**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.

Transportation experts have spent a lot of time and effort in recent years investigating what causes traffic jams and how to predict traffic flows on various road networks. While many studies have been conducted to predict traffic flows at signalized road intersections, only a minority of researchers have used a combination of heuristic and hybrid predictive models. In a previous study (Hu & Menkir, 2020), the authors developed an expert system that, by analyzing traffic data (including past and present vehicle speed), can detect traffic congestions on various road networks. Scalable approach to predicting gridlock in car traffic was influenced by research that used a similar grid-based method (Zhang & S., 2020). Anwar and his team of researchers used a spectral clustering strategy to keep tabs on traffic flows. In order to better understand traffic congestion, a novel prediction model was developed (Li et al., 2020) that uses only a single road traffic segment to predict the traffic volume at a later time step, taking into account factors such as traffic flow density and road type.

A support vector machine was used in new research (Olayo et al., 2021) to estimate traffic jams along road sections using forecasts of future traffic speed and volume (SVM). For density-based spatial clustering, these experts proposed a noise algorithm. This method was created to single out and examine a collection of grids that were chronically oversubscribed. Researchers investigated a deep learning-based prediction model (Troia et al., 2018) that used a restricted Boltzmann Machine and a Recurrent Neural Network to predict how traffic would flow on congested roads. A model was created that takes into account a large number of factors in order to make reliable predictions about future traffic conditions. Multiple predictive models, including an autoregressive model, were combined to make predictions. Troia et al., (2018) looked into a restricted Boltzmann machine and a recurrent neural network-based deep learning prediction model. (Li et al., 2018) to foretell how traffic would move on clogged roads. For both, the probability of a singularity was the criterion. Intelligent transportation systems have become increasingly popular, leading to an increase in traffic congestion that has researchers in this field working around the clock to find a solution. In order to better anticipate traffic patterns, some studies have shown encouraging results. There are two primary approaches used for predicting traffic volumes:

1. Traditional approaches to machine learning.
2. In-depth machine learning methods.

The historical average (HA) and the autoregressive integrated moving average (ARIMA) (Dutta et al., 2011) techniques are two examples of tried-and-true statistical methods. The ARIMA model is thus a hybrid of the Kalman-Adjusted KARIMA, time series model and the Seasonal Autoregressive Integrated Moving Average model (SARIMA). The inability to account for sophisticated traffic flow data presents a significant barrier for such a model.

The aforementioned models aren't as effective as more conventional machine learning methods when modeling complex, non-linear traffic data. A couple common examples are the systematic feature extractor (SFE) and the support vector machine (SVM) (Jiang et al., 2011). Conventional models can evaluate the features of traffic data for forecasting by using kernel functions to transform the data from low- to high-dimensional space. One of the most influential factors is the choice of kernel function employed by the predictive model, though there are many others. Predicting traffic flows is not limited to Bayesian models; k-nearest neighbors and ANN have also been used (Pascale & Nicoli, 2012). One major drawback of conventional machine learning is that it places an excessive emphasis on engineering and expert knowledge. However, it is a significant challenge for traditional methods to increase the accuracy of these predictive models when working with complex and highly non-linear data. Deep learning techniques have recently been successfully applied to the transportation industry, particularly in the areas of image processing and natural language processing (Pascale & Nicoli, 2012).

Researchers in the transportation sector are increasingly turning to deep learning methods for traffic data mining due to the success they have had in using temporal and spatial correlation. To better learn features from traffic flow data, prior studies have extended and deepened network layers using methods like DBN and Stacked Autoencoder Models (SAEs). According to the results, the DBN model's predictive performance can be enhanced by including traffic and weather information. To enhance traffic flow prediction, researchers have used models that take into account the temporal correlation of traffic data, such as the Long Short-Term Memory (LSTM) ( Graves, 2012)), the Gated Recurrent Unit Network (GRU)( Dey & Salem 2017), and the Nonlinear Autoregressive with External Input (Wunsch et al., 2018) (NARX). These prediction models did not, however, account for the traffic network's spatial organization. Transportation researchers have used Convolutional Neural Networks (CNNs) to predict traffic flows, going a step further by attempting to capture local spatial characteristics. Recent literature reviews on traffic flow prediction are extremely limited, especially when looking at the topic from the point of view of road transportation at intersections. For instance, (Yuan, et al., 2022) researched strategies and initiatives from the previous decade and provided a detailed account of the ten challenges and problems currently faced by pedestrians and motorists. When estimating vehicle traffic, the reviews of relevant literature mostly stuck to tried-and-true methods. The studies also detailed the steps necessary to collect traffic data and emphasized the use of conventional machine learning techniques. Methods and early versions of deep learning that can be used in practice are also summarized. For a thorough introduction to the use of deep neural networks in traffic prediction, see (Yuan, et al., 2022) Convolutional, recurrent, and feed-forward neural networks, three of the most popular types, were covered. However, they did not consider the most recent developments in graph-based deep learning. There has been prior study into the feasibility of employing graph-based deep learning algorithms to traffic flow problems. However, there are important facets of auto travel that have been left out of their studies. Only traffic flow forecasting research was conducted. There are commonalities with other studies on predicting traffic on roads. Consider the best routes for visitors to take to get to and from each destination. As a result, little research has been done on predicting traffic flows, particularly not using heuristics and algorithms that are inspired by natural phenomena.

