Science Journal of Applied Mathematics and Statistics
2024, Vol. 12, No. 5, pp. 72-79
https://doi.org/10.11648/j.sjams.20241205.11

Research Article

Traffic Flow Prediction: A Method Using Bagging-Based Ensemble Learning Model

Xinyue Cai[1], **Qinyu Jin**[2, *], **Wenyu Zhang**[2]

[1]International Division, The Affiliated High School to Hangzhou Normal University, Hangzhou, China

[2]School of Information Technology and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou, China

Abstract

For the development of the national economy, transportation is strategically significant. As the increasing ownership of automobiles, traffic jams are a common occurrence. Accurate prediction of traffic flow contributes to diverting traffic effectively and improving the quality of urban traffic, in turn improving the operation of the overall transportation system. The rapid development of artificial intelligence technologies, especially machine learning and deep learning, has provided effective methods for accurate prediction of traffic flow. Based on the above, in order to improve the accuracy of the prediction and to extend the application of machine learning and deep learning in the prediction of traffic flow, this study proposed a bagging-based ensemble learning model. Firstly, normalization method is used to preprocess the data. Subsequently, base prediction models including decision tree, random forest, logistic regression, convolution neural network, long short-term memory and multilayer perceptron are selected for training the prediction model, respectively. Finally, bagging-based ensemble learning method is used to integrate these base prediction models to further predict traffic flow. The results of comparison between the single base prediction models and the bagging-based ensemble learning model on the five evaluation indicators show that, for predicting the traffic flow, the bagging-based ensemble learning model outperforms the base prediction models. Meanwhile, this study explores the potential in the application of machine learning, deep learning, and especially bagging-based ensemble learning to predict traffic flow.

Keywords

Traffic Flow, Prediction, Bagging, Ensemble Learning Model

1. Introduction

With the acceleration of urbanization, traffic flows are increasing rapidly, leading to increasing difficulty in management of traffic, and transportation is deeply related to the national economy. Intelligent transportation systems (ITS) have been gradually developed. Traffic flow, as part of ITS, is an important indicator to reflect the traffic conditions, and accurate prediction of traffic flow can improve operation efficiency of the transportation system with better management of traffic.

In the earlier studies, traffic flow was predicted mainly using statistical modeling [1, 2], which was only applicable to simple short-term prediction. However, holidays, unexpected events,

*Corresponding author: jinqinyu@zufe.edu.cn (Qinyu Jin)

Received: 29 July 2024; **Accepted:** 2 September 2024; **Published:** 10 October 2024

and weather can all have an impact on traffic flow, making accurate prediction of traffic flow a complex and challenging problem [3]. Statistical models have difficulty in meeting the requirements of accurate prediction of traffic flow [4].

Some other studies have predicted traffic flow by simulating urban transportation networks and constructing traffic simulation models [5, 6]. However, the simulation model has some limitations in conducting large-scale simulations. With the development of ITS and artificial intelligence technologies, more and more studies utilize machine learning, deep learning, and other technologies to improve the model performance and predict traffic flow more effectively [3, 7, 8].

Following the existing research, to predict the traffic flow more accurately, this study proposes a bagging-based ensemble learning model. Firstly, traffic flow data from the U.S. is preprocessed using normalization methods. Secondly, based on historical data, the base prediction models including decision tree (DT) [9], random forest (RF) [10], logistic regression (LR) [11], convolution neural network (CNN) [12], long short-term memory (LSTM) [13] and multilayer perceptron (MLP) [14] are utilized to train and predict future traffic flow. Thirdly, with a bagged learning method, these single base prediction models are integrated to predict traffic flow. The comparison experiments between these single base prediction models and the bagging-based ensemble learning model demonstrate that the proposed bagging-based ensemble learning model has superior performance in solving the problem of traffic flow prediction.

The remainder of the study is structured as follows. Section 2 analyzes the literature related to machine learning, deep learning, and the prediction of traffic flow. Section 3 details the bagging-based ensemble learning model. Comparative experiments are conducted in Section 4 and the experimental results are analyzed. Summary of this study and perspectives of future are provided in Section 5.

2. Related Work

This section presents an introduction to the literature related to machine learning, deep learning, and the prediction of traffic flow to provide a better understanding of the previous research.

2.1. Machine Learning and Deep Learning

Machine learning is the process that automatically generalizes logic or rules from data through algorithms, and can make prediction based on the results of the generalization. Data prediction using machine learning algorithms is one of the common tasks in data analysis and has been widely used in various fields. For example, Ma et al. [15] proposed three DT models to predict the length of stay of patients to improve the management of hospital beds. Milanović et al. [16] used several machine learning models, such as RF and LR, to predict forest fires in order to obtain forest fire probability maps. Huang et al. [17] conducted a comparison between machine learning models and statistical models regarding the performance of landslide susceptibility prediction, and the results showed that the statistical models were limited by linear analysis, and the machine learning models had superior performance.

