International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Decision-Tree-based Ensemble Learning Models for Long-Term Traffic Intensity Forecasting and Analysis of Congestion Treatment Strategies

Abdulrahim Shaikh
School of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India
1032192249@mitwpu.edu.in

Vaishali Jaiswal
Centre for Development of Advanced Computing
Pune, India
vaishalij@cdac.in

Sumedha Sirsikar
Faculty of Computer Engineering and Technology
Dr. Vishwanath Karad MIT World Peace University
Pune, India
sumedha.sirsikar@mitwpu.edu.in

Abstract—Traffic intensity forecasting is a key factor in analyzing traffic patterns and making recommendations to overcome congestion. It can also prove helpful to the Intelligent Transportation System (ITS) application. In this work, we have made a detailed comparative evaluation of various ML regression algorithms aimed at solving the problem of long-term traffic intensity prediction. A lot of work focuses mainly on traffic flow prediction. However, work on traffic intensity prediction has not been done sufficiently. For this problem, ensemble learning methods like Random Forest Regression that use the outputs of individual trees (Decision Trees) proved to be more successful and efficient rather than the single models approach. This work also dictates the study of various features that may be used to express the traffic data and the various strategies that can be employed to make decisions on whether a solution to overcome traffic congestion is needed.

Keywords-Traffic, Decision trees, ensemble learning, Random forest regression, K-NN, Gradient Boosting regression

I. Introduction

The increasing patterns of the population give rise to the increase in urbanization, economy, goods, services, etc. As a result, the problem of traffic management has arisen. These increasing patterns make traffic congestion worse with time and show no signs of waning. Congestion always results in numerous adverse effects such as elevated journey times, environmental pollution, intensified fuel usage, an increase in the number of road accidents, etc. Nevertheless, there are many ways by which traffic congestion might be weakened. For example, widening of the road, promotion of public transport, urban planning, road pricing, etc. However, to implement such an action, a proper understanding of traffic intensity is necessary. The Internet of Things (IoT) has been a primary impact in gathering traffic intensity data by storing data collected from a web of connected devices (sensors, cameras, GPS) that can interact with each other. The increasing utilization of IoT is anticipated to play an essential role in upgrading the efficiency of transportation systems and has the aptitude to revolutionize the methods, ideas, and strategies of extracting traffic intensity data. To manage and solve all these arising problems, the need for traffic intensity forecasting has emerged.

The data that represents traffic intensity entails non-linear relationships between the features. These features are also complexly related. Hence, an individual model may suffer to perceive all the gradations of the data. This work expresses the comparison between different regression algorithms and gives emphasis to the decision tree-based ensemble learning models like Random Forest Regression which are more suitable and favourable in solving the problem of traffic intensity forecasting and the strategies that may be implemented to finalize the decision of treatment of traffic congestion..

The following topics are covered in the remaining section of the paper: The review of recent fieldwork is contained in Section II. Section III holds the expression of traffic intensity followed by different representations of traffic data. Section IV holds the methodology of this research. The outcomes of the evaluated algorithms are covered in Section V. The study of various approaches is presented in Section VI to help determine whether or not a recommendation to reduce congestion is necessary. And the final section VII concludes the work.

II. Literature Review

Surely, advancements in the technological and IT fields have transmuted systems and various departments of the world to be more systematic, structured, ordered, well-planned, and well-managed.

Around the 2000s, wireless networking technologies called Vehicle-to-Vehicle (V2V) [1] and Vehicle-to-Everything (V2X) [1] were being developed. V2V standardly entails proximate-range wireless networking protocols of Dedicated Short-Range Communication (DSRC) or general cellular communication. This technology enables automobiles to interchange their location, direction, speed, possible route, and several other factors. Simultaneously, when the data is exchanged, it gets analyzed and presented before the user in order to allow him/her to make a decision on whether to change the current route. The main aim here was to allow automobiles to interchange live information and assess the conditions of traffic on the prevailing route.

