**TRAFFIC PREDICTION ANALYSIS**

**GROUP MEMBERS**

# Introduction

Traffic prediction analysis is a project concerned with predicting future traffic measurements, including but not limited to volume, and speed in a network road using either historical, time series, or both types of datasets. The primary purpose of this project was to implement various predictive models using different machine learning models. The most performing model among the ones implemented will be considered and selected as the best model for this task. Datasets inputs fed into our predictive models were collected from various sources to make quality, accurate predictions. The traffic prediction analysis project is essential to BI/A because the insights driven from the best model and the visualizations of our dataset provide an initial idea of what we expect in the next few days and as a result of a given problem.

# Related Works

Various topics related to traffic prediction analysis have been implemented over time to prevent and reduce traffic congestion in different cities worldwide. In this paper, we discussed multiple sources and found out the importance of each source. We could identify gaps in the authorities about what was done and what was not done at all the way. Using the gaps found in this source project, we were able to define and come out to solve some of them in this project.

(Zhang & S., 2020)focused on traffic congestion in urban areas as the main problem. They found out that traffic is dynamic but not static, as it spreads once congestion happens at one road segment. With the help of the similarity function, (Zhang & S., 2020)implemented a neural network that could cluster the road points affected by traffic. This was achieved as a result of using affinity propagation clustering. With a ready-made long-short-term neural network model, they could predict the effect of traffic on the cluster formed. The advantage of this research is that the paper explains how the implemented model could predict peak and non-peak hours using the collected open-source dataset. Despite achieving that, the article needs to indicate where traffic will likely occur.

Researchers in the field of transportation have devoted a great deal of time and energy in recent years to studying the causes of traffic congestion on roads and developing methods for forecasting traffic patterns on different road systems. Nonetheless, only a small fraction of researchers has used a combination of heuristic and hybrid predictive models to forecast traffic flows at signalized road intersections. (Hu & Menkir, 2020)developed an expert system in a previous study that uses traffic data consisting of past and present vehicle speed to detect traffic congestions at various road networks. Scalable, a method proposed in related work (Zhang & S., 2020)for predicting vehicle traffic congestion in a grid framework, was developed in a manner similar to. Anwar and colleagues used a technique based on spectral clustering to monitor traffic jams. (Li, et al., 2020) developed a novel prediction model that can estimate the next-time-step traffic volume using a single road traffic segment, considering factors such as traffic flow density and road type, in order to shed light on traffic congestions by analyzing variables such as the current inflow, outflow, traffic volume, etc.

However, (Olayode, Tartibu, Okwu, & Severino, 2021) research improved the model of by employing a support vector machine (SVM) for the prediction of the next time-step traffic speed and traffic volume and subsequently applying this to the estimation of traffic congestion along road segments. Density-based spatial clustering using a noise algorithm was proposed by researchers such as. This method was created to locate and examine a persistently crowded group of grids. In order to foretell how traffic will move on clogged roads, the researchers looked into a deep-learning-based prediction model (Troia, Alvizu, Zhou, Maier, & Pattavina, 2018) that used a restricted Boltzmann Machine and a Recurrent Neural Network. In order to accurately predict future traffic conditions, a model that considers a wide range of parameters was developed. Predictive models, including an autoregressive model, were combined with additional models. The authors (Narejo & Pasero, 2018) combined RMS error with artificial neural networks to create a predictive model. Singular point probabilities were used as a metric for both. Researchers in this field are working around the clock to find solutions to the growing problem of traffic congestion caused by the widespread implementation of intelligent transportation systems. Some studies have shown promising outcomes for predicting traffic patterns. There are two main types of methods used to forecast traffic flows:

1. Methods of machine learning that have been around for a while.
2. Machine learning techniques of a deep nature.

The historical average (HA) method and the autoregressive integrated moving average (ARIMA) method are examples of classic statistical methods. Thereafter, the ARIMA model's characteristics are a synthesis of other models, such as the Kalman-Adjusted ARIMA (KARIMA) time series model and the Seasonal Autoregressive Integrated Moving Average (SARIMA). The primary drawback of such a model, however, is its restricted ability to process non-linear and difficult traffic flow data.

