National College of Ireland

Traffic Flow Forecasting using DeepAR

MSc Research Project
Data Analytics

Prajwal Joshi
Student ID: 22111034

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Prajwal Joshi
Student ID:	22111034
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Abid Yaqoob
Submission Due Date:	31/01/2024
Project Title:	Traffic Flow Forecasting using DeepAR
Word Count:	3943
Page Count:	10

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Traffic Flow Forecasting using DeepAR

Prajwal Joshi
22111034

Abstract

Improved traffic volume forecasting techniques are required due to disruptions in to travel patterns caused by the COVID-19 pandemic . During pandemics, traditional approaches find it difficult to handle complicated dynamics. This work investigates the use of an RNN architecture called DeepAR to forecast traffic volume both during and after the epidemic. DeepAR is a good fit for simulating pandemic-induced traffic patterns because of its capacity to manage temporal dependencies and long-range interactions. DeepAR models were trained for various prediction horizons using historical data covering pre-pandemic, pandemic, and post-pandemic eras. The results show that DeepAR works better than conventional techniques at capturing pandemic-induced changes in traffic patterns. Its flexibility makes proactive traffic control techniques possible. The implications of these findings for traffic management after the pandemic are noteworthy. Transportation authorities can improve traffic flow, manage infrastructure, and lessen congestion by utilising DeepAR. Because of its adaptability, DeepAR is a useful tool for transportation systems to adjust to constantly shifting traffic patterns.

1 Introduction

Transportation systems throughout the world have been disrupted by the COVID-19. This has presented serious issues for traffic management organisations. Traditional traffic forecasting techniques, which are mostly statistical in nature have been unable to determine the pandemic's complicated effects on mobility trends, temporal dynamics, and the intrinsic non-stationarity of traffic data. This has led to increase in need for a traffic forecasting system based on machine learning.

Due to the pandemic's widespread lockdowns, remote work policies, and travel restrictions, traffic patterns have fundamentally changed, resulting in significant variations in traffic volume both daily and on different days of the week. The development of innovative and reliable traffic forecasting techniques that can precisely assess the pandemic's impact and offer insightful information for enhancing traffic management strategies has become necessary due to this extraordinary change in traffic patterns.

Deep learning has become a potent tool for time series forecasting, especially for complicated datasets with temporal dependencies and nonlinear relationships. The efficacy of Deep Autoregressive Neural Networks(DeepAR)(Salinas et al.; 2020), created especially for time series data forecasting, in predicting a variety of time series datasets, including traffic volume, has drawn a lot of attention. Long-term dependence of data and seasonal nature of variables are handled excellently by it (Cheng et al.; 2023), resulting in excellent long term forecasting. Due to this, DeepAR has become popular for forecasting in

transportation (Lunacek et al.; 2021) and various industries. (Dong et al.; 2020) (Liao and Liang; 2021) (Mahdy et al.; 2020) (Santos Escriche et al.; 2023).

This research examined the use of DeepAR to forecast traffic volume during the COVID-19 pandemic in boroughs of New York City. Training and assessing DeepAR models for various prediction horizons and time intervals, the author uses historical traffic data from the pre-pandemic, pandemic, and post-pandemic periods. By comparing DeepAR's performance to conventional statistical methods, the results offer significant insights into the technology's capacity to capture changes in traffic patterns brought about by pandemics.
The findings of this study have significant implications for post-pandemic traffic management. By leveraging DeepAR's forecasting capabilities, transportation agencies can optimize traffic signal timing, adjust public transportation schedules, and implement road capacity enhancements to improve traffic flow and reduce congestion. DeepAR's adaptability to diverse traffic conditions makes it a valuable tool for adapting transportation systems to the ever-changing dynamics of traffic patterns.
The study's conclusions will have a big impact on traffic management after the pandemic. Transportation agencies can improve traffic flow and reduce congestion by implementing road capacity enhancements, adjusting public transportation schedules, and optimising traffic signal timing by utilising DeepAR's forecasting capabilities. DeepAR is a useful tool for adjusting transport systems to the constantly shifting dynamics of traffic patterns because of its adaptability to a wide range of traffic conditions.

