Fuzzy Systems and Soft Computing
ISSN : 1819-4362

Hybrid Traffic Forecasting Models Leveraging Weather and Sensor Data

Sitanath Biswas,Pratyush Ranjan Mohapatra,Priyabrata Nayak
Dept. of Computer Science and Engineering, Gandhi Institute For Technology, Bhubaneswar, 752054, India
Email: pratyush@gift.edu.in

ABSTRACT

Traffic flow prediction is an integral part of the intelligent transportation system (ITS) that helps in making well-informed decisions. Traffic flow prediction helps in alleviating traffic congestion as well as in some connected vehicles applications such as resources allocation. However, most of the existing models do not consider external factors such as weather data. Traffic flow in road networks is affected by weather conditions which affects the periodicity of traffic. These effects introduce some irregularity to the traffic pattern, making traffic flow prediction a challenging issue. In this paper, we present a detailed investigation on the impact of weather data on different traffic flow prediction models. The investigation presented in this paper demonstrates how adding weather data could improve the models' prediction accuracy and efficiency.

1. INTRODUCTION

In recent years, with the rapid growth in the number of vehicles, the road infrastructure capacity and resources cannot keep up with the rapid increase in demand. This has led to several problems like traffic congestion and road accidents. Therefore, intelligent transportation system (ITS) was introduced to help alleviate these problems. Traffic prediction is an integral part of ITS which contributes to making prediction-based decisions in traffic control management. Therefore, accurate and timely prediction leads to better decision making. In short-term traffic prediction historical and real-time traffic data is used to predict few seconds to few hours into the future. There have been a lot of research work on traffic prediction.

Traffic prediction models can be divided into 3 categories: statistical methods, traditional machine learning methods, and deep learning methods. The statistical methods allowed researchers to capture the regularity of historical data and drive a model to predict future traffic. These methods include historical average (HA) and autoregressive integrated moving average (ARIMA). ARIMA model have been used widely in traffic prediction because of its ability to capture periodic change [1]–[3].

However, traffic data is complex and has nonlinear features and spatio-temporal dependencies which makes statistical methods not suitable for modelling traffic flow accurately. Traditional machine learning methods, like support vector machine (SVM) and support vector regression machine (SVR), have shown great performance in modelling nonlinear data accurately. Traditional machine learning methods use kernel function to capture the inner characteristics of traffic data. These methods show good prediction performance in some traffic prediction applications [4]–[6]. In addition to traditional machine learning methods, basic feedforward neural networks (FFNNs) with backpropagation have been used for traffic prediction in [7], [8]. However, these methods depend on human-engineered features which make them

struggle with the complexity of traffic data. In addition, with recent advancement in road networks and vehicles the amount and complexity of collected traffic information has been increasing rapidly. This huge amount of collected traffic information led to the emergence of traffic big data, which introduces more challenges and complexity for short-term traffic prediction. To address these challenges and complexity, deep learning-based models have been applied to short-term traffic prediction such as recurrent neural network (RNN) or its variants Long short-term memory (LSTM)/Gated recurrent units (GRU) [9]–[12]. Moreover, these models sometimes are combined with convolutional neural networks (CNNs). However, most of the existing short-term traffic prediction methods focuses mainly on modelling traffic data, and little research on external factors like weather conditions. In [13], Hall and Barrow discussed how traffic flow is affected by weather conditions. From the findings of Karlaftis [14] regarding the impact of adverse weather conditions on traffic flow, it appears that adverse weather affects the short-term predictability of lane speed patterns. They also suggest the need for a modeling strategy that can efficiently make use of weather and traffic data to enhance prediction. Recently, some work has been done to use weather data with traffic data. In [15], Hou *et al.* propose a combined framework of stacked autoencoders (SAE) and radial basis function (RBF) neural network; their framework leverages weather data to capture the disturbance of weather factors. Zheng *et al.* [16] proposes a deep learning with embedding approach that can leverage traffic information, route structure, and weather conditions to train a traffic flow prediction model. In [17], Koesdwiady *et al.* studies the correlation between weather parameters and traffic flow and then proposes a deep learning approach that uses deep belief networks. It uses traffic flow with weather conditions to enhance prediction accuracy. In a report made by the Federal highway administration (FHWA), where they evaluated a prediction system that leverages weather data and transportation operations data, it states that the system helped them prioritize and focus attention on particular roadway sections or areas [18]. Bao *et al*. [19] proposed a model that uses deep belief network (DBN) with SVR for traffic prediction, where the model is trained on traffic and weather data collected from Internet of Vehicles (IoV) and shows a prediction error of 9%.