Similarly, around the 2010s, another similar technology called Connected and Automated Vehicles (CAVs) [2] was in the works. This application employs sensors, cameras, and radios to engage with their surroundings and with nearby automobiles in order to make automated forecasts of traffic intensity on the current route. Several top-notch car manufacturers (such as Tesla, Volvo, Audi, and BMW) are working on installing the CAVs. The CAVs technology has the potential to transform how transportation can be upgraded. Therefore, a significant amount of research and development is being focused on this specific field, with many researchers and scientists predicting that it will have an intensifying key role in the future of automobiles and transportation.

Modern innovations like V2V, V2X, and CAVs, however, are only helpful for forecasting short-term traffic densities. In a different investigation, Sharma et al.'s[3] created a short-term traffic flow forecast model based on Artificial Neural Networks (ANN). An SVR (Support Vector Regression) model was suggested by Neto et al. primarily for estimating short-term traffic [4]. Later, Chan et al. proposed a better ANN model using a hybrid approach that incorporated exponential smoothing and the Levenberg-Marquardt method [5]. A short-term traffic flow forecasting technique using Bayesian networks was also proposed by Sun et al. [6]. It has even been done to apply probabilistic graphical models in short-term forecasting methods [7]. However, These applications cannot give out predictions of weeks, months, or years in advance. Hence, with the aim of getting long-term predictions for urban planning, a study of more advanced algorithms and technologies is necessary.

Numerous studies have been conducted in order to anticipate traffic flow. Though, traffic intensity and traffic flow are two distinct notions. Traffic flow alludes to the amount of vehicles that traverse a predetermined spot on a carriageway over a predetermined time period. Traffic intensity alludes to the amount of vehicles per unit time that traverse a predetermined point on a carriageway. To state another way, traffic flow pertains to the rate at which automobiles traverse through a predetermined checkpoint on a carriageway, while traffic intensity refers to the net amount of vehicles that traverse through that predetermined point per unit time. For traffic flow prediction, several more features like weather conditions, day of the week, speed, etc. are included unlike in traffic intensity prediction which does not need additional information. Yet both traffic flow and traffic intensity predictions are determined on the basis of past data.

In an effort to address the issue of traffic flow prediction, academics have looked into a wide range of approaches throughout the previous ten years. Kirby et al. recommended that though the efficiency of a model is of utmost significance, it should not be the only factor to stick to when selecting the right methodology for predictions [8].

Three typical modeling approaches—ARIMA (Auto-Regressive Integrated Moving Average) [9], ANN (Artificial Neural Networks) [5, 10], and Non-Parametric Regression (NPR) [11, 12, 13, 14]—were compared by Rong et al. in their study. ARIMA stands as a statistical analysis and predictive model, harnessing historical time series data to grasp and predict extended patterns [9]. It comprises three primary components: the AR (Auto-Regressive) module, the MA (Moving Average) module, and the differencing module. The auto-regressive module alludes to the utilization of past data points in the temporal series data to forecast future magnitudes. The MA module alludes to the utilization of previous bias errors in prediction to forecast future magnitudes. The differencing module alludes to the differencing of the temporal series to stabilize it, meaning that its statistical parameters remain fixed over time. The ARIMA method still remains the most popular method to employ in cases of both short-term and long-term traffic forecasting [15].

Yet, these are data-driven models and they generally fail to give out promising outcomes unless they are fed some decent amount of data. Luckily, ANN models tend to show more promising results in the forecasting of traffic predictions [16].

Specifically, DNN (Deep Neural Networks) and CNN (Convolutional Neural Networks) [16, 19] were used in the majority of situations to answer the challenge of predicting traffic intensity. In cases where the data contains numerous non-linearities, the Back Propagation technique seeks to capture these non-linear patterns [20]. Hence, BP-Neural Networks seemed accurate in some cases. It is further differentiated into two modules: forward propagation and backward propagation. A DNN-BTF [21] model that uses different types of temporal data (weekly, daily) to boost the efficiency of traffic flow forecasting was suggested by Wu et al [21]. The LSTM (Long Short-Term Memory) [17] based models have also been employed widely to tackle forecasting tasks related to temporal series. This LSTM-based approach has also shown promising results in terms of forecasting but it starts to fail when there is a huge amount of traffic data [16]. Following the growth in adopting LSTM-based [17] models, forecasting in complex and dynamic traffic conditions was not given much attention. Later, a model called Gated Recurrent Units Neural Network (GRU-NN) [17] was put forward as a modification of existing LSTM by Cho et al. [22]. In order to improve traffic flow forecast, Wali et al. [23] developed an approach called SSGRU that concentrated on particular road segments. This method was more efficient than the then-existing GRU-NN and LSTM models. Likewise, in the following years, a lot of work and research was carried out by tuning and modifying these GRU and LSTM models, which indicated promising outcomes. Table 1 shows the comparison of RMSE values of existing Deep Learning models. Root Mean Square Error (RMSE) [24] is calculated as:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i}^{m}(h(x^{(i)}) - y^{(i)})^2} \quad (1)$$

