision Tree	criterion = 'mae' max_depth = 7
Random Forest	max_depth = Auto selected w.r.t min_sample_splits min_samples_split = 8 n_estimators = 99
KNN	algorithm = 'auto' weights = 'distance' n_neighbors = 17
XGBoost	booster = 'gbtree' max_depth = 12 learning_rate = 0.3 min_split_loss = 0 n_estimators = 100
LSTM	number of LSTM hidden layers = 1 number of nodes per layer = 15 loss = 'mse', optimizer = 'adam' epochs = 120, batch_size = 72

IV. PERFORMANCE EVALUATION

A. Performance Measures

To determine which of the six models performed the best, we compared their Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) scores. MAE is the performance metric that checks the accuracy of the implemented model. Absolute error is the difference between the predicted values and the "true" value; the MAE is the average of all absolute errors. The MAE equation is as follows where $\hat{y}_i$ is the prediction and y_i is the true value:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{2}$$

RMSE is defined as the standard deviation of the residuals as it indicates how concentrated the data is around the line of best fit. In general, the smaller the RMSE value is, the better the model. The RMSE equation is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{3}$$

The coefficient of determination or R-squared (R^2) is defined as the comparison of the residual sum of squares with the total sum of squares. It represents the goodness of fit of a regression model. The closer the value of R-squared to 1, 1 representing the perfect predictor, the better the model. The R^2 equation is as follows:

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{4}$$

Fig. 3. Comparison of MAE values for each model built for the nine signals.

B. Model Applications

The five machine learning and one deep learning models were trained with the gathered hourly volumes of traffic and corresponding factors such as peak hour, precipitation, and so on for 2016 and 2019. Each model performed differently. As seen in (Fig. 3) and (Fig. 4), the MAE and RMSE values were displayed and compared for all nine signals. To put the MAE and RMSE values in perspective, the lowest number of vehicles per hour is 0, and the highest number is between two and three thousand. Since the MAE and RMSE values range anywhere between 40 and 160, all the models performed well except the linear regression model. The linear regression model distinctively performs worse than other models (Fig. 3) and is not a good predictive model for the dataset as the traffic data is highly non-linear. MAE values for all the models can be seen in Fig. 3. As seen in Fig. 4, multiple models performed better than others: Random forest and XGBoost. A similar pattern is observed in Fig. 5, where R^2 values are compared. Looking at the ranges of R^2 values, all of the models but linear regression performed exceptionally. Again, Random Forest, XGBoost, and LSTM have performed consistently superior compared to the other three models. These results are consistent with our expectations as the aforementioned three models are more suitable for modeling highly non-linear data in which the traffic volume falls.

Taking a closer look at randomly chosen signals 1, 5, and 8, actual vs. predicted graphs are plotted and it can be seen visually how accurately Random Forest, XGBoost, and LSTM predicted the hourly vehicle volume in Fig. 6. As one can see, the more concentrated the data points around the $y = x$ line (the red line), means the better the model predicted hourly traffic volume. In all three models, their data points are all located tightly around the red line.

Compared to the best performing three models, Multiple Linear Regression, KNN, and Decision Tree did not perform well; this suggests that those models took a simpler approach. For multiple linear regression, it performs the worst; this is

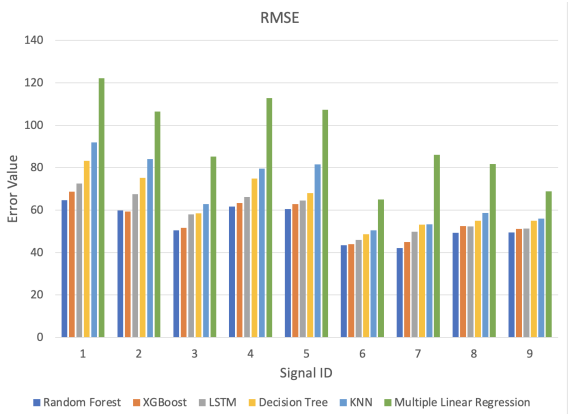

Fig. 4. Comparison of RMSE values for each model built for the nine signals.

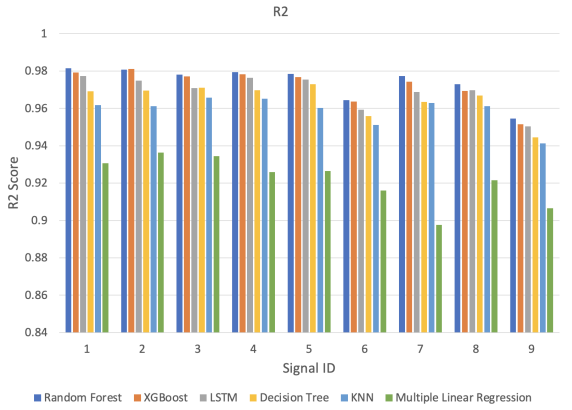

Fig. 5. Comparison of R^2 values for each model built for the nine signals.