In this report, the author starts by providing the abstract of the project. The abstract is succeeded by a brief introduction to the project. A few related works are critically reviewed, followed by the description of methodology and results. Evaluation methods are mentioned in the end, followed by the conclusion and any possible future works.Overall, the author attempts to provide the answer for the question: How is DeepAR better than traditional methods of traffic flow forecasting?

2 Related Work

Hou, Xing and Liang (2023) propose a traffic prediction method that combines the advantages of LSTM and ARIMA models. While LSTM excels at nonlinear relationships, linear trends are best captured by ARIMA. After pre-processing the data, the proposed model utilises LSTM and ARIMA are to handle nonlinear patterns and linear trends respectively . LSTM is refined by feeding predictions from the ARIMA model. The model performs better than traditional models as it captures both linear and non-linear relationships, control long term dependencies and adepts to various conditions of the traffic.
In the study, Liu (2022) came up a short term algorithm that predicts traffic flow based on multi-machine learning. The purpose of the algorithm is to enhance the accuracy of the prediction by correlating traffic flows on various related roads and decreasing the data dimension. Various machine learning models are deployed by algorithm for training and the best classifier is chosen by competition. The author tested the algorithm on METR-LA data set to demonstrate the superiority of proposed algorithm over traditional methods.
Ma, Dai and Zhou (2022) propose a short-term enhanced LSTM and time series analysis based prediction model. The model's ability to capture long-ranged time dependencies

are enhanced by the addition of a bidirectional LSTM, while it still retained its ability to effectively capture seasonal and random components. The model's superiority was established by testing on real world data, where it surpassed other models in terms of accuracy.

The research by Trinh, Tran and Do (2022)suggests the combination of various multivariate time series models in conjunction with distributed computing. The models were evaluated on a dataset that consisted of traffic data in Ireland. It was concluded that deep learning models are more precise than machine learning models. Furthermore, the models were also experimented under different situations, indicating their strengths and limitations.

Ma, Huang and Ullah (2020)propose an approach that utilises support vector regression along with chaos theory. The proposed model provides more accurate predictions than the traditional prediction techniques like ARIMA and SVM. Despite being tested on real time data and indicating potential, the model requires further research for expanding into a wider range of traffic data and external factors.

For the research, Cvetek, Muštra, Jelušić and Abramović (2020) conducted traffic volume forecasting by collecting the data from various bluetooth detectors. They employed various time series forecasting model such as random walk, unobserved component model, ARIMA, SARIMA and exponential smoothening. The researchers concluded that ARIMA was the best performing model for forecasting. The article indicated that the researchers were willing to implement more sophisticated models like neural networks, hence admitting the limitation of the model. Also the researchers did not take in account the data for weekends.

Qu, Qie, Li, Liu, Li and Shi (2022)introduce Time Slot Recurrent Neural Networks (TS-RNN) in this study. The proposed model handles the variations in traffic patterns by slotting the historical data into different slots and training an individual sub-model for each division. Due to this, the model is able to focus on more correlated data within slot, hence improving the accuracy of prediction. The proposed model's ability to focus on the correlated data segments indicates its suitability for real time traffic flow data.

Duan and Gong (2022) propose a modified Space-Time Auto-regressive Integrated Moving Average (STARIMA) model. The STARIMA model can determine complex relationships between the traffic flow data across the road network by incorporating spatial dependencies between traffic data from different locations. The researchers modified the traditional STARIMA model by implementing a new parameter, weighting scheme to tackle temporal lags and 2 new methods for selecting autoregressive parameters and moving average parameters respectively. The model is evaluated on Beijing Urban Road Network data set, where it significantly outperforms the traditional forecasting methods like ARIMA, SARIMA and BPNN.