where m denotes the number of instances, h indicates the hypothesis, $x^{(i)}$ denotes the array of all features and $y^{(i)}$indicates the label value [24].

TABLE I. COMPARISON OF ALGORITHMS (EXISTING WORK IN DEEP LEARNING)

Model	*RMSE*
SVR	7.02
CNN	9.34
CNN-GRU	9.09
CNN-LSTM	9.75
ARIMA	9.515
LSTM	11.14
GRU-NN	11.15
BP-NN	18

III. EXPRESSION OF TRAFFIC

A. *Traffic Flow Data*

As previously defined, traffic flow alludes to the locomotion of automobiles on a particular carriageway. Consequently, traffic flow data is expressed in more complex terms like traffic volume, traffic density, traffic flow rate, Level of Service (LoS), etc. LoS being a non-quantitative non-numerical scale of traffic congestion conditions, cannot be considered for the problem of traffic forecasting. Traffic flow accounts for interplay and mutual influence between individual automobiles on the carriageway. It is impacted by factors such as speed, density, the behaviour of the driver on the carriageway, etc.

B. *Traffic Intensity Data*

Whereas, traffic intensity alludes to the degree or quantity of traffic on a specific carriageway. Consequently, traffic intensity data is expressed in more detailed and quantitative ways such as Passenger Car Units (PCU), Vehicles per hour (VPH) / Vehicles per day (VPD), etc. PCU indicates the amount of automobiles in motion on a carriageway at a predetermined point of time. It accounts for the changing effect of the several kinds of automobiles in the traffic, by representing them with their equivalent capacity in passenger cars. For instance, a lorry can be counted as 2 or 3 PCUs, while a single car counted as 1 PCU. Using PCU rather than VPH/VPD makes more sense as it gives out a more efficient depiction of actual traffic intensity. Another important reason is that by using PCUs, analyses of different kinds of roads are obtained, which proves to be more helpful for the current domain.

C. *Data Representation*

In the context of using PCUs as a representation of traffic intensity, a dataset was created as a subset of a replication of real data collected from sensors on different junctions/carriageways at different points in time. The generated dataset (in comma-separated values format) consists of three series, namely date-time, PCU, and junction (name of the junction). Hence, a single instance or record of the dataset represents the PCU amount at the specified junction at the specified point in time. For example

TABLE II. REPRESENTATION OF CSV DATASET AND ITS FEATURES

DateTime	Junction	PCU
01-07-2023 00:00	junction_name	13
02-07-2023 00:00	junction_name	12
03-07-2023 00:00	junction_name	18
04-07-2023 00:00	junction_name	18
05-07-2023 00:00	junction_name	20
06-07-2023 00:00	junction_name	18

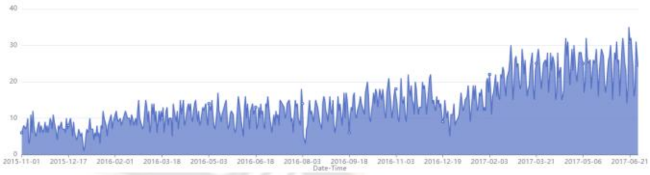

Figure 1. PCU vs DateTime (PCU values over date-time)

IV. METHODOLOGY

Ensemble learning alludes to an ML technique that combines multiple individual models to tune the efficiency and robustness of the forecasts. The basic ideology of ensemble learning is that by blending two or more individual models that vary in their presumptions, parameters, or features, the assessed/resultant forecast can be much more accurate and considerable than the forecast of an individual model.

