Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://jaida.ikcu.edu.tr

Vol. 1, No. 2, pp. 136-143, 2021
Research Article

Traffic Density Estimation using Machine Learning Methods

Sümeyye AYDIN [1,*], Murat TAŞYÜREK [1], Celal ÖZTÜRK [2]

[1*] Kayseri University, Engineering, Architecture And Design Faculty, Computer Engineering Department, Turkey
[2] Erciyes University, Engineering Faculty, Computer Engineering Department, Turkey

Abstract
In cities where population density is high and transportation systems are widely used, it is necessary to manage traffic systems more effectively not to affect the daily planned works. The Intelligent Transportation System (AUS) is expressed as a system that provides users with better information and safer, more coordinated, and smarter use of transportation networks with different transportation modes and traffic management. One of the most important components of AUS models is the determination of traffic density. The traffic density of intersections is a difficult problem as it affects other interconnected intersections and varies in time. Deep learning method is a widely used method in traffic density estimation in recent years. In this study, the long-term short memory network (LSTM) model, one of the deep learning methods, is proposed to estimate the traffic density of a certain region using open data of Istanbul Metropolitan Municipality. The performance of the proposed LSTM-based model is compared with machine learning methods such as linear regression, decision tree, random forest, and the classical deep learning method (DL). Experimental evaluations show that the proposed LSTM method is more successful in traffic density estimation than the compared methods.
Keywords: Deep learning, Intelligent transportation systems; Istanbul traffic density prediction; LSTM, Machine Learning.

1. Introduction

Intelligent Transportation Systems (ITS) technologies improve transportation safety and mobility and increase productivity by integrating advanced communication technologies into transportation infrastructure and vehicles. One of the most important components of ITS technologies is the estimation of traffic density. Traffic density is expressed as the traffic demand exceeding the transportation capacity. The traffic density of one intersection is a difficult problem as it also affects other interconnected intersections and varies in time. Traffic density is a serious problem that affects the planning of daily life and is becoming increasingly common all over the world. This paper proposes to estimate the intersection density in a particular region by using LSTM-based deep learning method with open data of Istanbul Metropolitan Municipality (IMM).

Deep learning methods have recently been widely used to solve many problems such as natural language processing and real-time object detection [1-3]. There are many different deep learning models in the literature, depending on the data set and the type of problem. Some deep learning methods are insufficient in time-dependent processes. Recurrent Neural Networks (RNN) are successful in datasets containing time-dependent relationships [4]. The RNN model was first introduced in the 1980s [5]. However, traditional RNN methods face the problem of capturing long-term dependencies in the input data. A long short-term memory network (LSTM) has been proposed to solve this problem [6]. Ma et al. have applied the LSTM model for estimating traffic speed with remote micro software sensor data [7]. Tian et al. proposed the LSTM RNN model for traffic flow prediction [8] and found that the LSTM RNN method outperformed most of the non-parametric models [8]. Li et al. evaluated the performance of the LSTM and GRU model for traffic flow prediction [9]. A better performance can be obtained with the LSTM model in traffic flow forecasting, as LSTM can automatically calculate optimal time delays and capture features of longer time interval time series [8].

In this study, it is proposed to estimate traffic density using LSTM-based deep learning method using IMM open transportation data. Istanbul is the city with the largest population density and the most significant traffic problem in Turkey [10,11]. The proposed method is aimed to predict the traffic density of certain intersections during the time periods when the traffic density is high, such as commuting and rush hours. The performance of the proposed LSTM-based model is compared with machine learning methods such as linear regression, decision tree, random forest, and the classical deep learning method (DL) on actual observed data.

*Corresponding author e-mail address: sumeyyeaydin@kayseri.edu.tr

Received: 29/11/2021; Accepted: 21/12/2021

2. Methods

2.1. Linear regression method

Linear regression aims to establish a linear relationship between one or more independent variables and a dependent variable or a numerical result. This method estimates the value according to the independent variables. In simple linear regression, the target value can be estimated with one independent variable, while in multiple regression, the target is estimated by two or more independent variables. A multiple linear regression model was used because the number of independent variables in the traffic dataset was more than one.

2.2. Decision tree

Decision trees are a hierarchical machine learning method consisting of nodes, branches, and leaves. The output of the decision tree is in the form of a tree or rules, so it is easy to understand. The tree expands gradually according to the questions asked to the root node, and the growth is completed when the last leaves are formed. It is divided into regression trees and classification trees. If the dependent variable is categorical, classification trees are used, and if the dependent variable is continuous, regression trees are used. In this article, regression trees were used because it works with constantly changing data.