In this paper, we aim to investigate how weather data can affect the accuraces traffic flow prediction models. The main contributions of this paper can be summarized: i) We study the correlation between traffic flow data and weather data using Performance Measurement System (PeMS) dataset and weather information. ii) We study the impact of weather data on three deep learning-based traffic prediction methods LSTM, GRU, CNN-LSTM, and Stacked-LSTM auto encoders [20]–[22]. iii) The results demonstrate how correctly combining weather data with traffic data could improve prediction accuracy.

2. PROBLEM DESCRIPTION

In this research, each time step represents 5 minutes of data where 12 timesteps accumulate to one hour of data. The timesteps represents historical data used for each prediction where 12 timesteps or more are used to predict 1 timestep or more. In our study, we use past 12 consecutive time steps with 5-minute interval to predict one future time step to reduce the complexity of the processed data. We formulated the problem as a supervised learning problem, where the previous time steps are input features $[X_1, X_2, X_3, X_4, \dots, X_{12}]$ and the subsequent time step is the output value y.

$$y = ([X_1, X_2, X_3, X_4, \dots, X_{12}]) \tag{1}$$

$$X_i = \text{concatenate}(x_i^{traffic}, x_i^{weather}) \tag{2}$$

To train the model with traffic and weather data we concatenate both dataset into one to create a timeseries dataset that has traffic flow and weather parameters as features. Where $x^{traffic}$ represents the traffic flow data and $x^{weather}$ represents the weather data. The traffic data and weather data are feed to the prediction model with respect to their timestamps. After adding the weather data, the prediction problem can be expressed as in (3) , where f represents the prediction model, and y represents the output predicted traffic flow in the subsequent time steps.

$$y = ([X_i^{traffic+weather}]) \tag{3}$$

3. DATA PROCESSING

In this research, PeMS traffic dataset is used for training and testing because they are widely used in traffic prediction tasks [23]. PeMS is an abbreviation from the California Transportation Agency PeMS,

which contains 6 months of data recorded by 325 traffic sensors ranging from January 1st, 2017, to June 30th, 2017, in the Bay Area. The weather data is an open-source datasets available in [24]. The location and collection time of weather data corresponds to the traffic data. The data is processed and cleaned to make sure all data points in both datasets corresponds with each other. Weather type is a non-numerical parameter, we ordered the weather type according to the severity of its condition, then we assigned numerical values for the weather types. Finally, we use Pearson correlation coefficient (PCC) to study the correlation between traffic flow and weather parameters. The PCC is calculated using formula in (4). Based on the PCC value irrelevant weather parameters which have the lowest correlation with the traffic flow will be dropped. Table 1 and Table 2 shows a sample of the traffic flow dataset and weather dataset respectively, Table 3 shows the processed weather data after dropping least relevant values.

$$\rho_{X,} = \frac{cov(X,Y)}{\sigma X \sigma Y} \tag{4}$$

Table 1. Traffic flow dataset

5 Minutes	Lane 1 Flow (Vehicle/5 Minutes
2017-01-01 00:00:00	70.5
2017-01-01 00:05:00	70.6
…	…
2017-06-30 23:55:00	68.4

Table 2. Weather dataset

5 Minutes	Humidity (%)	Pressure (mbar)	Temperature (K)	Wind direction (°)	Wind speed (km/h)	Weather type
2017-01-01 00:00:00	81.0	1015.0	284.47	140.0	2.0	moderate rain
2017-01-01 00:05:00	81.0	1015.0	284.47	140.0	2.0	moderate rain
…	…	…	…	…	…	…
2017-06-30 23:55:00	66.0	1016.0	296.80	140.0	3.0	haze