Despite the fact that Deep Learning proves to be a more favourable and powerful method for several applications along with traffic forecasting, ensemble learning has several more pluses over deep learning for this particular problem. Ensemble learning is known to perform well with scanty data, whereas deep learning models need large quantities of data in order to escape overfitting. Ensemble learning models can prove to be more robust to exceptions and anomalies. Ensemble learning models are also known to be computationally faster and more efficient than deep learning models, especially for smaller and scanty datasets. Therefore ensemble learning tends towards being a proper choice for solving the problem of traffic intensity prediction.

In the context of temporal data, ensemble learning can help identify and solve the exceptional challenges regarding this kind of data. The dataset created contains complex patterns and relations that can cause overfitting if an individual model is used. Ensemble learning reduces this minus risk by integrating two or more models having different results, features, or parameters. Thus, pulling down the risk of overfitting and enhancing the generalization performance. The dataset also exhibits dynamic non-uniform behaviour like certain trends, sudden changes in the data, seasonality, etc which can be handled well by ensemble learning models. Overall, ensemble learning can turn out to be a powerful technique for upgrading the accuracy and robustness of temporal-series forecasting by blending multiple individual models that assess the various aspects and different problems concerned with this type of data.

Coming to Decision Trees, a decision tree model is an ML algorithm, that comprises a tree-like structure where the branches represent the decisions and the leaves represent the forecasted assumption. The decision tree is established by recursively splitting the dataset into smaller slices. At every split, the decision tree algorithm chooses the feature with the largest information gain or reduction in noise. The goal is to create a tree that is capable of accurately predicting the target feature for fresh data points. Ensemble learning based on decision trees is a widely used and effective approach that utilizes the aforementioned decision trees as the foundational models. The concept underlying this method involves creating an ensemble of decision trees with diverse parameters and structures, followed by combining their forecasts to yield a conclusive prediction. Here, each tree in this ensemble is trained using various input data slices, to create a cluster of diverse models.

Random forest regression constitutes a specific variant of ensemble learning based on decision trees. It employs multiple decision trees to predict continuous numeric values. The final forecast is obtained by taking the mean value of the forecasts from all the decision trees formed. The random forest regression technique constructs many decision trees on the basis of various subsets of the training portion of the data and a subset of the input features.

$$y = \frac{\sum_N^{i=1} f_d\left(i\right)}{N} \qquad (2)$$

Here, y represents the ultimate forecasted output, N represents the count of decision trees generated by the Random forest regressor, and $f_d(i)$ symbolizes the prediction from the individual decision tree.

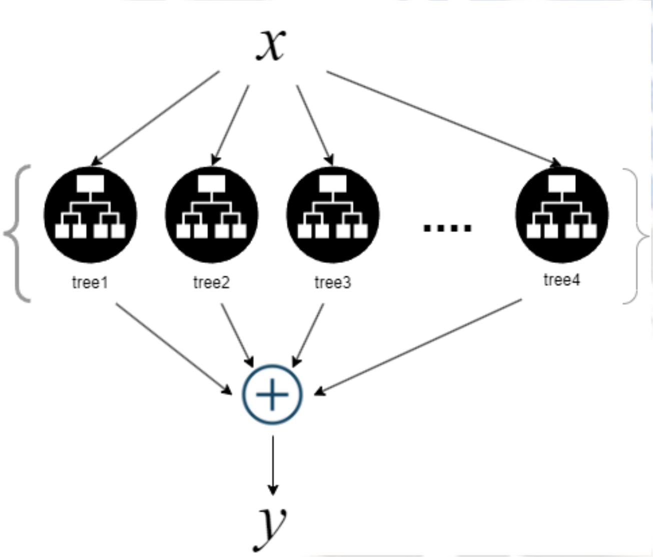

Figure 2. Random Forest Regression

A randomly selected subset of the training data section (also referred to as the train-dev set) and a randomly selected subset of input characteristics are used to train each decision tree that is created. Hence, the decision trees generated have very less correlation between them, which can reduce the risk of overfitting.

V. RESULTS

The dataset was trained with 10 different regression algorithms, namely, Linear Regression, Lars Regression, ARD (Automatic Relevance Determination) Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Decision Tree Regression, K-Nearest-Neighbours (K-NN) Regression, Gradient Boosting Regression, and Random Forest Regression. As shown in Table 3, it is clear that the decision tree-based ensemble learning model of RFR outperforms all other individual models. Above all, the Random Forest Regression proved to be dominant in terms of efficiency.