By comparing different model specifications, we demonstrate the value of testing results in establishing the credibility of a stated prediction model. Short-term traffic flow forecasting has been the basis for a lot of recent comparative research, and that research has been examined to see how effective and useful it can be. They pointed out that it's not always helpful to compare a complicated non-linear model to a straightforward linear one. A model's accuracy, simplicity, and suitability are all closely related. Model accuracy is crucial, but it shouldn't be the only factor considered when deciding on the best strategy for predicting traffic volumes based on vehicles. Before investing too much time and resources into developing the model, techniques, and expertise, it is crucial to think about the challenges of portability and adaptability to shifts in the temporal habits of traffic flow.

Although testing and comparing multiple baseline models to determine which one is "best" is essential, in order to combine traffic flow predictions in a way that is more practically relevant, one option is to use a heuristic or metaheuristic approach.. A unified, accurate predictive model may be impossible to construct from a collection of separate models. In this sense, traffic volume forecasting is a well-known application. Researchers have used this method of combining predictive models in traffic flow forecasting, with one study (Yuan, et al., 2022) providing statistical recommendations for traffic flow by rapidly switching between models. Their research has one major limitation: it does not provide unified traffic flow predictions.

A dynamic traffic awareness system was used (Wang & Thulasiraman, 2019). This system's objective is to gather all of the locations anywhere along road that are affected by traffic at any given time in one place. 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.

A short-term traffic flow prediction model was put forth in the paper "Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space-Time Analysis and GRU" (G. Dai et al., 2019). Combining a Gated Recurrent Unit with the spatiotemporal analysis, the model was created (GRU). The algorithm used to define the ideal time and the volume of spatial data as inputs is the spatiotemporal feature selection. In order to target the prediction from the spatial feature data, the GRU could be used in combination. This model was later compared with the CNN (Convolutional Neural Network) model, according to (G. Dai et al., 2019). The suggested model performed better than 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" (Feng, 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 (Qi, 2012), 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. (Qi, 2012) Proposes a hyper parameter 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 (Qi, 2012) 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 (Qi, 2012). 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.

The highest total vehicle counts were found at Junction-1, as shown in the group by below. Additionally, it is on the rise, so traffic has been growing over time.

Junction-2 has the most peak bars, which indicates some unusually high vehicle counts for a given month that should be treated as outliers but require further investigation.

Over the years, Junction-3 has had a somewhat constant flow despite a slight upward trend.

We observed limited data for Junction-4 in our above exploratory data analysis section, with only 2017 partial data available.

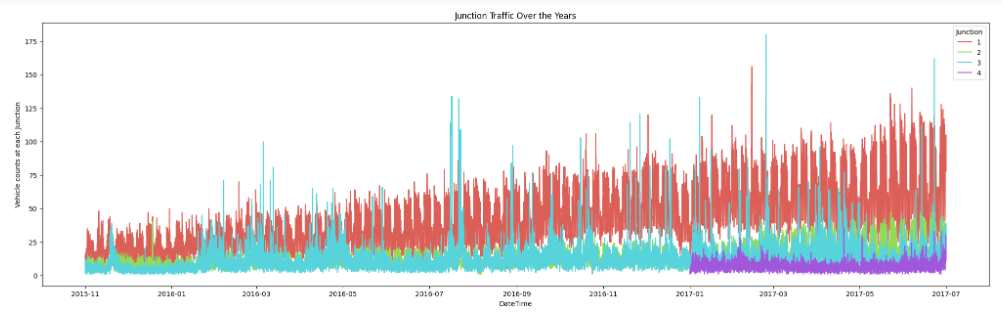


Figure 1 Junction traffic experience over the years

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

The finding that weekend traffic at all of the Junctions is smoother than weekday traffic supports the results above (Mon-Fri)