Table 2 shows the weather dataset with six weather parameters, but not all of them impact traffic flow directly. Therefore, parameters with low correlation will be dropped so that they do not cause the model performance to degrade. Figure 1 shows the correlation heatmap between traffic data and weather data; we can see that weather type, humidity, and temperature have the highest correlation values. These three parameters are selected to create a new dataset with weather type encoded ordinally into numerical values.

Figure 1. Correlation heatmap

Table 3. Pre-processed weather dataset

5 Minutes	Humidity (%)	Temperature (K)	Weather type
2017-01-01 00:00:00	**81.0**	**284.47**	**16**
2017-01-01 00:05:00	**81.0**	**284.47**	**16**
…	**…**	**…**	**…**
2017-06-30 23:55:00	**66.0**	**296.80**	**2**

4. PROPOSED METHODOLOGY

To study the impact of weather data on traffic flow prediction we used three prediction models and compared their performance with and without weather data. To capture the features of traffic flow and weather conditions by concatenating both datasets with respect to their timesteps. We feed the data from traffic datasets and weather datasets into the model as sequence of timesteps. First, we train the LSTM model on 3 months' worth of traffic and its corresponding weather data. Then, we test the model using the next month data. During the training process, in each iteration, the model learns how to predict one timestep using the 12 former timesteps. We also trained GRU, CNN-LSTM, and Stacked-LSTM models in the same manners. Figures 2-5 show the architecture of the deep learning models used to perform the experiment.

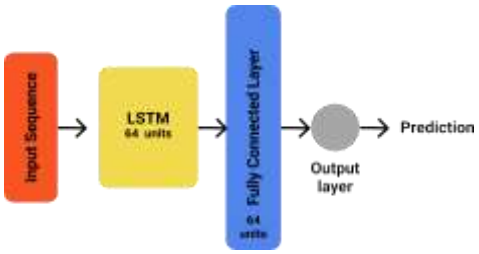


Figure 2. LSTM model structure

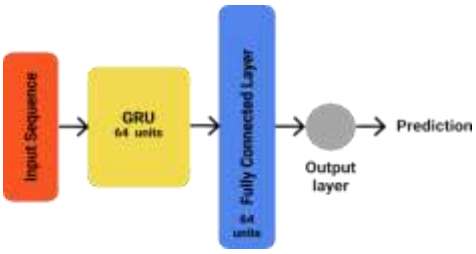


Figure 3. GRU model structure

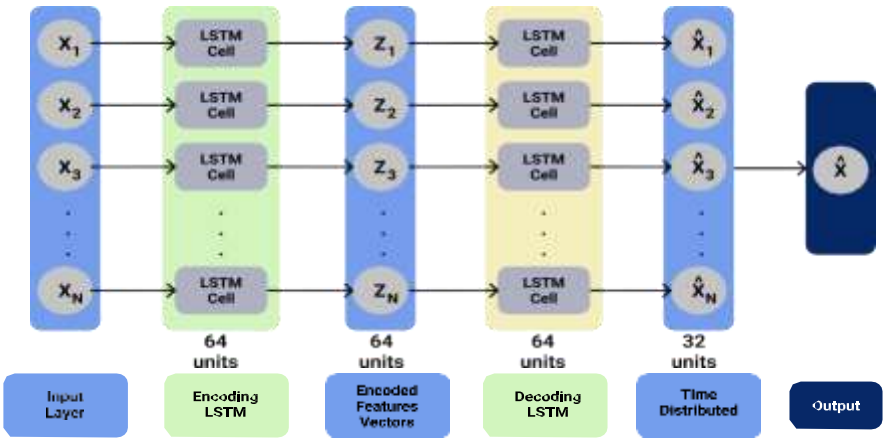


Figure 4. Stacked LSTM architecture

Figure 5. CNN-LSTM architecture

5. EXPERIMENTAL RESULTS

To test the performance of the proposed model, a variety of evaluation matrixes were used in the experimental stage, including mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE). These metrics mainly reflect the distance between the real values and the predicted values. The specific formulas of these metrics are:

MSE:

$$MSE = \frac{1}{n}\sum_{i=1} (y_i - y_i')^2 \quad (5)$$

RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1} (y_i - y_i')^2} \quad (6)$$

MAPE:

$$MAPE = \frac{100\%}{n}\sum_{i=1} \left|\frac{y_i - y_i'}{y_i}\right| \quad (7)$$

Mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1} |y_{:,i} - y'_{:,i}| \quad (8)$$

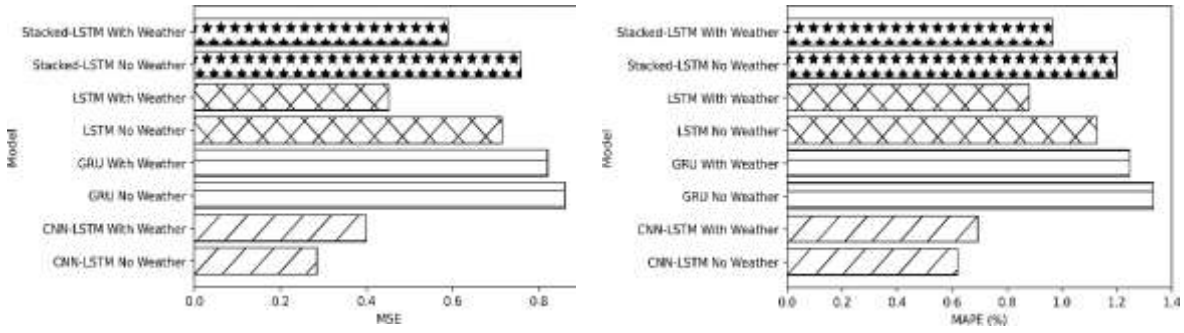
The traffic prediction models were implemented using TensorFlow [25] framework, and the models were trained for 100 epochs, the data was normalized using MinMaxScaler function, the batch size is 128, the function optimizer used is Adam, and the loss function is MSE. Figure 6 shows the performance comparison of different traffic prediction models with and without weather data using four evaluations matrixes. In Figure 6(a), we can see that CNN-LSTM model has the lowest MSE value compared to the other models. Moreover, we notice that GRU model has the highest MSE value which indicates the low performance of the model. GRU compared to the other models has less complex structure, but GRU unlike LSTM exposes the complete hidden content affecting far future predictions. LSTM model shows better performance compared to GRU due its capability to remember longer sequences using a memory unit.

In Figure 6(a), we can see how weather data helped to improve the GRU, LSTM and, Stacked- LSTM models performance. We see that the impact of weather data on each model varies. LSTM model has the highest impact by weather data in term of performance. we notice that the GRU model has the lowest impact by weather data. In contrast to the other models, we can see that CNN-LSTM model performance degraded with weather data. Figures 6(b), 6(c), and 6(d) show the graphs of other Key Performance Indicators (KPIs) to provide a clearer analysis on the impact of weather data on the model performance. Figure 6(b) shows the MAPE value graph of each model which is commonly used and easy to interpret KPI, but MAPE can be skewed with high errors. Figure 6(c) and Figure 6(d) shows MAE and RMSE graphs respectively which provide a more accurate and trusted measure for model performance.