2.3 Random forest

Random forest algorithm is a machine learning method used to solve regression and classification problems[12]. The algorithm is an ensemble learning algorithm consisting of the output of multiple decision trees. Each node branches by choosing the best among the randomly selected variables in the node [13]. It is not pruned as in decision trees, so it is fast. In this study, a forest was created using 30 decision trees.

2.4. Deep learning

Deep learning is a subset of machine learning. It has many neural nodes. An artificial neural network is formed by the combination of nerve cells. A neural network is called ‘deep’ when it has more than one hidden layer[14].This design was inspired by the biological structure of the human brain. So in deep learning, there can be many layers between the input and output layers. These layers are called hidden layers. In the model consisting of consecutive layers, the previous layer's output is the input of the next layer[15]. Deep learning has been a widely used prediction method in recent years.

2.5. Long short-term memory (LSTM)

The LSTM deep learning algorithm is known as a recurrent neural network introduced by Hochreiter and Schmidhuber in 1997 to eliminate the disadvantages of the RNN architecture [16]. The most important feature of the LSTM network is that it also evaluates the background information about the network. The basic LSTM diagram is given in Figure 1 [17, 18].

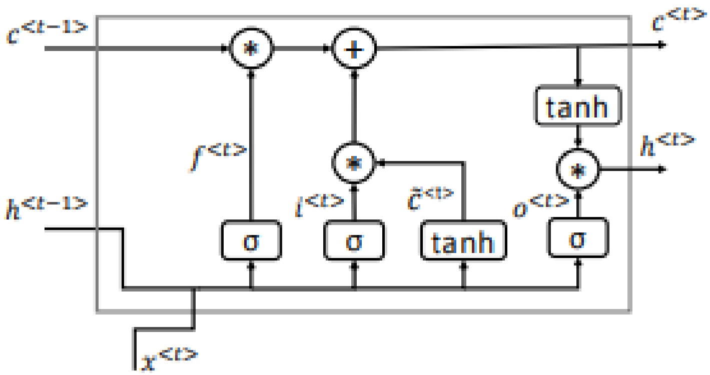

Figure 1. *Basic LSTM block diagram.*

LSTM networks address the problem of vanishing gradients of the RNN model by dividing into three inner cell gates, creating pseudo-memory cells to store information in a long-range context [19]. A typical LSTM neural network cell is basically configured with four gates, namely the input gate, the input modulation gate, the forget gate, and the output gate: The cell is controlled by gates and can keep the value or reset the value according to the state of the gate. In particular, three gates are used to control whether the current cell value will be forgotten (forget gate f_t), whether it will read its input (input gate i_t), and output (output gate o_t) the new cell value; In addition, there is an input modulation gate called $\bar{c}_t$. The LSTM gates and their operations are presented in Eqs. (1)-(6).

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{1}$$

$$\mathrm{i}_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2}$$

$$\bar{\mathrm{c}}_t = tanh(W_{xc}x_t + W_{ch}h_{t-1} + b_c) \tag{3}$$

$$\mathrm{c}_t = f_t c_{t-1} + i_t\bar{\mathrm{c}}_t) \tag{4}$$

$$\mathrm{o}_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{5}$$

$$\mathrm{h}_t = \mathrm{o}_t * \tanh(\mathrm{c}_t) \tag{6}$$

The gateway receives a new entry point from the outside and processes the newly incoming data. The memory cell input gate receives input from the output of the LSTM neural network cell in the last iteration. The forget gate decides when to forget the output results and thus selects the optimal time delay for the input sequence. The output port receives all the calculated results and generates the output for the LSTM neural network cell [6]. LSTM models redesign computing nodes based on the structure of the RNN network.

3. Experimental Evaluation

3.1. Dataset

In this study, traffic density data of August 2020 belonging to IMM were used [20]. IMM open dataset includes time information, locations of intersections, number of vehicles based on intersections, average speed, maximum speed and minimum speed of these vehicles. The dataset used in this study contains 1.048.575 rows of records. The spatial locations of the intersections used in this study are shown in Figure 2 and the summary information of the dataset is presented in Table 1.