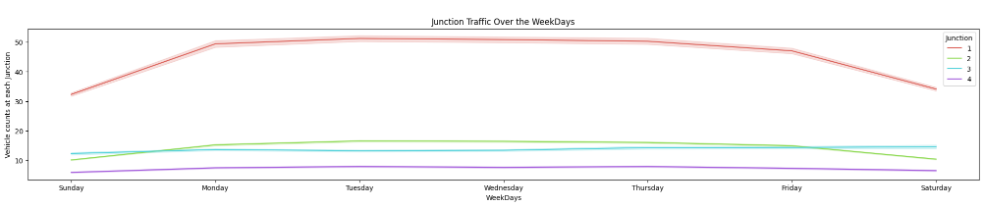


Figure 2 Junction traffic experience over the weekdays

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.

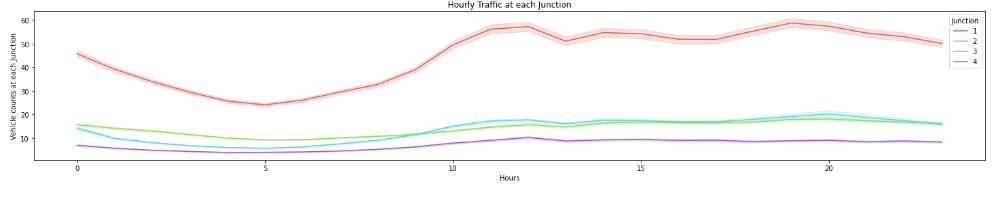
****

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.

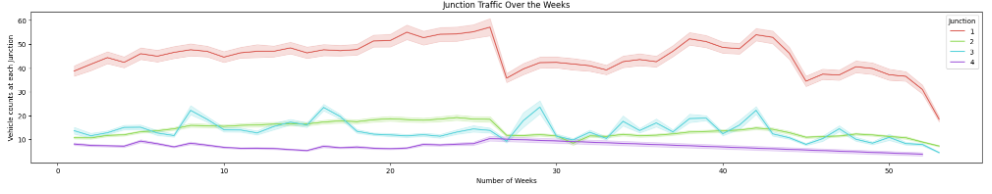


Figure 4 Junction traffic experience over the weeks

# 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.

Back propagation in time (BPTT) is used with the Adam optimizer, a stochastic gradient descent (SGD) variant with adaptive learning rates, to decrease training error while avoiding local minimal points. 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.

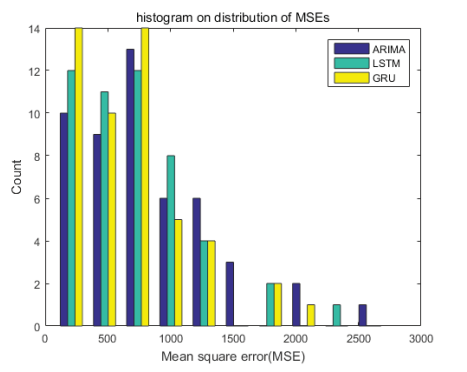


Figure 6 Show the MSES and MAES distribution of traffic.

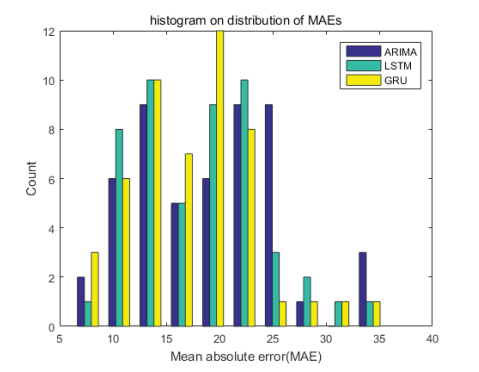


Figure 7 Show the MSES and MAES distribution of traffic.

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.

Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* (pp. 1597-1600). IEEE.

Dutta, D., Patel, N. R., & Kundu, A. (2011). Analyzing the performance of auto regressive integrated moving average (ARIMA) model for predicting agricultural productivity in eastern Rajasthan. *Res. J. Agric. Sci*, *2*, 555-559.

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.

Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.

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

Jiang, X., Xia, Y., Hu, J., Zhang, Z., Shu, L., & Sun, C. (2011). An S-transform and support vector machine (SVM)-based online method for diagnosing broken strands in transmission lines. *Energies*, *4*(9), 1278-1300.

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, C., Wang, J., & Ye, X. (2018). Using a recurrent neural network and restricted Boltzmann machines for malicious traffic detection. NeuroQuantology, 16(5).

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.

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.

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.

Wunsch, A., Liesch, T., & Broda, S. (2018). Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). *Journal of Hydrology*, *567*, 743-758.

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.