A model named PANGO was proposed by Zhou and Xu (2022) in the study, for prediction of traffic flows to certain venues. It utilises a combination of Long Short Term Memory(LSTM) networks and clustering in order to enhance the accuracy. What differentiates PANGO from the traditional models is its ability to process the long term cycle characteristics of the data. The factor that limits PANGO is that it's only able to predict traffic flow to and from a particular venue.

In order to address disadvantages of regular RNNs for long term predictions, Qin, Niu, Wang and Qian (2021) introduce a model that combines the JANet architecture and attention mechanism. The complex architecture of LSTM network is simplified without compromising the accuracy and weights are assigned to relevant features by attention

mechanism, hence laying more emphasis on more relevant features of the data. The model has the potential to be utilised for traffic flow optimisation as it performed better than baseline models on test data set.
For short-term traffic prediction, Fang, Cai, Fan, Yan and Zhou (2021) proposed a model that combined Kalman's filter and LSTM. Kalman's filter's ability to carry out dynamic updates in real-time data is complimented by LSTM's ability, which is capturing long term dependencies in a data set. Experiments indicated that the proposed model is robust to noise and can handle the traffic flow patterns of non-linear nature.
Zhang, Liu and Feng (2021) proposes an enhanced neural network for short term traffic prediction. Genetic algorithm is used in conjunction, as conventional neural networks are prone to overfitting. Optimisation of neural networks structure by preventing overfitting is first carried out by genetic algorithm. Back propagation helps with weight distribution.
Dong, Lei, Jin and Hou (2018) utilise XGBoost model in order to conduct a short term traffic prediction. A wavelet de-noising algorithm is employed to remove the noise from the traffic data and assign the frequency label, i.e. high or low. The label is then used to train the model. The model was able to surpass support vector machine in the evaluations.
Yu, Zhao, Gao and Lin (2019) proposes a bidirectional long short term memory network reinforced with particle swarm optimisation for predicting short term traffic flow with high accuracy. PSO algorithm considers both, the current best and global best while iteratively modifying the arrangement of particles. LSTM is capable of processing both past and future data. The model was tested on real-time data for a location in Korea and was experimentally proven to be superior to other time series forecasting methods and some deep learning models too.
Xie, Jin and Li (2021) argue that time series forecasting methods like exponential smoothening are incapable of dynamically capturing the complexities in traffic flow data. A fuzzy clustering algorithm was implemented on the data to address this problem. The algorithm clustered the data based on it's similarity and was then used to train prediction models. The argued strategy was implemented on real time data and was superior to base model in terms of accuracy.

3 Methodology

The researcher carried out a KDD methodology in order to conduct time series prediction of traffic data in various boroughs of New York. The image below describes the process that was followed throughout, refer figure1.

3.1 Data Acquisition

In order to carry out the research, Automated Traffic Volume Count dataset was used. It was obtained from New York City's open data repository. Source: https://data.cityofnewyork.us/Transp
Traffic-Volume-Counts/7ym2-wayt. The dataset consisted of following columns: RequestID, Boro, Yr(Year),M(Month), D(Day), HH(Hour), MM(Minute),Vol(traffic volume), SegmentID, toSt,fromSt, wktGeom and direction. The required data is imported into a Python environment with the aid of Google Colab notebook.