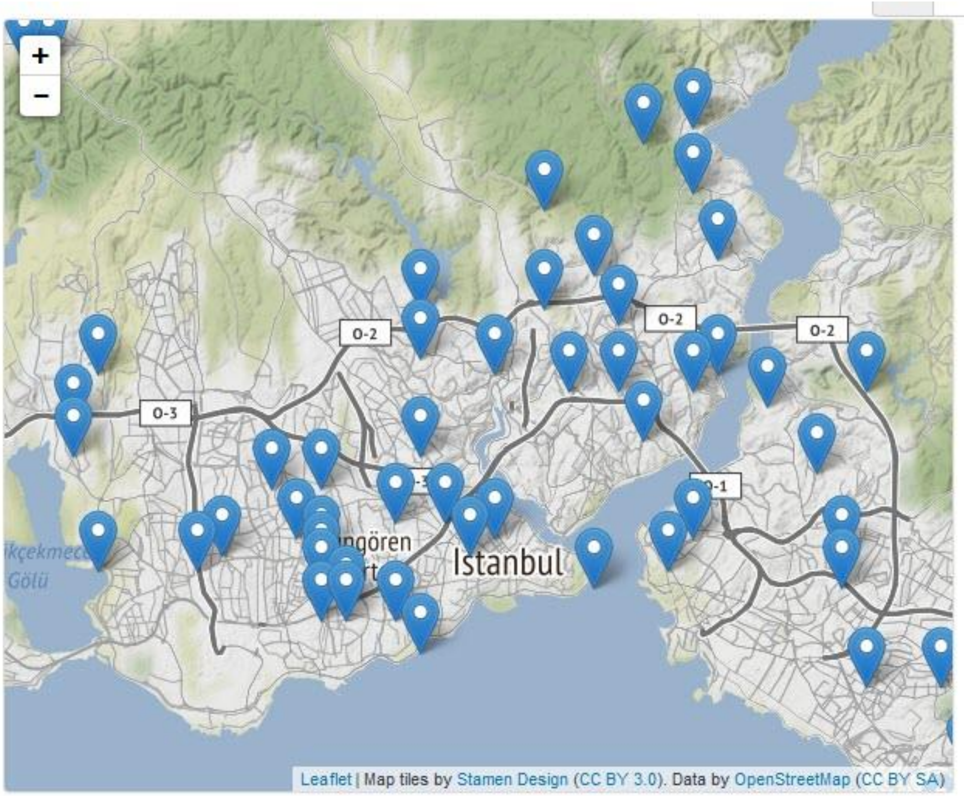

Figure 2. *Location of target stations in the used dataset*

When the IMM open dataset, the general information of which is presented in Table 1, is examined in detail, an increase was observed in the number of vehicles, especially during the commuting and exiting hours, while a decrease was observed during the night hours. This paper shows that the number of vehicles in traffic changes over time. Since the traffic density changes according to time, the data were sorted on the basis of date and time before the training process. In addition, the clock data is added to the dataset as an input. In determining the number of vehicles in traffic, it is not sufficient to know only the vehicle number data of a certain time before the relevant station. An intersection is also affected by the density of intersections close to it. An intersection usually consists of 4 branches, so it is important to know the four intersections close to each intersection. Therefore, giving the number of vehicles in four intersections close to the target intersection where the traffic density will be estimated as an input to the LSTM model will strengthen the prediction ability. For this reason, the 4 intersections that are spatially closest to each intersection are determined by using the Euclidean distance as the spatial distance. After determining the four intersections, which are the closest neighbours of each intersection, the number of vehicles in the same date and time zone in these neighbouring intersections is used as input data for the intersection whose density will be estimated.

Table 1. *Data Dictionary.*

Column	Type	Tag	Definition
DATE_TIME	text	Date	It is the field that contains the date and time information. The data format is YYYY-MM-DD HH24:MI:SS. * Date breakdown is hourly.
LONGITUDE	text	Longitude	It is the field that contains the longitude information.
LATITUDE	text	Latitude	It is the field that contains the latitude information.
GEOHASH	text	geohash	Geohash Value of Latitudes and Longitudes
MINIMUM_SPEED	text	Minimum Speed	Minimum speed (km/h) in the geohash area at the relevant hour.
MAXIMUM_SPEED	text	Maximum speed	Maximum speed (km/h) in the geohash area at the relevant hour.
AVERAGE_SPEED	text	Average Speed	Average speed (km/h) in the geohash area at the relevant hour.
NUMBER_OF_VEHICLES	text	Number of Individual Vehicles	Number of different vehicles in the geohash area at the relevant hour.