Traditional machine learning techniques outperform the aforementioned models in their ability to model complex, non-linear traffic data. Systematic feature extractor (SVR) and support vector machine (SVM) are two common examples. In order to evaluate the characteristics of traffic data for prediction, conventional models can map low-dimensional non-linear data to high-dimensional space using kernel functions. Predictive model accuracy is heavily dependent on many factors, but the kernel function chosen is one of the most crucial. K-nearest neighbors, and Artificial Neural Networks (ANN)] are two other methods that have been used to predict traffic flow alongside Bayesian models (Pascale & Nicoli, 2012)Traditional machine learning has the major flaw of depending too heavily on engineering and the knowledge of specialists. Improving the effectiveness of these predictive models when dealing with complex and highly non-linear data, however, presents a significant challenge for conventional approaches. Good results have been seen recently with the use of deep learning methods in transportation (Pascale & Nicoli, 2012) particularly in the areas of image processing and natural language processing.

Using temporal and spatial correlation, transportation researchers are increasingly employing deep learning techniques in traffic data mining. Prior studies extended and deepened network layers to learn features from traffic flow data using Deep Belief Networks (DBN) and Stacked Autoencoder Models (SAEs). The DBN model's predictive performance was then improved through the use of traffic flow and weather data, as demonstrated. In order to enhance traffic flow prediction, several models were applied to the temporal correlation of traffic flow data, including Long Short-Term Memory (LSTM), Gated Recurrent Unit Network (GRU), and Nonlinear Autoregressive with External Input (NARX). These forecasting models, however, ignored the significance of the traffic network's spatial organization. Researchers in the transportation sector have applied Convolutional Neural Networks (CNNs) to traffic flow prediction, taking this technique one step further by attempting to capture local spatial characteristics. Only a small fraction of the available literature on traffic flow prediction has been reviewed recently, especially from the perspective of road transportation at intersections. For instance, (Yuan, et al., 2022) looked into the methods and programs of the past decade and provided a comprehensive explanation of the ten difficulties and problems that pedestrians and drivers face today. The involved literature reviews relied heavily on tried-and-true strategies for estimating vehicular traffic. Furthermore, studies conducted provided a thorough breakdown of how to gather traffic data and focused their efforts on traditional machine learning approaches. Models that can be used in practice and are based on traditional methods and early iterations of deep learning are also summarized. An extensive survey of deep neural networks for traffic prediction was outlined by (Yuan, et al., 2022) In their study, they talked about the convolutional, recurrent, and feed-forward neural network architectures that are commonly used today. Recent technological advances in graph-based deep learning, however, were not covered in their study. Similar studies have looked into the use of graph-based deep learning architectures for traffic flow problems. The research they've conducted, however, hasn't touched on some key aspects of vehicular travel. They looked into the topic of traffic flow forecasting exclusively. Other studies on traffic flow forecasting in road transportation share commonalities. It's helpful to think about how people will get to and from every location. Thus, there is a lack of studies that advance the field of traffic flow prediction, especially those that focus on heuristics and algorithms inspired by nature.

Testing results are crucial in demonstrating the value of a proposed prediction model, as demonstrated by a comparison of different model specifications. Recent comparative research has been based on short-term traffic flow forecasting, and this has been explored, which looks into the usefulness and effectiveness of this type of study. They noted that comparing a complex non-linear model to a simple linear one is not always fruitful. Also, the distinction between a model's accuracy, simplicity, and suitability is extremely thin. Model accuracy is highly important, and that it should not be the sole criterion used to select the most effective approach to predicting vehicle traffic flows. Challenges in terms of transferability and suitability to changes in the temporal behavior of traffic flow should be considered when deciding how much time and effort to put into developing the model, techniques, and expertise.

While it is important to select the "best" model from a set of baseline models through testing and comparison, a more practically applicable option would be to pick a heuristic or metaheuristic approach to combine traffic flow predictions. Multiple predictive models may not be able to be combined into a single, accurate model. Predicting future traffic volumes is a well-known example of this. This method of combining predictive models has been used by a number of researchers in traffic flow forecasting; a research (Yuan, et al., 2022) was conducted studies in which they provided statistical recommendations for traffic flow by switching between models in real time. Their study has only one flaw: it does not offer consolidated traffic flow forecasts

A dynamic traffic awareness system was implemented by (Wang & Thulasiraman, 2019). This system was made to find all the road points affected by the traffic at some specific road point at a given time and cluster them. Using these road data points, the system can forecast the impact of traffic on the associated road points. The plan was made using IOT technologies which collect real-time data with the help of sensors, analyze the data, and perform the prediction for peak and non-peak hours. The significance of this research paper is that the system makes predictions based on the real-time flow of traffic data. This is advantageous as we are working with the current condition of the roads. However, its drawback is that the implementation of the system is much costlier as the sensors used are expensive to purchase and install in all cities.