With the improvement of computing power and abundant data, deep learning has become a shining star in the field of machine learning. In particular, deep learning has achieved outstanding results in the field of time series prediction. For example, Hu et al. [18] combined the convolutional LSTM and the bidirectional LSTM as a hybrid deep learning model to predict traffic speed. Alharkan et al. [19] combined CNN and LSTM to improve the accuracy of short-term solar power prediction. Saboor et al. [20] used machine learning models and deep learning models to predict the stocks and then compared the results, finding that the recurrent neural networks (RNN)-based deep learning model outperforms the common machine learning models.

2.2. Prediction of Traffic Flow

In order to provide a scientific basis for intelligent transportation planning, traffic flow prediction has been widely studied. Earlier studies focused on using statistical models to predict traffic flow. For example, Kim [1] proposed a statistical model for real-time traffic flow prediction by using Markov random field (MRF) to modify the changes in traffic flow and applying it to a Korean expressway. Chen et al. [2] proposed an improved autoregressive integrated moving average (ARIMA) model to improve accuracy of prediction and reduce complexity of computation for predicting traffic flow. However, statistical models for predicting traffic flow can be limited by linear analysis and have weak performance in making complex nonlinear predictions.

Some studies predicted the traffic flow by constructing simulation models. Wang et al. [5] constructed a traffic simulation model that integrates driver behavior and environment to predict traffic flow with heterogeneous behaviors. Ma et al. [6] suggested that the predicted traffic flow through the simulation of urban traffic generated from geo-population distribution data is referential.

Table 1. General information about the dataset utilized in this study.

Dataset	Sensor	Area	Time Span	Number of records
PeMSD4	400038	Bay Area	1 June 2017 to 30 June 2017	8640

With the development of artificial intelligence technologies, and for more accurate and effective traffic flow predictions, machine learning and deep learning have begun to be combined into prediction models. Chen et al. [3] dynamically combined spatio-temporal features with external factors and then extended the CNN to a multi-gated spatio-temporal CNN model to predict citywide traffic flow. Aljuaydi et al. [7] used several deep learning models (e.g., MLP, CNN, LSTM, CNN-LSTM, and autoencoder LSTM) to predict traffic flow and comprehensively compared and analyzed the prediction performance of each model. Redhu et al. [8] proposed an extended deep learning prediction model by combining the particle swarm optimization algorithm and LSTM to predict traffic flow. However, the above traffic flow prediction models are mostly single prediction models, and the accuracy and stability of the prediction still need to be improved.

3. Methodology

In this session, the bagging-based ensemble learning model will be described in detail.

3.1. Data Description and Preprocessing

The dataset used for the experiments in this study is from Performance Measurement System (PeMS) in the United States [21],[1] and 8640 records obtained from the sensor 40038 (located in Bay Area) from 1 June 2017 to 30 June 2017 were selected for the traffic flow prediction, where each traffic flow data was recorded at 5-minute intervals. The general information about the dataset is presented through Table 1.

To improve the quality of the raw data and reduce the bias during the modeling, the preprocessing to the raw data is necessary. In this study, data normalization and missing data imputation are utilized to preprocess the raw data.

(1) Data normalization

In order to scale the raw data to the range from 0 to 1, and to reduce the differences between the data, Min-Max normalization is employed, which is calculated as shown in Eq. (1).

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_{new} indicates the value of the normalized data, which ranges from 0 to 1. x_{old} indicates the value of the raw data. x_{max} and x_{min} indicates the maximum and minimum values of the raw data in the dataset, respectively.

(2) Missing data imputation

For missing data values in the dataset, 0 is used for imputation to ensure the integrity of data.

3.2. Bagging-Based Ensemble Learning Model

To overcome the shortcoming of a single base prediction model in terms of both accuracy and stability of prediction, this study proposed a bagging-based ensemble learning model to predict the traffic flow. Firstly, the preprocessed dataset is divided into three parts, i.e. the training set, validation set and test set, accounting for 64%, 16%, and 20%, respectively. Secondly, bagging-based ensemble learning model is divided into two phases, i.e., the training and optimization of the base prediction models and the training and prediction of the ensemble learning model.

(1) The training and optimization of the base prediction models

In this phase, six base prediction models that have relatively better predictive performance were selected, including three machine learning models (i.e., DT, RF, and LR) and three deep learning models (i.e., CNN, LSTM, and MLP). The training set is randomly sampled six times to obtain six independent training subsets. Each training subset was used to train each base prediction model respectively, and the validation set is used to verify the training results and obtain the best parameters of the trained base prediction model, and then the trained optimal base prediction model is obtained.