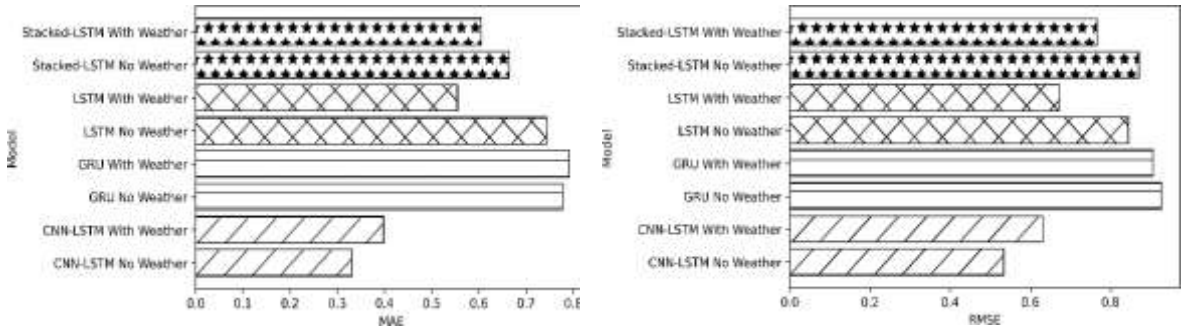
Table 4 shows a summary of the model's performance with and without weather data in term of improvement. We can see that the MAPE value for the GRU model with only traffic data is 1.33% and combining weather data with the traffic data lead to 6.49% reduction in MAPE. However, we can notice that the MSE value increased with weather data. This is because MSE aims for prediction that is correct on average, but since the other KPIs shows improvement, we can assume the model overall performance improved. In addition, we can see that LSTM model has the highest improvement of 21.85% in MAPE value. CNN-LSTM model shows MAPE value of 0.62% which is the lowest compared to the other models without weather data. However, the MAPE value increased to 0.69% with weather data, which indicates that adding weather data degraded the model performance.

Figure 7 shows the traffic flow prediction of 12 hours against the true value. In Figure 7(a), we see the traffic flow prediction of the LSTM model. We notice how the model prediction tends to undershoot the true value, we also notice that adding weather data helps the model to reduce the prediction error and reduce its bias. Figure 7(b) shows the traffic flow prediction of the GRU model, and it tends to overshoot the true value. The weather data also helped the model to be more precise and unbiased. According to Figure 7(c), the traffic flow predictions of the Stacked-LSTM model are more accurate compared to LSTM and GRU, and the weather data does not show a significant impact on the model prediction. Moreover, Figure 7(d) shows the traffic flow prediction of the CNN-LSTM model which has the highest accuracy compared to the other

models. We also notice that adding weather data affected the model accuracy and lead to less accurate results. Figure 7 is shown in Appendix.



(a) (b)



(c) (d)

Figure 6. Performance comparison of different models with weather data (Error Matrix) in terms of (a) MSE, (b) MAPE, (c) MAE, and (d) RMSE

Table 4. Performance of different model with weather data

Model	NO Weather	Weather	Error Reduction (%)
		GRU	
MAPE	1.3335	1.2469	6.4927
MSE	0.7773	0.7911	-1.7748
MAE	0.8609	0.8208	4.6542
RMSE	0.9278	0.9060	2.3548
		LSTM	
MAPE	1.1284	0.8818	21.8597
MSE	0.7444	0.5547	25.4820
MAE	0.7150	0.4527	36.6803
RMSE	0.8456	0.6729	20.4263
		Stacked-LSTM	
MAPE	1.1986	0.9668	19.3394
MSE	0.6646	0.6045	9.0478
MAE	0.7590	0.5887	22.4389
RMSE	0.8712	0.7673	11.9312
		CNN-LSTM	
MAPE	0.6209	0.6941	-11.7928
MSE	0.3310	0.3979	-20.2168
MAE	0.2849	0.3985	-39.8897
RMSE	0.5337	0.6313	-18.2750

6. CONCLUSION

In this paper, we studied the impact of weather data on traffic prediction. The experiments performed showed how the model accuracy increases when weather data is combined with traffic data. This increase in models' performance is due to the correlation between weather conditions and traffic flow. The weather data was not ready to be combined with traffic data due some non-numeric values and missing

values. In this experiment, lots of work has been done for data processing to prepare the weather data to be used with traffic data, where every data point in weather data need to have its corresponding values in traffic data. The data has been processed to make it useful for traffic prediction using deep learning models.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Higher Education, Malaysia FRGS/1/2019/TK08/MMU/03/1 and TMRND Grant MMUE/190012.