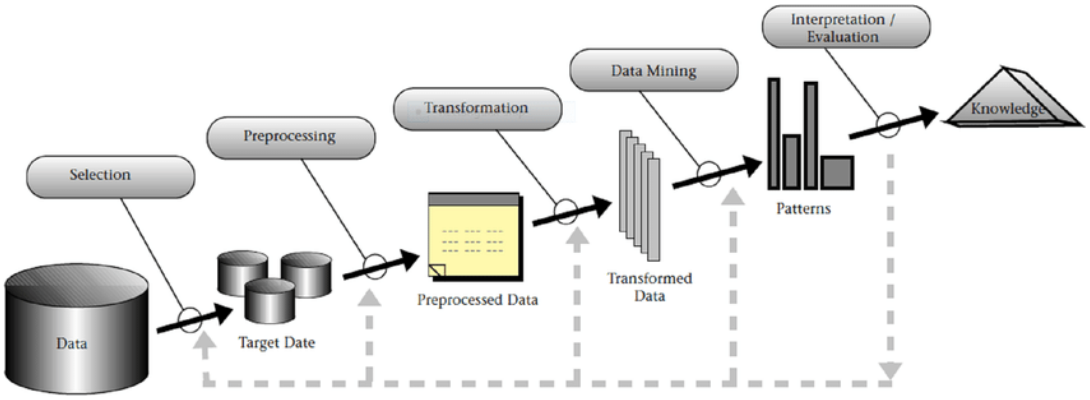

Figure 1: KDD

3.2 Data Preprocessing and Transformation

The dataset was checked for missing and garbage values. There were none in the time related columns and Vol. All the columns that were not Boro, Yr(Year),M(Month), D(Day), HH(Hour), MM(Minute),Vol(traffic volume) were dropped as they held no significance for the research. After cleaning up the data, feature engineering was used in order to create a new feature called Datetime, which was the timestamp column created by combining all the time indicating columns. As the research only pertains to data during covid time, all the records that were not between 01/03/2020(1st March 2020) and end of 2021 were dropped from the dataframe.The data was then aggregated on Boro column and then split into 5 different dataframes. Now that time series were established for all the boroughs, the Boro column was dropped from all the dataframes.

4 Design Specification

The research is a 3 step process, i.e. data preperation, data modelling and visualisation of the data. The process of data collection, pre-processing and exploratory analysis and data transformation has been described in subsection 3.2. In the modelling step, various algorithms namely DeepAR, ARIMA and LSTM are implemented on the time series data sets. They will also be evaluated with metrics like root mean square error. At last, visualisations are performed, if required, refer figure 2

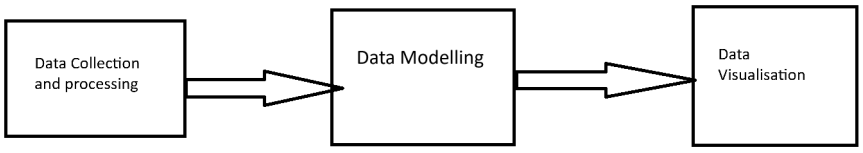

Figure 2: Design Specification of the Research

5 Implementation

The implementation of research is discussed in this section. Data is collected from the NYC opendata website and is loaded into the python environment, that being Google colab. Pandas read_csv() function is used to accomplish that. Once the data is imported, pre-processing steps are implemented on it. Null values are checked for and removed, if found. Columns that are not relevant for the study are dropped using drop() method. After the columns are dropped, feature engineering is implemented to obtain a timestamp column, which is crucial for the research. Once the timestamp column is obtained, all the features that represented time were dropped. The only columns in the dataframe are Boro, datetime and Vol. As the study requires only COVID Data, all the records that do not lie between 01/03/2020 and 31/12/2021 are dropped. The dataframe is then aggregated on boroughs and is split, hence creating a time series dataframe for all boroughs. Boro column is then dropped from the dataframe. 3

	RequestID	Boro	Yr	M	D	HH	MM	Vol	SegmentID	WktGeom	street	fromSt	toSt	Direction	Datetime
0	32110	Bronx	2020	3	3	0	0	6	88955	POINT (1022455.0901136672 260806.54150942338)	EAST 218 STREET	Barnes Avenue	White Plains Road	WB	2020-03-03 00:00:00
1	32110	Bronx	2020	3	3	0	15	15	88955	POINT (1022455.0901136672 260806.54150942338)	EAST 218 STREET	Barnes Avenue	White Plains Road	WB	2020-03-03 00:15:00
2	32110	Bronx	2020	3	3	0	30	11	88955	POINT (1022455.0901136672 260806.54150942338)	EAST 218 STREET	Barnes Avenue	White Plains Road	WB	2020-03-03 00:30:00
3	32110	Bronx	2020	3	3	0	45	3	88955	POINT (1022455.0901136672 260806.54150942338)	EAST 218 STREET	Barnes Avenue	White Plains Road	WB	2020-03-03 00:45:00
4	32110	Bronx	2020	3	3	1	0	5	88955	POINT (1022455.0901136672 260806.54150942338)	EAST 218 STREET	Barnes Avenue	White Plains Road	WB	2020-03-03 01:00:00