3.2 Model Settings

By working on the number of layers in the LSTM model, the most suitable LSTM input layer was determined to be one output layer and 20 neurons. The structure of the LSTM model used in the study is given in Figure 3.

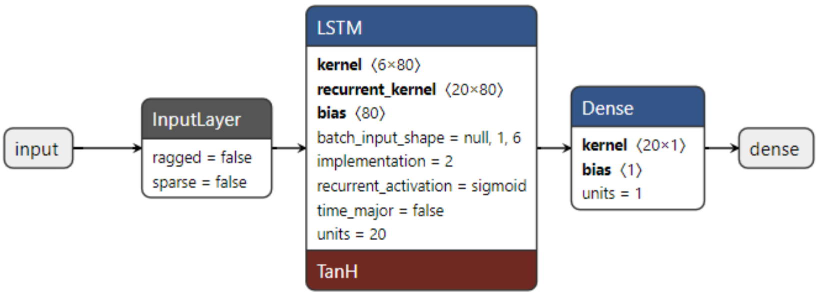

Figure 3. *LSTM model structure in the study.*

In the study with the LSTM method, the Adam algorithm was used as the optimization method. Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update iterative Evaluation Metrics mesh weights based on training data.

3.3 Evaluation Metrics

In this study, MAE and RMSE measurements were used as evaluation metrics for performance comparison. The study was carried out by normalizing the data, in order not to make an incorrect comparison, the data was converted to real values by reverse scaling and evaluated.

The MAE is the mean of the absolute error between the predicted value and the true value. It ranges from zero to infinity, it is expected to give a low result. It is a linear value. The MAE formula is shown in Eq. (7). In the formula, p_i represents the predicted value and $o_i$original value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \tag{7}$$

The RMSE is the root mean square error between the predicted value and the true value. It ranges from zero to infinity, it is expected to give a low result. The RMSE formula is shown in Eq. (8). In the formula, p_i represents the predicted value and $o_i$original value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \quad (8)$$

3. Experiment Results

Experimental evaluations of linear regression, decision trees, random forest, classical deep learning, and LSTM methods were made using real data from IMM. RMSE and MAE values were used as evaluation criteria to examine the performance of the methods. The station information and the values of the methods that were experimentally evaluated are presented in Table 2 and Table 3.

When the experimental evaluation results (Table 2) created according to the MAE error values are examined, the LSTM method generally gave lower error results than the other compared methods.

Table 2. *MAE Values of the Methods.*

İstasyon Geohash	Decision Tree	Deep Learning	Linear Regression	LSTM	Random Forest
sxk3q9	5,325	3,799	4,237	3,816	**3,773**
sxk3r9	25,678	29,115	19,625	**15,152**	16,536
sxk3rf	12,864	12,387	10,574	9,96	**9,878**
sxk3w2	5,019	4,883	4,315	**4,282**	4,431
sxk3w6	39,309	32,604	25,265	**23,649**	24,429
sxk90y	17,365	16,137	14,353	**13,251**	13,531
sxk91n	5,837	5,311	5,112	4,745	**4,702**
sxk91u	9,289	7,3	7,637	**7,111**	7,493
sxk91w	8,821	6,984	6,386	**5,317**	6,939
sxk92b	10,835	**8,438**	9,205	8,753	8,764
sxk92c	7,1	6,053	5,633	**5,582**	5,998
sxk92e	7,117	**4,547**	4,592	4,606	4,757
sxk92f	8,301	6,155	6,166	**5,71**	6,012
sxk92q	8,926	8,218	8,231	**7,65**	7,739
sxk93s	13,367	13,304	11,497	**8,803**	11,49
sxk966	8,387	5,696	5,752	**5,532**	5,948
sxk96e	6,952	5,193	5,444	**5,118**	5,52
sxk96h	24,372	24,022	19,673	**17,577**	21,22
sxk99b	9,475	**7,279**	8,53	7,291	7,449
sxk99y	10,687	8,745	7,205	**6,225**	7,664

Table 3 shows statistics on the RMSE values of the methods.