(2) The training and prediction of the bagging-based ensemble learning model

In this phase, the trained optimal base prediction models obtained in the previous phase use the validation set to get the predictor factors of the validation set, which are integrated as a matrix. Then, the ensemble learning model takes the predictor factors as training inputs, and the trained ensemble learning model is obtained. Finally, the test set is also inputted into trained optimal base prediction models for getting the predictor factors of the test set, and further processed by the trained ensemble learning model to get the final prediction results. Figure 1 presents the detailed structure of bagging-based ensemble learning model.

1 http://pems.dot.ca.gov

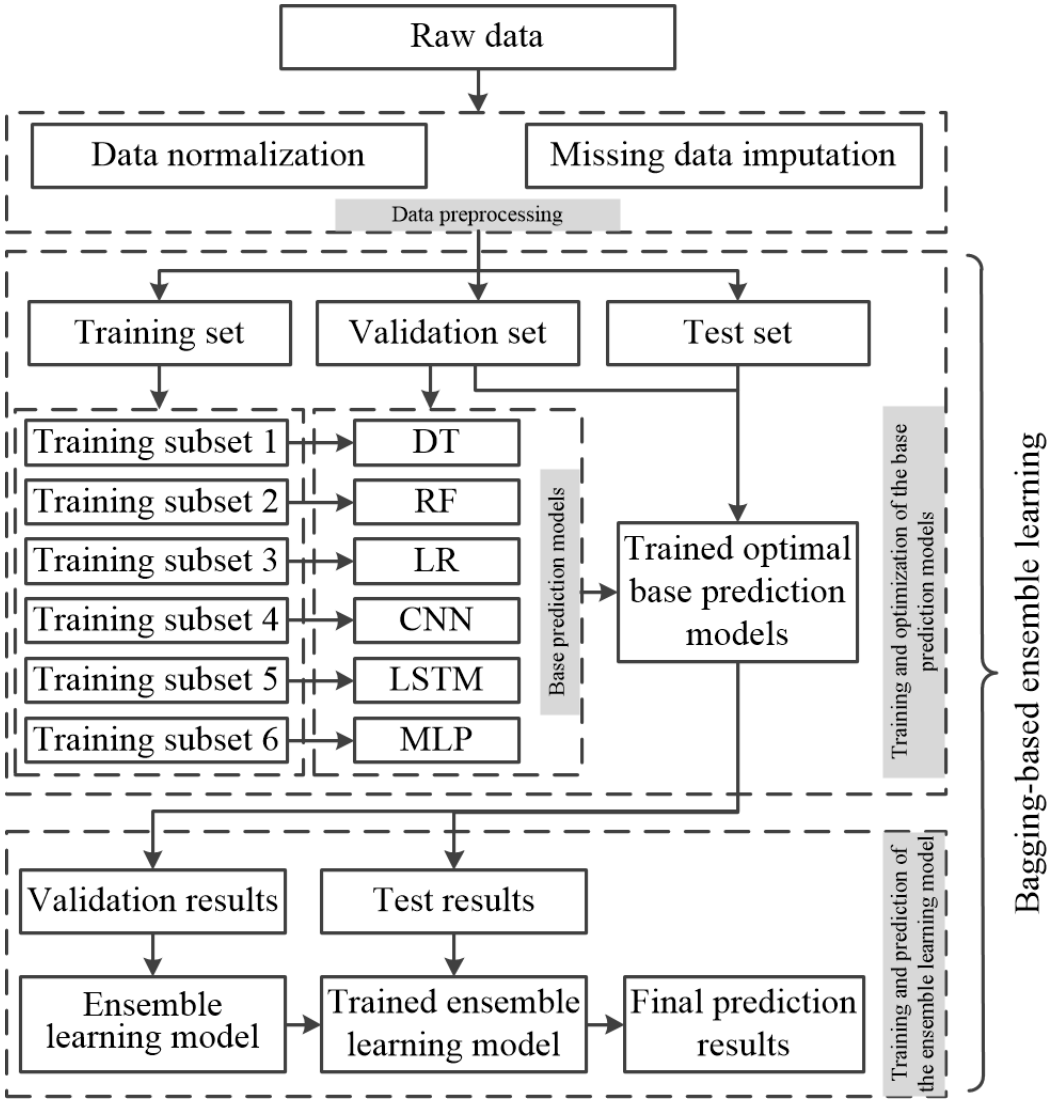


Figure 1. *Detailed structure of bagging-based ensemble learning model.*

4. Experiment

The comparison experiments between the single base prediction models and the proposed bagging-based ensemble learning model are conducted and the performance of the proposed model is analyzed in this section. All experiments were implemented with the Python programming language on a personal computer with an Intel Core i5, 1.60GHz CPU, 16GB of RAM, and an Intel (R) UHD Graphics GPU.