Figure 3: Sample of Data used for the research

An ARIMA model is created using statsmodels library. The data in the dataframes is at timestamp of 15 minutes. Grid search was implemented to find the best value of p, d and q. The model was then fitted using fit function. The fitted model is then used to conduct forecasts for all the time series models.The model was also tested using the 20% of records in the data sets. Root mean square error for every implementation was calculated for model evaluation.

For implementation of LSTM, keras library is used. The data is first minimised using MinMaxScaler() function to bring all features to an equal scale. The data is then split into train and test set and a LSTM model is fitted on the training data using fit() function with implementation of Adam optimiser. Predictions are performed using test data and RMSE is calculated for evaluation.

DeepAREstimator() is imported from GluonTS along with functions trainer() and make_evaluation_prediction(). Optuna library is used for hyperparameter optimisation, which is carried out by creating a study of 5 trials each. The models were trained for each data set ensuring the most optimal hyperparameters were used. The dataframes are then converted into ListDataSet, a data structure specific to GluonTS which can be used for time series analysis and forecasting. Forecasts are generated and rmse is calculated for evaluation.

6 Evaluation

In order to evaluate the models, root mean square error(RMSE) was utilised as the metric. Improvement in the model is indicated by minimisation of it.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

6.1 ARIMA Model

ARIMA Models for every data set were trained with gradient descent, which ensured the optimal parameters used for training. The root mean square error for ARIMA model is

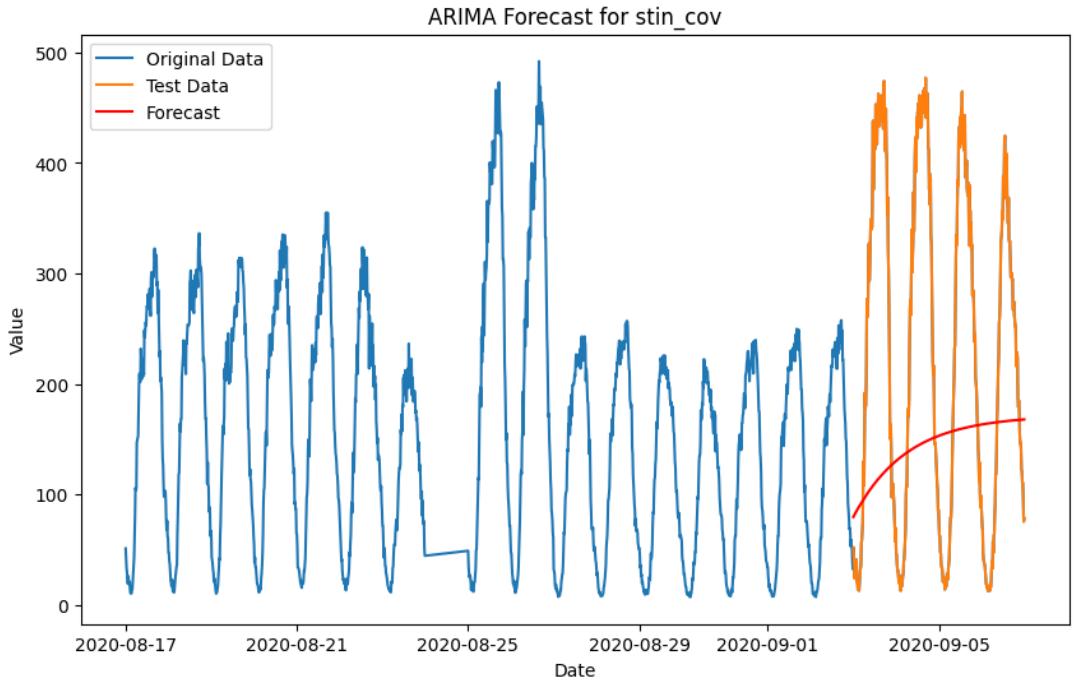

Figure 4: ARIMA Forecasts

as follows:

1. Manhattan Data Set: 391.15
2. Bronx Data Set: 149.22
3. Staten Island data Set: 178.68
4. Brooklyn Data Set: 76.96
5. Queen Data Set: 41.22

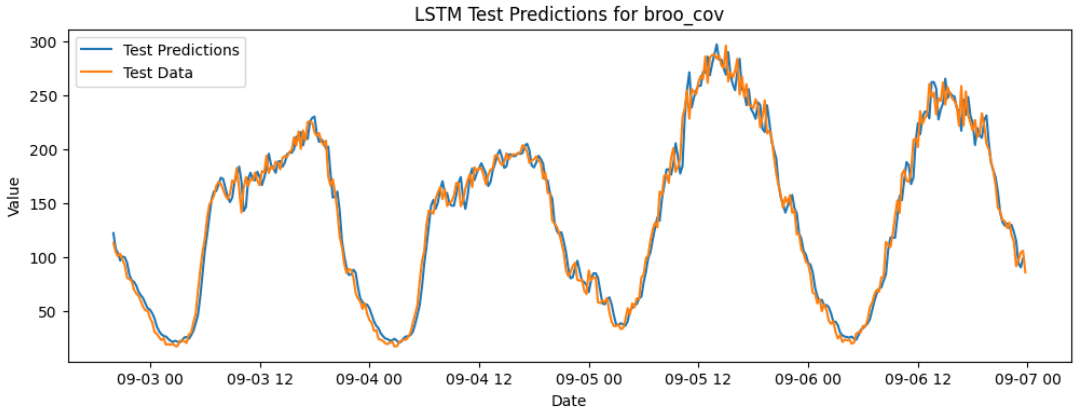

Figure 5: LSTM Test Data vs Test Predictions

6.2 LSTM Model

The LSTM model was fitted with 50 units and 4 time steps. Adam optimiser was also implemented. The root mean square error for ARIMA model is as follows:

1. Manhattan Data Set: 549.94
2. Bronx Data Set: 170.88
3. Staten Island data Set: 267.20
4. Brooklyn Data Set: 149.21
5. Queen Data Set: 65.07

6.3 DeepAR Model

The DeepAR model was trained after implementation of hyperparameters using optuna. Following is the list of root mean squared error.

1. Manhattan Data Set: 54.60
2. Bronx Data Set: 2.89
3. Staten Island data Set: 10.13
4. Brooklyn Data Set: 18.30
5. Queen Data Set: 1.74

6.4 Discussion

It was indicated with the experiments that DeepAR was the best performing model as it had the lowest RMSE over all the data sets. ARIMA was the second best performing model. One of the reasons for LSTM's poor performance may be the lesser number of data points in the data sets. The data set had data of some months missing from it. LSTM's performance could be improved with the completed data.

7 Conclusion and Future Work

The research concluded with being proven that DeepAR model performed better on time series data that two other traditional models, i.e. ARIMA and LSTM. However, the lackluster performance of LSTM model may not be its own fault but could be due to low data points in the datasets. If efficiently trained, DeepAR models can be a vital tool in prediction of traffic. There is plenty literature evidence to prove it.
Further work could be done to compare DeepAR models to the likes of other state of art models like deep neural networks. Data with higher frequency may also be needed.

References

Cheng, Y., Xing, W., Pedrycz, W., Xian, S. and Liu, W. (2023). Nfig-x: Nonlinear fuzzy information granule series for long-term traffic flow time-series forecasting, *IEEE Transactions on Fuzzy Systems* **31**(10): 3582–3597.

Cvetek, D., Muštra, M., Jelušić, N. and Abramović, B. (2020). Traffic flow forecasting at micro-locations in urban network using bluetooth detector, *2020 International Symposium ELMAR*, pp. 57–60.