Table 3. *RMSE Values of the Methods.*

Station Geohash	Decision Tree	Deep Learning	Linear Regression	LSTM	Random Forest
sxk3q9	7,429	**4,748**	5,135	4,797	4,778
sxk3r9	35,96	37,584	24,058	20,58	**20,449**
sxk3rf	17,084	16,447	13,687	13,353	**13,093**
sxk3w2	6,405	6,165	**5,317**	5,405	5,471
sxk3w6	54,811	41,776	32,909	31,34	**30,639**
sxk90y	21,559	20,193	18,814	**16,812**	17,948
sxk91n	7,658	**5,763**	6,465	6,074	5,978
sxk91u	11,643	**8,658**	9,506	8,872	9,303
sxk91w	10,953	**7,07**	7,897	7,164	8,358
sxk92b	13,44	**10,741**	11,763	11,321	11,339
sxk92c	9,411	7,757	7,387	**7,358**	7,92

sxk92e	8,838	6,276	**5,948**	6,16	5,954
sxk92f	10,008	**7,234**	7,844	7,468	7,643
sxk92q	12,272	**9,66**	10,427	9,777	10,162
sxk93s	17,194	16,429	14,337	**13,007**	14,445
sxk966	11,249	**6,577**	7,471	7,261	7,95
sxk96e	8,927	7,934	7,067	**6,706**	7,309
sxk96h	30,717	31,072	23,533	**21,894**	26,93
sxk99b	12,262	9,885	10,876	9,834	**9,803**
sxk99y	14,047	14,418	8,935	**8,205**	9,734

When the result values were examined (Table 2 and Table 3), the LSTM method made predictions with lower error rates compared to other methods. In addition, comparison of MAE values with statistical data is presented in Table 4. Standard deviation value of the LSTM method gave better results compared to other methods (Table 4). This shows that the LSTM method works more stable.

Table 4. *Comparison of MAE error values.*

Method	Total	Avg	Best	Worst	Std.
Decision Tree	245,03	12,25	5,02	39,31	8,53
Deep Leaning	216,17	10,81	3,80	32,60	8,38
Linear Regression	189,43	9,47	4,24	25,27	5,89
LSTM	170,13	8,51	3,82	23,65	5,16
Random Forest	184,27	9,21	3,77	24,43	5,65

In Figure 4, the success of LSTM is seen with the low total RMSE value in the traffic density of the stations.

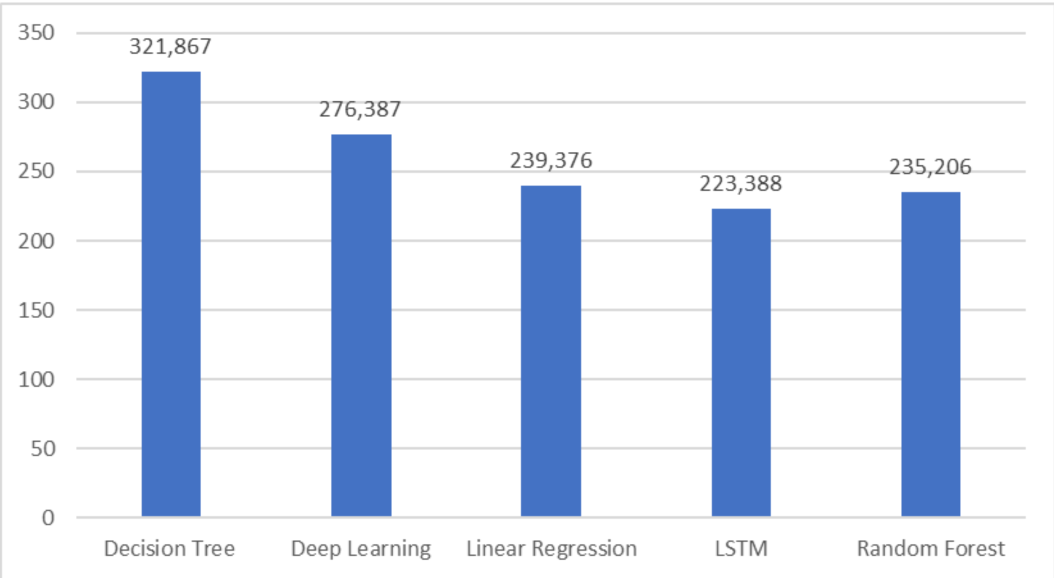

Figure 4. *Total RMSE values of the method.*

Figure 5 shows the actual number of vehicles and the number of vehicles estimated by decision trees, deep learning, linear regression, LSTM and random forest methods in the sxk99y geohash station. Comparison of methods in the sxk99y geohash station is given in figure 5.The graph with the highest overlap between the predicted value and the actual value belongs to the study using the LSTM method.