The "Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space-Time Analysis and GRU" paper (G. Dai, C. Ma, & Xu, 2019) proposed a short-term traffic flow prediction model. The model was arrived at by combining the Spatiotemporal analysis with a Gated Recurrent Unit (GRU). The spatiotemporal feature selection was employed as an algorithm to define the optimal time and the volume of spatial data as inputs. In combination, the GRU could be used in preprocessing the spatial feature data to target the prediction from that data. According to (G. Dai, C. Ma, & Xu, 2019), this model was later compared with CNN (Convolutional Neural Network) model. The proposed model outperformed the CNN. The prediction was made to help people select the transportation and the suitable time that they can travel without congestion on the road. However, the model proposed was only made for short-term forecasts.

The journal article "A Period-Specific Combined Traffic Flow Prediction Based on Travel Speed Clustering" (B. Feng, J. Xu, Y. Lin, & P. Li, 2020) presents their prediction on traffic data flow for a period-specific. The vehicle's speed was the determining factor of the predictive model made. This article was also made for short-term forecasting of the traffic flow. The traffic data used in the prediction was collected by loop detectors where there are intersections of signals. This article aims at presenting a time-based prediction of traffic flow.

According to (Mahdavian, et al., 2021), truck traffic volumes on state and interstate highways are the sources of traffic congestion. However, these truck traffic volumes are critical for highway authorities and federal organizations. (Mahdavian, et al., 2021) proposes a hyperparameter framework of optimization after a study review was conducted, indicating that there needed to be more research performed on the utilization of automatic framework for the prediction of truck traffic volumes. The framework proposed was implemented using the Florida interstate highways' historical traffic flow data. These data were meant to perform a test on the proposed framework. The research was effective as the random forest technique of machine learning that was used produced an accuracy of 86%. However, this research was only done for the volume factor. Other factors that cause traffic congestion like speed were not considered. Also, the implementation was only made using one machine-learning algorithm.

According to (Z. Gao, et al., 2019), traffic congestion on a highway during holiday times might result from speeding. (Z. Gao, et al., 2019) uses the highway traffic speed dataset to build a predictive model. In their paper, they proposed an algorithm for Redundant Data Reducing. This algorithm is selected among many since it can reduce the data size used during the model's training process. The proposed algorithm was implemented alongside the LSTM (long short-term memory) machine learning technique. The implementation is made, and the model produces a prediction. Despite the model making a prediction based on traffic speed on a highway during the holiday, the model cannot be effective for the traffic data collected on weekdays.

In ., "Truck Traffic Speed Prediction Under Non-Recurrent Congestion: Based on Optimized Deep Learning Algorithms and GPS Data," article, (J. Zhao et al., 2019) say that heavy and oversized trucks have been restricted from traveling in the city and instead traveling along the expressways around the towns. This causes congestion in these areas as a result. (Z. Gao, et al., 2019) uses the Beijing Sixth Ring Road dataset and proposes the GRU algorithm to make a model for predicting the travel speed of the heavy trucks which drive around the cities on expressway roads under non-recurrent overcrowded conditions. These datasets include both weekdays and weekends, and holidays.

To summarise the related works section, it is evident that most reviews were based on speed and volume datasets as the two main factors resulting in congestion. In addition, most researchers only focused on urban areas, except the few based their research on highways and expressways away from the city. In many cases, the historical data was the mainly used input as only a few researchers could implement data collection means. For example (R. Thulasiram 2019), in

his paper "A Dynamic Traffic Awareness System for Urban Driving," was in a position to implement sensors and other data collection internet of things to collect real-time data. Many of the above-discussed research papers implemented predictive models, but only one machine learning was recognized. Also, their traffic data point collection was basically within urban centers or on the highways. We implemented three distinct predictive models in this paper using three other machine-learning techniques. We used traffic data collected at junction points. This region, we realized, is another source of traffic congestion and has yet to be researched. The idea of implementing the three different machine learning techniques on the junction dataset is to help us select the best model among the three, which can be used in the future in traffic prediction analysis, mostly at junction points.

# Major obstacles to the use/Implementation/management of Traffic prediction Analysis

Once the model is implemented, it is deployed to web and mobile applications; internet connectivity is the only obstacle to using this kind of model. The user must have internet connectivity to access the results of a specific point at the junction. The major obstacle to the implementation of the traffic prediction analysis model is the act of getting enough data and having the skilled technologies which are needed for the performance of a successful model.