Dong, M., Wu, H., Hu, H., Azzam, R., Zhang, L., Zheng, Z. and Gong, X. (2020). Deformation prediction of unstable slopes based on real-time monitoring and deepar model, *Sensors* **21**(1): 14.
URL: *http://dx.doi.org/10.3390/s21010014*

Dong, X., Lei, T., Jin, S. and Hou, Z. (2018). Short-term traffic flow prediction based on xgboost, *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 854–859.

Duan, W. and Gong, Z. (2022). Research and application of urban real-time traffic flow prediction based on starima, *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pp. 1521–1525.

Fang, W., Cai, W., Fan, B., Yan, J. and Zhou, T. (2021). Kalman-lstm model for short-term traffic flow forecasting, *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Vol. 5, pp. 1604–1608.

Hou, T., Xing, H. and Liang, X. (2023). Traffic prediction method for time series networks based on arima-lstm model, *2023 IEEE 16th International Conference on Electronic Measurement Instruments (ICEMI)*, pp. 384–388.

Liao, Y. and Liang, C. (2021). A temperature time series forecasting model based on deepar, *2021 7th International Conference on Computer and Communications (ICCC)*, IEEE, pp. 1588–1593.

Liu, M. (2022). Short-term traffic flow prediction based on multi-machine learning, *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, pp. 289–292.

Lunacek, M., Williams, L., Severino, J., Ficenec, K., Ugirumurera, J., Eash, M., Ge, Y. and Phillips, C. (2021). A data-driven operational model for traffic at the dallas fort worth international airport, *Journal of Air Transport Management* **94**: 102061.
URL: *http://dx.doi.org/10.1016/j.jairtraman.2021.102061*

Ma, C., Dai, G. and Zhou, J. (2022). Short-term traffic flow prediction for urban road sections based on time series analysis and $\mathrm{lstm}_b ilstmmethod, IEEETransactionsonIntelligentTransportationSystems\mathbf{23}(6) : 5615--5624$.

Ma, Q., Huang, G. H. and Ullah, S. (2020). A multi-parameter chaotic fusion approach for traffic flow forecasting, *IEEE Access* **8**: 222774–222781.

Mahdy, B., Abbas, H., Hassanein, H., Noureldin, A. and Abou-zeid, H. (2020). A clustering-driven approach to predict the traffic load of mobile networks for the analysis of base stations deployment, *Journal of Sensor and Actuator Networks* **9**: 53.

Qin, G., Niu, X., Wang, J. and Qian, Q. (2021). Traffic flow prediction based on janet and attention mechanism, *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 131–134.

Qu, L., Qie, L., Li, X., Liu, Z., Li, X. and Shi, Y. (2022). Time slot recurrent neural networks for short-term traffic flow prediction, *2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 265–271.

Salinas, D., Flunkert, V., Gasthaus, J. and Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks, *International Journal of Forecasting* **36**(3): 1181–1191.
URL: *http://dx.doi.org/10.1016/j.ijforecast.2019.07.001*

Santos Escriche, E., Vassaki, S. and Peters, G. (2023). A comparative study of cellular traffic prediction mechanisms, *Wireless Networks* **29**(5): 2371–2389.
URL: *http://dx.doi.org/10.1007/s11276-023-03313-9*

Trinh, N.-P., Tran, A.-K. N. and Do, T.-H. (2022). Traffic flow forecasting using multivariate time-series deep learning and distributed computing, *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 665–670.

Xie, H., Jin, F. and Li, H. (2021). Short term traffic flow prediction method based on fuzzy clustering algorithm, *2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 601–607.

Yu, L., Zhao, J., Gao, Y. and Lin, W. (2019). Short-term traffic flow prediction based on deep learning network, *2019 International Conference on Robots Intelligent System (ICRIS)*, pp. 466–469.

Zhang, L., Liu, W. and Feng, L. (2021). Short-term traffic flow prediction based on improved neural network with ga, *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 312–315.

Zhou, H. and Xu, P. (2022). Pango: Prediction model based on clustering of time series for traffic flow to venues, *2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pp. 21–25.