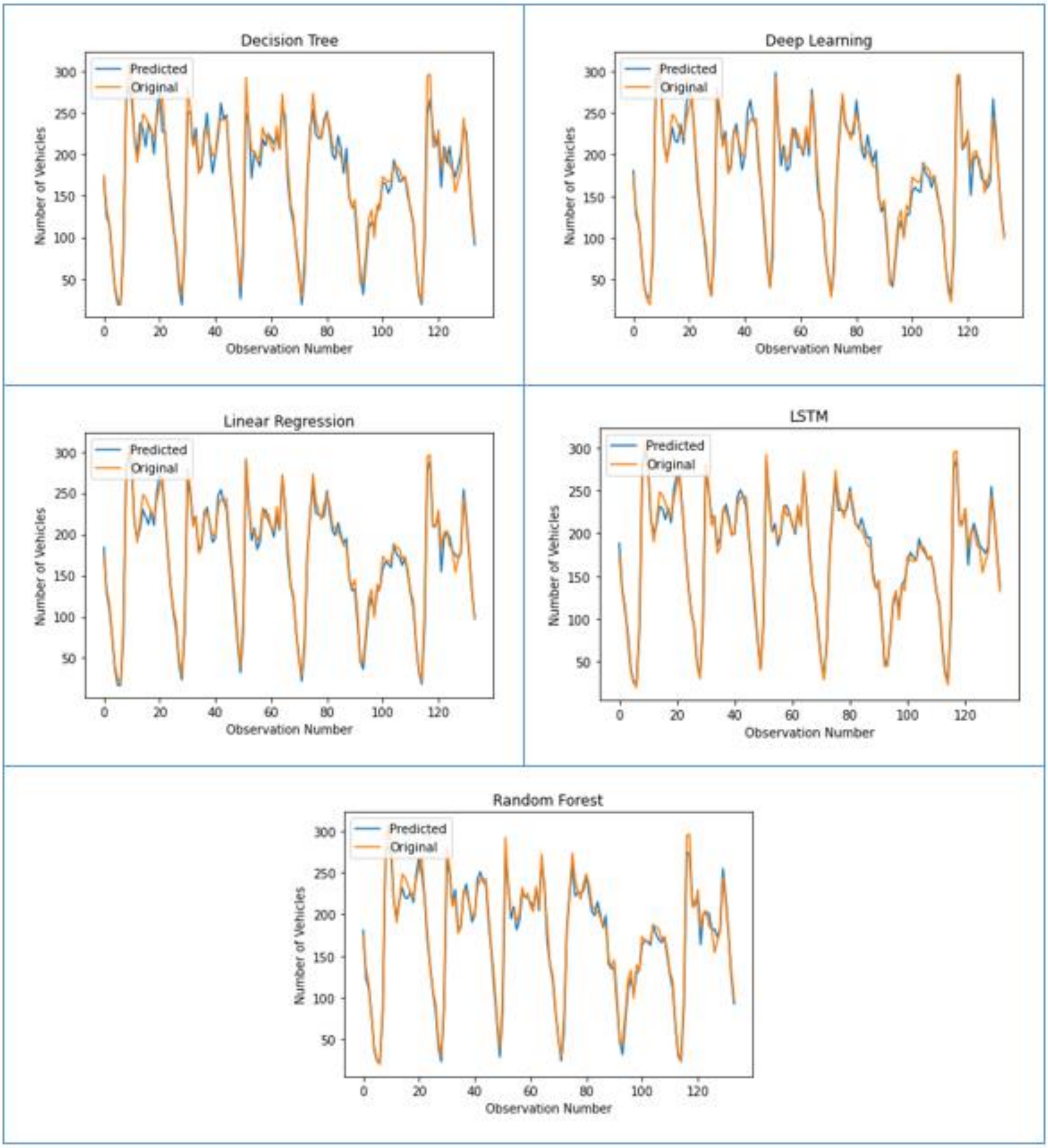

Figure 5. *Comparison of methods test and prediction data for station sxk3r9.*

Conclusion

The traffic problem takes a large part of people's time in last years. In addition, it is very important to manage traffic systems in places with high population density. Intelligent Transportation Systems (ITS) are used to manage traffic systems to overcome the traffic density problem. One of the most important components of AUS models is that traffic density can be determinated in advance. For this purpose, a model has been proposed to predict the number of vehicles in the traffic by working on the data of the city of Istanbul, which is the most crowded and has the most traffic in Turkey.