The major obstacle that might be experienced in managing the traffic prediction analysis model is the data collection process. This model must receive data inputs at the junction points around the country. This means installing IoT equipment, such as servers and digital cameras, is necessary for data collection. This is expensive as far as the installation is concerned. Many states may need to be in a condition to implement the same.

# Software that facilitated the implementation

This project was based on machine learning techniques. This requires training and testing of models developed. Jupyter software served as the Integrated Development Environment for the implementation of the project. We considered Jupyter since it is the most used machine learning development tool and provides a ready-made environment with almost all model development required packages installed; therefore, less time is spent waiting for package installation.

# What we learned

We learned a lot. We were able to face new technologies upon implementing our research topic. The use of machine learning techniques, data collection, data preparation, data exploration, and data cleaning in general and among others. More so in machine learning techniques, Keras, TensorFlow and Random Forest implementation, and many more.

# Data Exploration and Visualization

In this section, we explored our junction dataset to get information in graphical ways. Below is the diagram showing the junction traffic experience over the years. From the diagram, we find out the following trends.

Junction-1 has the highest total vehicle counts which was evident in the below group\_by too. It also has an upward trend meaning traffic keep increasing over the years

Junction-2 has the most peak bars meaning some high count of vehicles for a certain month - can be treated as outliers but need to look further

Junction-3 has slight upward trend but kind of constant flow over the years

Junction-4 has limited data that we saw in our above exploratory data analysis part as well; only 2017 partial data present

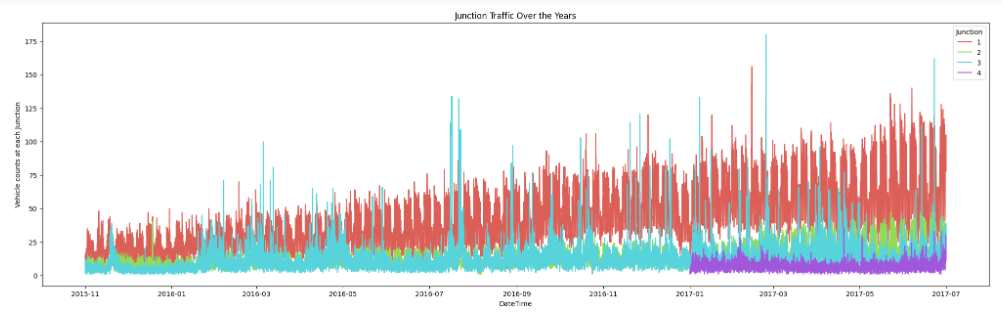


Figure 1 Junction traffic experience over the years

The next diagram below shows the junction traffic experience over the weekdays. From the diagram, we find out the following trends.

The results above are supported by the observation that weekend traffic is lighter than weekday traffic at all the Junctions (Mon-Fri)

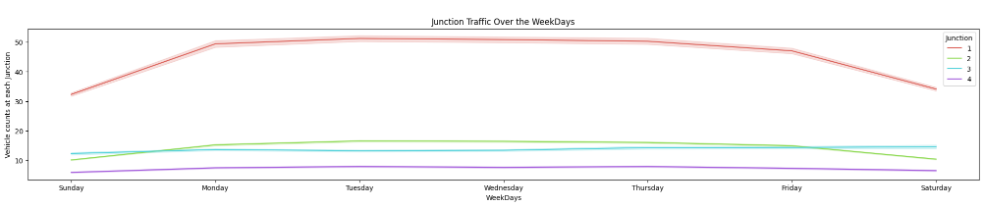


Figure 2 Junction traffic experience over the weekdays

The next diagram below shows the junction traffic experience at hourly. From the diagram, we find out the following trends.

Consistent with these observations, we discover that the least amount of traffic occurs at all Junctions between the hours of 12 and 10 in the morning.

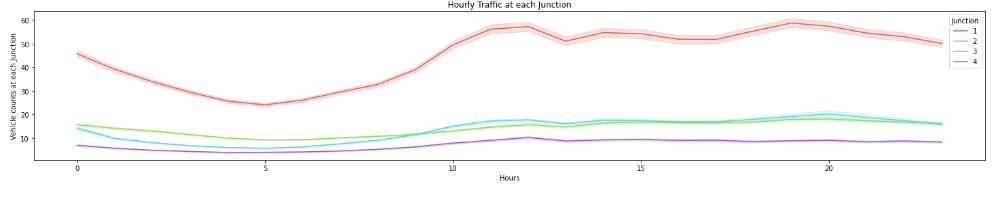
****

Figure 3 Junction traffic experience at hourly

The next diagram below shows the junction traffic experience over the weeks. From the diagram, we find out the following trends.