TABLE III. COMPARISON OF ALGORITHMS (RESULTS)

Algorithm	Accuracy Score
Other individual models	~0.63
KNN-Regression	0.78
GB Regression	0.79
Decision Tree Regression	0.904
RF Regression	0.96

In general, a Random Forest regressor can be a good choice if the data has trends, patterns and correlations that can be identified by the regressor. The traffic data has non-linear yet complex relationships between the predictors and the dependent variable which the Random Forest regressor can capture and then identify the ideal predictor that contributes to the prediction.

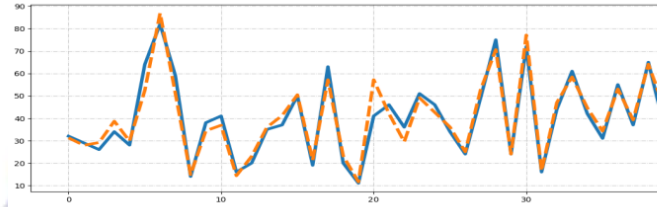

Figure 3. Difference between actual and predicted values of the test set (blue - predicted, yellow - actual)

Other than the ensemble models and the decision tree regression models trained, the K-Nearest Neighbors (K-NN) regressor also gave out decent results. From this, we can observe a similarity/relationship between the working of random forests and the K-NN algorithms as pointed out by Lin and Jeon [25s]. These models make forecasts on the basis of the neighbourhood of the points. According to Lin & Jeon, the entire random forest functions as a weighted neighborhood system, where the weights are determined by the average of the individual trees. The training dataset's layout determines the tree structure, which in turn affects the vicinity of the prediction points.

Lastly, using the RFR algorithm, predictions in terms of weeks were made which showed good results with a correlation between the trend of existing data and trend in the forecasts.

VI. RECOMMENDATION STRATEGIES

After obtaining predictions, comes the important task to decide whether a recommendation to improve the congestion is necessary. After thorough discussion, two ways to finalize congestion treatment recommendations were considered.

A. Method 1

Defining a constant for ideal relative change and then calculating a relative change trend line that stands out to be the limit of the PCU capacity that the specified junction can have. If a defined percentage of forecasts crosses this line trend, it can be assumed that urban planning/treatment to overcome congestion for the predicted date-time is recommended.

B. Method 2 (preferred)

Beforehand define a maximum PCU value for the junction and calculate a threshold PCU value that is 70% of the max PCU value. If the percentage of forecasts that cross this threshold value is greater than or equal to 60%, it can be finalized that urban planning/congestion treatment is needed. Figure 4 shows the representation of the maximum capacity (18) line and threshold (13) line.

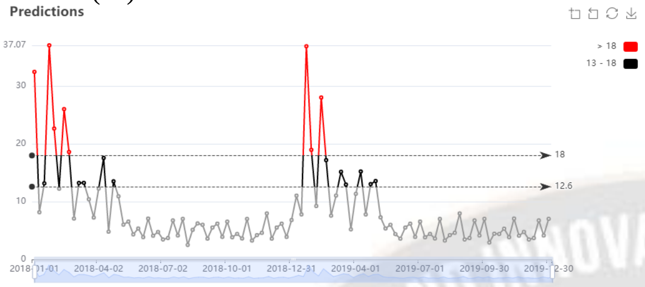

Figure 4. Predictions with the maximum PCU capacity and the threshold value

VII. Conclusion and Future Works

In conclusion, there are two independent standards for predicting traffic: flow prediction and intensity prediction. Both norms are complicated yet compelling tasks. Diverse modeling techniques are available for predicting both traffic flow and traffic intensity. However, the selection of a suitable model relies on the specific characteristics of the data.

This study demonstrates that decision tree-based ensemble learning models, such as Random Forest Regression, can also present a viable solution for addressing the challenge of traffic intensity prediction. This is attributed to their capability to capture complex patterns and correlations within the data's features, and their resilience in handling outliers and noisy data—issues frequently encountered in traffic data analysis. Following the assessment of multiple algorithms, it was noted that the K-NN algorithm exhibited superior performance, thereby validating the connection and resemblance between the random forests and the K-NN algorithms.