LSTM model, which is a deep learning approach, is proposed for traffic density estimation. Since LSTM is able to automatically calculate optimal time delays and capture the characteristics of longer time-lapse time series, it has achieved higher performance in traffic flow forecasting. Experimental evaluations have shown that LSTM-based methods are more successful than linear regression, decision trees and random forest, and the classical deep learning methods.

Declaration of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

An earlier version of this paper was presented at the ICADA 2021 Conference and was published in its Proceedings (Title of the conference paper: "Uzun Süreli Kısa Bellek Ağı Yöntemi ile Trafik Yoğunluk Tahmini").

We would like to thank IMM for sharing the hourly traffic density data used in this study.

References

[1] D. Ravì, C. Wong, F. Deligianni, M. Berthelot; J. Andreu-Perez, B. Lo and G. Yanget, "Deep Learning for Health Informatics," in IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4-21, doi: 10.1109/JBHI.2016.2636665, Jan. 2017

[2] T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," in IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55-75, doi: 10.1109/MCI.2018.2840738, Aug. 2018.

[3] A. Şeker, B. Diri ve H. H. Balık, "Derin Öğrenme Yöntemleri Ve Uygulamaları Hakkında Bir İnceleme", Gazi Mühendislik Bilimleri Dergisi (GMBD), c. 3, sayı. 3, ss. 47-64, Aralık 2017

[4] A. J. P. Samarawickrama, T. G. I. Fernando, "A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market," 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), pp. 1-6, doi: 10.1109/ICIINFS.2017.8300345, 2017

[5] N. Buduma and N. Locascio, Fundamentals of Deep Learning. Designing Next-Generation Machine Intelligence Algorithms, O'Reilly Media, 172-217, 2017.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[7] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data" Transportation Research Part C: Emerging Technologies, 54, pp. 187–197, 2015.

[8] Y. X. Tian and P. Li, "Predicting Short-term Traffic Flow by Long Short Term Memory Recurrent Neural Network", 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), pp. 153-158, 2015.

[9] R. Fu, Z. Zhang and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction", Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 324-328, 2016.

[10] TUİK, Adrese Dayalı Nüfus Kayıt Sistemi Sonuçları 2013, Türkiye İstatistik Kurumu Haber Bülteni, Sayı: 37210, 04.Şubat.2021.

[11] TUİK, "İllere göre motorlu kara taşıtları sayısı", Available: https://data.tuik.gov.tr/Bulten/Index?p=Road-Motor-Vehicles-December-2020-37410, [Accessed Aralık, 2020].

[12] A Liaw, M. Wiener, 2002, Classification And Regression ByRandom Forest, R News, Vol.2/3, December.

[13] L. Breiman, "Random forests", Machine Learning, Volume 45, pp. 5-32, 2001.

[14] N. Kriegeskorte, T. Golan, Neural network models and deep learning, Current Biology, 29(7), R231–R236, 2019.

[15] L. Deng and D. Yu, "Deep Learning: Methods and Applications," Found. Trends® Signal Process., vol. 7, no. 3–4, pp. 197–387, 2014.

[16] K. Chakraborty, K. Mehrotra, C. K. Mohan and S. Ranka, "Forecasting The Behavior of Multivariate Time Series Using Neural Networks", Neural Networks 5(6):961-970, 1992.

[17] Y. Duan, Y. L.V. and F. Wang, "Travel time prediction with LSTM neural network," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016, pp. 1053-1058, doi: 10.1109/ITSC.2016.7795686, 2016.

[18] S. Samui, I. Chakrabarti, S.K. Ghosh, "Tensor-Train Long Short-Term Memory for Monaural Speech Enhancement" arXiv preprint arXiv:1812.10095, 2018

[19] F. A. Gers, J. Schmidhuber, F. A. Cummins, "Learning to forget: Continual prediction with LSTM" Neural Computation, 12 (10), pp. 2451-2471, 2000.

[20] Ulaşım Daire Başkanlığı, "Saatlik Trafik Yoğunluk Veri Seti, Ağustos 2020 Trafik Yoğunluk Verisi" 13 Aralık, 2020.