APPENDIX

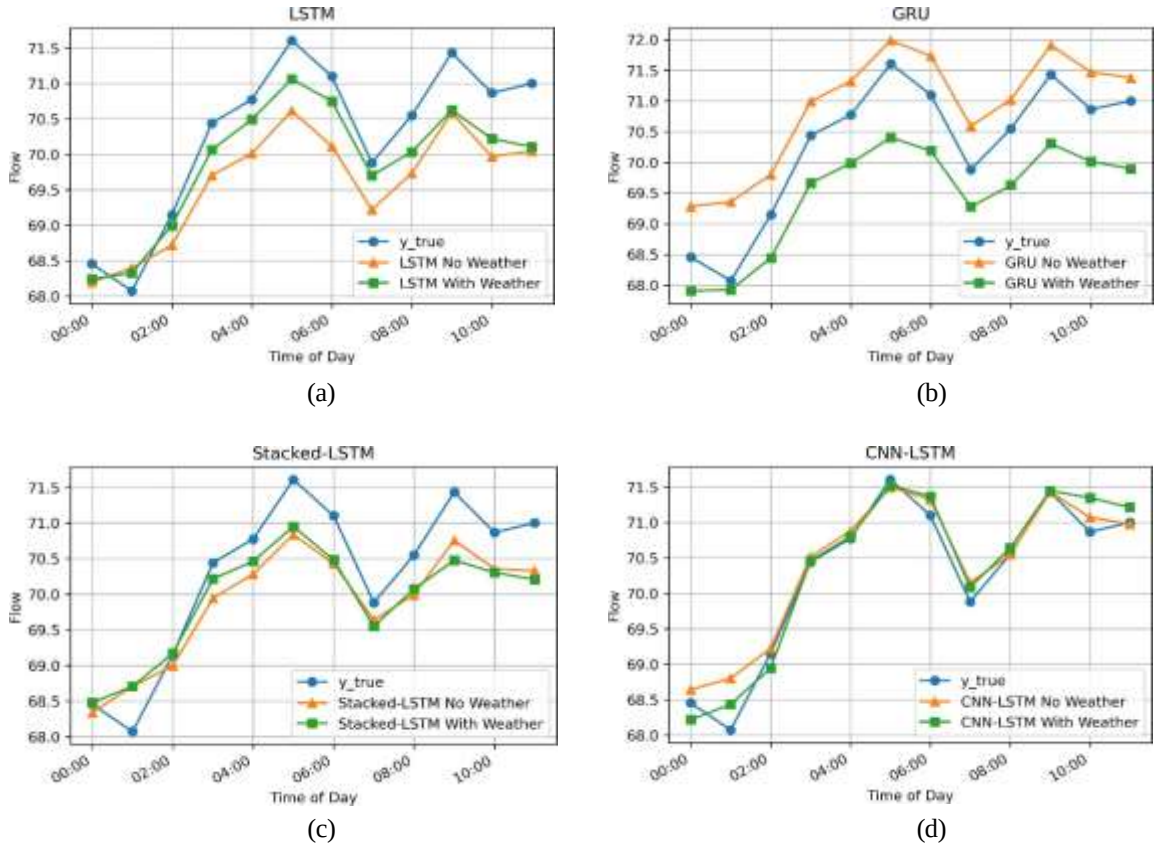


Figure 7. Performance comparison of different models with weather data (Traffic flow prediction) (a) LSTM, (b) GRU, (c) Stacked-LSTM, and (d) CNN-LSTM

REFERENCES

[1] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, Nov. 2003, doi: 10.1061/(ASCE)0733-947X(2003)129:6(664).

[2] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 5, pp. 307–318, Oct. 1996, doi: 10.1016/S0968-090X(97)82903-8.

[3] M. S. Ahmed and A. R. Cook, "Analysis of Freeway Traffic Time-Series Data By Using Box-Jenkins Techniques.," *Transportation Research Record*, no. 722. pp. 1–9, 1979.