As mentioned above, the data show that traffic at Junction-4 drops dramatically during Week-27. There may be fewer vehicles on the road because it is the end of June/beginning of July and people are on vacation. In a similar vein, all intersections have reduced traffic in the last months of the year (November and December) due to holidays like Thanksgiving, Christmas, and annual leave.

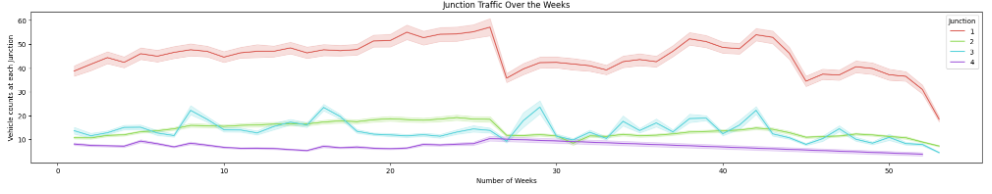


Figure 4 Junction traffic experience over the weeks

# Long Short-Term Memory and GRU Neural Network For Prediction

As a result of their superior capacity to memorize long-term dependencies, RNNs found early application in language models. Gradients of RNNs, however, may disappear as RNNs unfold into extremely deep feed forward neural networks as time delays increase. To address the issue of gradient vanishing, RNNs with forget units—such as long short-term memory (LSTM) and generalized recurrent neural networks (GRU)—have been proposed as a solution. In contrast to GRU NN, which has not been used for traffic flow prediction until recently, LSTM was proposed for language models in 1997 and has been used for this purpose since 2015.

Adam optimizer is a variant of stochastic gradient descent (SGD) optimizer with adaptive learning rates, and it is used for back propagation in time to reduce training error while avoiding local minimal points (BPTT). Despite their impressive expressive power, neural networks are notoriously susceptible to overfitting. Since training neural networks has always been challenging, many regularization techniques have been proposed to mitigate overfitting. Dropout was suggested in 2012 as a highly effective way for training neural networks to get more accurate picture characteristics. Dropout has been challenging to implement in RNN language models because of their recurring nature. However, it wasn't until 2014 that it was revealed that dropout approaches had been successfully applied to RNNs.

# Results

We conduct a comparative analysis between the three distinct machine learning models (LSTM, GRU, and ARIMA). Mean square error (MSE) and mean absolute percentage error (MAPE) are used to evaluate the performance of our models in making predictions as follows.

**MSE =**

**MAE =**

|  |  |  |  |
| --- | --- | --- | --- |
|  | ARIMA | LSTM NN | GRU NN |
| MSE | 801.0065 | 712.0502 | 660.6452 |
| MAE | 18.1753 | 16.127758 | 14.7621 |

Figure 5 MSE and MAE for the three tested models

As can be seen in table above, RNNs are not far off at all. Figures 5 and 6 show the MSES and MAES distribution of traffic. Figure 5 shows that the majority of MSES fall within the range [0, 1200] and that the distribution peak occurs earlier for GRU and LSTM NN than for ARIMA, indicating that the MSES of the two RNNs are often smaller. MAES appear to have a more centrally distributed distribution, and the same property holds true for MAES distribution.

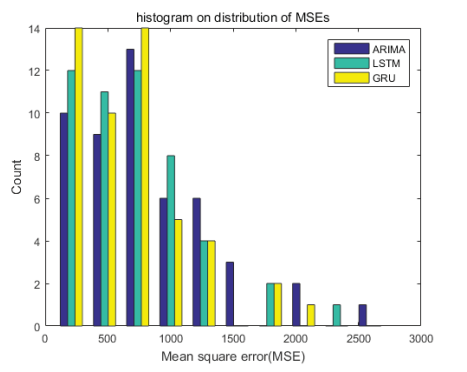


Figure 6 Show the MSES and MAES distribution of traffic.