because this type of model establishes a linear relationship between the independent variables and the dependent variable, but the underlying patterns lean more towards a non-linear relationship. KNN uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This algorithm is one of the simplest algorithms in general, but it does not recognize or learn traffic patterns. Lastly, the Decision Tree algorithm is a great algorithm to use, but Random Forest and XGBoost utilize multiple decision trees rather than just one, taking it one step further. Although LSTM is highly capable to learn non-linear patterns and achieves a high R^2 score, it is not the best performing model. One of the advantages of LSTM is to capture patterns in long time series of data so the last 6 hours of data might not be enough the realize the power of LSTM. Another requirement for best performance for any deep learning model is a large amount of data and even though two years of data can be considered large enough for other machine learning models, deep learning models are data-hungry and potentially improve with more data.

V. Conclusion

In traffic control systems, accurate volume prediction is a critical component to ensure efficient traffic operations at an intersection and/or corridor [1]. Traffic models have been consistently developed to predict volume with real data and it plays major roles in planning and management techniques in traffic controls [5]. Machine learning techniques have become more popular in this field, their demonstrated ability to capture sharp discontinuities in traffic flows using nonlinear functions (e.g., tanh, sigmoid) or yes/no decision mechanism [18]. These models have displayed their ability to learn the task of predicting traffic volume from past data, which is especially important and difficult due to the high complexity and dimensionality of the traffic patterns [15]. More complex machine learning models, as well as deep learning models, were demonstrated to capture the non-linearity of traffic data. Models such as random forest, XGBoost and LSTM outperformed other models when calculating the volume at an intersection. Other models such as KNN, decision tree, and multiple linear regression, are simpler models that although had a prediction potential, did not predict the traffic volume as well as more complex models.

It was proven possible to predict the hourly volume of traffic with the best three models selected based on a 6-hour pattern of traffic volume attached to some critical factors with these models applied. These factors made a substantial impact on the analysis and prediction of traffic volume. Each factor was seen as significant in the fluctuation of traffic volume, and it helped the machine learning models to account for the sharp changes seen in the data set. These numerous factors show the higher dimensionality of the traffic volume pattern.

The predictions were possible due to big data sets from IoT devices in real traffic signal systems. Correctly analyzing data sets was essential in improving the current system and getting ready for future city planning and management.

VI. Acknowledgements

The support for this work was provided by the National Science Foundation REU program under Award No. 1852002. The ATSPM data were provided by the Florida Dept. of Transportation's District 5 office.

References

[1] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transport reviews*, vol. 24, no. 5, pp. 533–557, 2004.

[2] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[3] C. M. Day, D. M. Bullock, H. Li, S. M. Remias, A. M. Hainen, R. S. Freije, A. L. Stevens, J. R. Sturdevant, and T. M. Brennan, "Performance measures for traffic signal systems: An outcome-oriented approach," Purdue University, West Lafayette, Indiana, Tech. Rep., 2014.

[4] Charles R. Lattimer, "Automated Traffic Signal Performance Measures Installation Manual," Federal Highway Administration, Tech. Rep., 2016. [Online]. Available: https://ops.fhwa.dot.gov/publications/fhwahop20002/fhwahop20002.pdf