However, it is still significant to meticulously prepare the data, choose the right attributes, and properly tune the model’s execution and effectiveness. Finally, two strategies that aim to calculate the trigger limit for a recommendation of traffic congestion treatment were discussed.

In future, advanced technologies like big data analytics are expected to be integrated with traffic prediction. Extensive research is being dedicated to traffic flow prediction. Also, in the case of traffic intensity prediction, there exists a scarcity of available data. Hence, traffic intensity prediction has now become a task to be looked at more.

References

[1] Qing X, Tony M, Raja S. Vehicle-to-Vehicle Safety Messaging in DSRC. VANET. 2004.

[2] Steven ES. Connected and automated vehicle systems: Introduction and overview. Journal of Intelligent Transportation Systems, 2018.

[3] Sharma B, Kumar S, Tiwari P, Yadav P, Nezhurina MI. ANN based short-term traffic fow forecasting in undivided two lane highway. J Big Data. 2018.

[4] Castro-Neto M, Jeong YS, Jeong MK, et al. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Systems with Applications. 2009

[5] Chan KY, Dillon TS, Singh J, et al. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. IEEE Trans. Intelligent Transportation Systems. 2012.

[6] Sun S, Zhang C, Yu G. A Bayesian network approach to traffic flow forecasting. IEEE Trans. Intelligent Transportation Systems. 2006.

[7] Lippi M, Bertini M, Frasconi P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. IEEE Trans. Intelligent Transportation Systems, 2013.

[8] Kirby H, Watson S and Dougherty M. Should we use neural networks or statistical models for short-term motorway traffic forecasting? Int J Forecasting 1997

[9] Williams BM, Hoel LA. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. Journal of Transportation Engineering. 2003.

[10] Vlahogianni EI, Karlaftis MG, Golias JC. Short-term traffic forecasting: where we are and where we’re going. Transportation Research Part C: Emerging Technologies. 2014.

[11] Hardle W. Applied nonparametric regression. Cambridge: Cambridge University Press, 1990.

[12] William H, Tang Y and Tam M. Comparison of two non-parametric models for daily traffic forecasting in Hong Kong. J Forecasting 2006.

[13] Smith B, Williams B and Oswald R. Comparison of parametric and nonparametric models for traffic flow forecasting. Transport Res C: Emer 2002.

[14] Haworth J and Cheng T. Non-parametric regression for space–time forecasting under missing data. Computers, Environment and Urban Systems 2012.

[15] Yaping R, Xingchen Z, Xuesong F, Tin-kin H, Wei W, and Dejie X. Comparative analysis for traffic flow forecasting models with real-life data in Beijing. Advances in Mechanical Engineering 2015.

[16] Noor AMR, Nuraini S, Khairul KI, Suzaimah R, Mohd FMA, and Sazali S. Gap Techniques and evaluation: traffic flow prediction using machine learning and deep learning. Journal of Big Data 2021

[17] Rui F, Zuo Z, and Li Li. Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction. 31st YAC. 2016

[18] Nicholas G. Polson, Vadim O. Sokolov. Deep Learning for Short-Term Traffic Flow Prediction. Transportation Research Part C: Emerging Technologies, 2017.

[19] Hongsuk Y, HeeJin J, Sanghoon B. Deep Neural Networks for Traffic Flow Prediction. IEEE International Conference on Big Data and Smart Computing (BigComp). 2017.

[20] Yisheng L, Yanjie D, Wenwen K, Zhengxi L, and Fei-Yue W. Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems. 2015

[21] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang. A hybrid deep learning based traffic flow prediction method and its understanding. Transportation Research Part C: Emerging Technologies. 2018.

[22] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches. Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. 2014.

[23] Elleuch W, Wali A, Alimi AM. Neural congestion prediction system for trip modelling in heterogeneous spatio-temporal patterns. Int J Systems Science. 2020.

[24] Aurelion G. Hands on Machine Learning with Scikit-Learn, Keras & TensorFlow. O’Reilly Media Inc, 2017

[25] Yi L and Yongho J. Random Forests and Adaptive Nearest Neighbors. Journal of American Statistical Association. 2006.