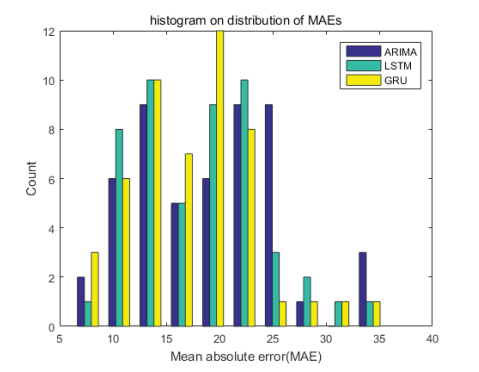


Figure 7 Show the MSES and MAES distribution of traffic.

Based on the outcomes of our suggested research, we can see that both LSTM and GRU NNs perform better than the ARIMA model, with GRU slightly outperforming LSTM NN. Meanwhile, GRU NNS outperform LSTM NNs on 84% of all time series.

# Conclusion

The data visualization graphs provide a clear picture of how traffic changes accordingly based on various factors. Using the visualization graphs in the data visualization section, one can tell the trends of traffic based on hourly, weekdays, holidays and so on. Using such information, one can better plan for the next following days.

In order to analyze and anticipate traffic , the three distinct machine learning models (LSTM, GRU, and ARIMA) models are proposed in this study. In this study, we tested the accuracy of predictions made by three distinct machine learning models (LSTM, GRU, and ARIMA) and discovered that LSTM NNs and GRU NNS are superior to ARIMA model. On average, GRU NNs have a 5% lower MAE than the LSTM NN model and a 10% lower MAE than the ARIMA model.

There are numerous ways in which this proposed strategy has enhanced traffic management. The model analyzes data and updates its predictions of how traffic will behave every hour. People also receive a monthly forecast that considers the upcoming week, weekend, and holidays. Congestion will lessen as a result of earlier information and better planning based on such information. The length-variable time sequence inputs may aid RNNs in automatically determining the ideal time lags in future work testing RNNs with more hidden states.

# References

B. Feng, J. Xu, Y. Lin, & P. Li. (2020). A Period-Specific Combined Traffic Flow Prediction Based on Travel Speed Clustering. 85880-85889.

G. Dai, C. Ma, & Xu, X. (2019). Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space-Time Analysis and GRU. 143025-143035.

Hu, B., & Menkir, H. M. (2020). Traffic Congestion Prediction using Decision Tree, Logistic Regression and Neural Networks.

J. Zhao et al. (2019). Truck Traffic Speed Prediction Under Non-Recurrent Congestion: Based on Optimized Deep Learning Algorithms and GPS Data. 9116-9127.

Li, W., Ban, X. J., Zheng, J., L., H. X., Gong, C., & Li, Y. (2020). Real-time movement-based traffic volume prediction at signalized intersections. . *Journal of Transportation Engineering, Part A: Systems,*, 146.

Mahdavian, A., Shojaei, M. Salem, H. Laman, N. Eluru, & Oloufa, A. A. (2021). A Universal Automated Data-Driven Modeling Framework for Truck Traffic Volume Prediction. 205341-105356.

Narejo, S., & Pasero, E. (2018). An application of internet traffic prediction with deep neural network. . *In Multidisciplinary Approaches to Neural Computing*, 139-149.

Olayode, I. O., Tartibu, L. K., Okwu, M. O., & Severino, A. (2021). Comparative Traffic Flow Prediction of a Heuristic ANN Model and a Hybrid ANN-PSO Model in the Traffic Flow Modelling of Vehicles at a Four-Way Signalized Road Intersection. Sustainabilit.

Pascale, A., & Nicoli, M. (2012). Adaptive Bayesian network for traffic flow prediction. . *Statistical Signal Processing Workshop*, 177-180.

Troia, S., Alvizu, R., Zhou, Y., Maier, G., & Pattavina, A. (2018). *Deep learning-based traffic prediction for network optimization.* In 2018 20th International Conference on Transparent Optical Networks.

Wang, Z., & Thulasiraman, P. (2019). Foreseeing Congestion using LSTM on Urban Traffic Flow Clusters. *6th International Conference on Systems and Informatics (ICSAI)*, 768-774.

Yuan, C., Li, Y., Huang, H., Wang, S., Sun, Z., & Li, Y. (2022). *Using traffic flow characteristics to predict real-time conflict risk: A novel method for trajectory data analysis. Analytic Methods in Accident Research.*

Z. Gao, X. Yang, J. Zhang, H. Lu, R. Xu, & W. Diao. (2019). Redundancy-Reducing and Holiday Speed Prediction Based on Highway Traffic Speed Data. 31535-31546.

Zhang, & S. (2020). Representation of traffic congestion data for urban road traffic networks based on pooling operations.