[4] W. Li *et al.*, "A general framework for unmet demand prediction in on-demand transport services," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2820–2830, Aug. 2019, doi: 10.1109/TITS.2018.2873092.

[5] J. Guan, W. Wang, W. Li, and S. Zhou, "A Unified Framework for Predicting KPIs of On-Demand Transport Services," *IEEE Access*, vol. 6, pp. 32005–32014, 2018, doi: 10.1109/ACCESS.2018.2846550.

[6] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. L. Tsui, "Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, Oct. 2019, doi: 10.1109/TITS.2018.2879497.

[7] M. S. Dougherty and M. R. Cobbett, "Short-term inter-urban traffic forecasts using neural networks," *International Journal of Forecasting*, vol. 13, no. 1, pp. 21–31, Mar. 1997, doi: 10.1016/S0169-2070(96)00697-8.
[8] T. Pamula, "Impact of Data Loss for Prediction of Traffic Flow on an Urban Road Using Neural Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1000–1009, Mar. 2019, doi: 10.1109/TITS.2018.2836141.
[9] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2018, doi: 1707.01926v3.
[10] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul. 2019, pp. 1720–1730, doi: 10.1145/3292500.3330884.
[11] Y. Li, "Forecaster : A Graph Transformer for Forecasting Spatial and Time-Dependent Data," pp. 1–8.
[12] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for Short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.
[13] F. L. Hall and D. Barrow, "Effect of weather on the relationship between flow and occupancy on freeways," *Transportation Research Record*, no. 1194. pp. 55–63, 1988.
[14] E. I. Vlahogianni and M. G. Karlaftis, "Comparing traffic flow time-series under fine and adverse weather conditions using recurrence-based complexity measures," *Nonlinear Dynamics*, vol. 69, no. 4, pp. 1949–1963, Sep. 2012, doi: 10.1007/s11071- 012-0399-x.
[15] Y. Hou, Z. Deng, and H. Cui, "Short-Term Traffic Flow Prediction with Weather Conditions: Based on Deep Learning Algorithms and Data Fusion," *Complexity*, vol. 2021, pp. 1–14, Jan. 2021, doi: 10.1155/2021/6662959.
[16] Z. Zheng, Y. Yang, J. Liu, H.-N. Dai, and Y. Zhang, "Deep and Embedded Learning Approach for Traffic Flow Prediction in Urban Informatics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3927–3939, Oct. 2019, doi: 10.1109/TITS.2019.2909904.
[17] A. Koesdwiady, R. Soua, and F. Karray, "Improving Traffic Flow Prediction with Weather Information in Connected Cars: A Deep Learning Approach," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9508–9517, Dec. 2016, doi: 10.1109/TVT.2016.2585575.
[18] J. K. Garrett, J. Ma, H. Mahmassani, M. Neuner, and R. Sanchez, "Integrated Modeling for Road Condition Prediction Phase 3 Project Report," no. December. 2020, [Online]. Available: https://rosap.ntl.bts.gov/view/dot/55717.
[19] X. Bao, D. Jiang, X. Yang, and H. Wang, "An improved deep belief network for traffic prediction considering weather factors," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 413–420, Feb. 2021, doi: 10.1016/j.aej.2020.09.003.
[20] K. Cho, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation Kyunghyun," *Journal of Biological Chemistry*, vol. 281, no. 49, pp. 37275–37281, 2006.
[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[22] C. V Dec, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," pp. 1–14, 2016.
[23] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep Learning on Traffic Prediction: Methods, Analysis and Future Directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4927–4943, Jun. 2021, doi: 10.1109/TITS.2021.3054840.
[24] D. Beniaguev, "Historical Hourly Weather Data 2012-2017," *Kaggle*. 2017, [Online]. Available: https://www.kaggle.com/selfishgene/historical-hourly-weather-data.
[25] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." 2016, [Online]. Available: http://arxiv.org/abs/1603.04467.