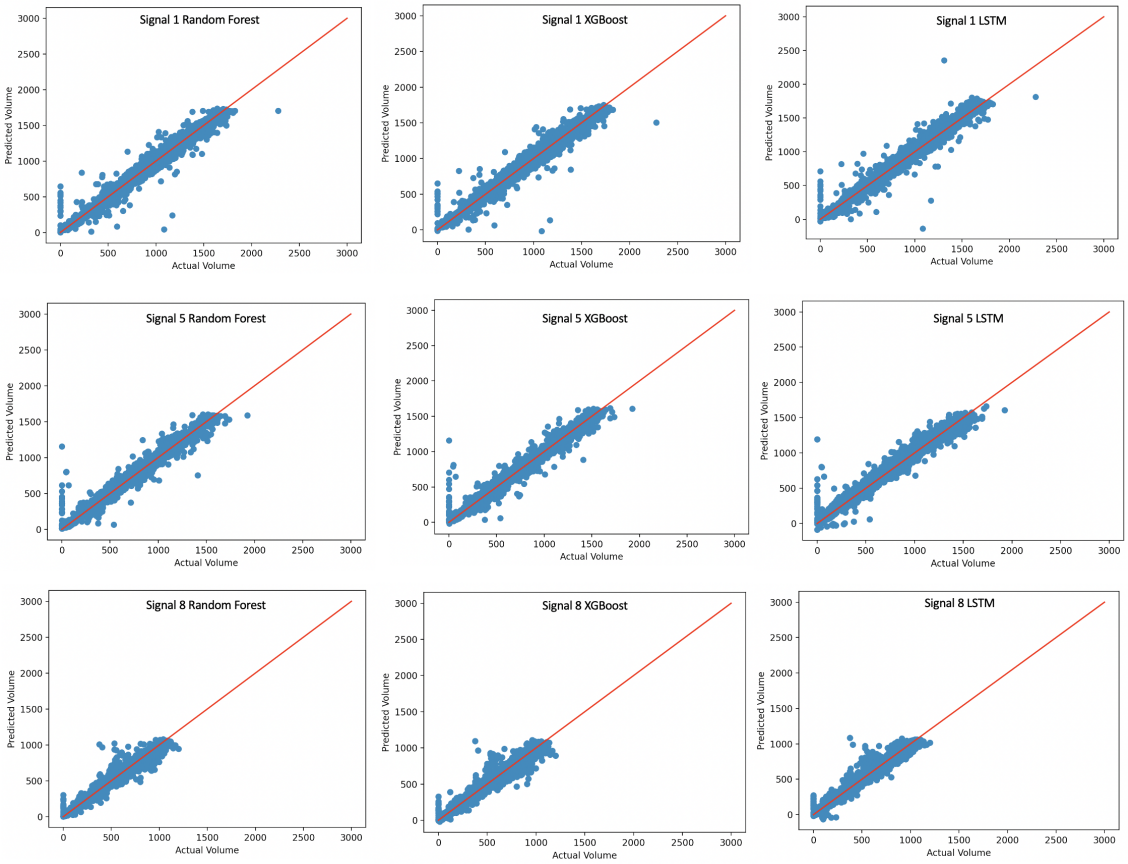

Fig. 6. Actual vs. predicted volume for the best performing models for Signals 1, 5, and 8 for northbound direction.

[5] S. Shukla, K. Balachandran, and V. Sumitha, "A framework for smart transportation using big data," in *IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 2016, pp. 1–3.

[6] Y. Tian and L. Pan, "Predicting short-term traffic flow by long short-term memory recurrent neural network," in *IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, pp. 153–158.

[7] W. Burghout, H. N. Koutsopoulos, and I. Andreasson, "Incident management and traffic information: tools and methods for simulation-based traffic prediction," *Transportation research record*, vol. 2161, no. 1, pp. 20–28, 2010.

[8] Q.-J. Kong, Y. Xu, S. Lin, D. Wen, F. Zhu, and Y. Liu, "Utn-model-based traffic flow prediction for parallel-transportation management systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1541–1547, 2013.

[9] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, "Short-term travel-time prediction on highway: A review on model-based approach," *KSCE Journal of Civil Engineering*, vol. 22, no. 1, pp. 298–310, 2018.

[10] C. Luo, C. Huang, J. Cao, J. Lu, W. Huang, J. Guo, and Y. Wei, "Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm," *Neural processing letters*, vol. 50, no. 3, pp. 2305–2322, 2019.

[11] X. Luo, D. Li, Y. Yang, and S. Zhang, "Spatiotemporal traffic flow prediction with knn and lstm," *Journal of Advanced Transportation*, vol. 2019, 2019.

[12] D. Zeng, J. Xu, J. Gu, L. Liu, and G. Xu, "Short term traffic flow prediction using hybrid arima and ann models," in *Workshop on Power Electronics and Intelligent Transportation System*, 2008, pp. 621–625.

[13] D. Xu and Y. Shi, "A combined model of random forest and multilayer perceptron to forecast expressway traffic flow," in *IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2017, pp. 448–451.

[14] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal arima model with limited input data," *European Transport Research Review*, vol. 7, no. 3, pp. 1–9, 2015.

[15] R. Rahman and S. Hasan, "Short-term traffic speed prediction for freeways during hurricane evacuation: a deep learning approach," pp. 1291–1296, 2018.

[16] ——, "Real-time signal queue length prediction using long short-term memory neural network," *Neural Computing and Applications*, vol. 33, no. 8, pp. 3311–3324, 2020.

[17] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.

[18] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.

[19] W. Alajali, W. Zhou, S. Wen, and Y. Wang, "Intersection traffic prediction using decision tree models," *Symmetry*, vol. 10, no. 9, p. 386, 2018.

[20] N. Mahmoud, M. Abdel-Aty, Q. Cai, and J. Yuan, "Predicting cycle-level traffic movements at signalized intersections using machine learning models," *Transportation research part C: emerging technologies*, vol. 124, p. 102930, 2021.

[21] H. Li, A. M. Hainen, J. R. Sturdevant, T. Atkison, S. Talukder, J. K. Mathew, D. M. Bullock, D. Nelson, D. M. Maas Jr, J. Fink *et al.*, "Indiana traffic signal hi resolution data logger enumerations," 2019.

[22] N. Nezamuddin, N. Jiang, T. Zhang, S. T. Waller, and D. Sun, "Traffic operations and safety benefits of active traffic strategies on txdot freeways," Tech. Rep., 2011.