4.1. Evaluation Metrics

To analyze the performance of the prediction models, the following five evaluation indicators are selected, including mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), R-squared (R2), and running time (RT).

MAE is used to describe the gap between the predicted value and the true value, which can be calculated by Eq. (2).

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|x_n^P - x_n^T\right| \tag{2}$$

where n indicates the n-th time point, $n = 1, 2, \cdots, N$, where N is the total number of time points. x_n^P and x_n^T indicate the predicted and true values of traffic flow at the n-th time point, respectively.

RMSE is used to describe the degree of bias between the predicted value and the true value, which is more sensitive to the outliers in the data, and can be calculated by Eq. (3).

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(x_n^P - x_n^T\right)^2} \tag{3}$$

MAPE is the average percentage of relative error between the predicted value and the true value, which can be calcu-

lated by Eq. (4).

$$MAPE = \frac{100\%}{N}\sum_{n=1}^{N}\left|\frac{x_n^P - x_n^T}{x_n^T}\right| \quad (4)$$

The closer the values of the above evaluation indicators and RT are to 0, the more superior the model is.

R^2 is the coefficient of determination, which is used to describe the fit of the predictive model to the data. The closer the value of R^2 is to 0, it indicates that the model has a higher fit, i.e., has superior ability of prediction. R^2 can be calculated by Eq. (5).

$$R^2 = 1 - \sum_{n=1}^{N}\left(x_n^T - x_n^P\right)^2 \Big/ \sum_{n=1}^{N}\left(x_n^T - \bar{x}_n^T\right)^2 \quad (5)$$

where $\bar{x}_n^T$ indicates the average of the true values.

4.2. Experimental Results and Analysis

In order to analyze the performance of the bagging-based ensemble learning model in predicting traffic flow, comparison experiments are conducted between six single base prediction models and the ensemble learning models. Then the five indicators mentioned above are used to describe the prediction performance of the models. The results of the experiments are shown in Table 2.

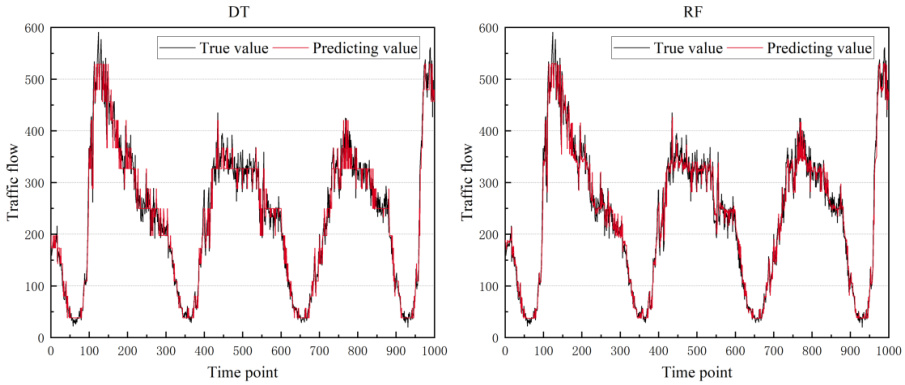
Table 2. *Results of comparison experiments for all models.*

Models	**MAE↓**	**RMSE↓**	**MAPE↓**	**R^2↑**	**RT↓**
DT	21.216	28.563	0.110	0.957	0.021
RF	22.284	29.952	0.115	0.951	0.487
LR	24.761	34.926	0.134	0.939	0.898
CNN	26.232	39.820	0.136	0.920	3.762
LSTM	27.353	41.996	0.138	0.911	2.934
MLP	26.082	38.882	0.137	0.924	2.843
Ensemble learning	20.242	28.474	0.103	0.959	2.582

Note: The significant values are bolded, "↓" indicates that the smaller value is better, "↑" indicates that the bigger value is better.

As shown in Table 2, in most of the indicators, the prediction results of the bagging-based ensemble learning model are better than those of the single base prediction models, which indicates that the bagging-based ensemble learning model outperforms the single base prediction models in terms of prediction accuracy and stability, and has better performance in predicting traffic flow.

To more intuitively present the performance of each model in predicting traffic flow, Figure 2 visualizes the degree of fit between the prediction and true values of each model. As shown in Figure 2, the predicted values obtained by the ensemble learning model are closer to the real values, indicating that in predicting traffic flow, the bagging-based ensemble learning model has superior performance than single base prediction models and is more accurate and stable in capturing the change trend of traffic flow.



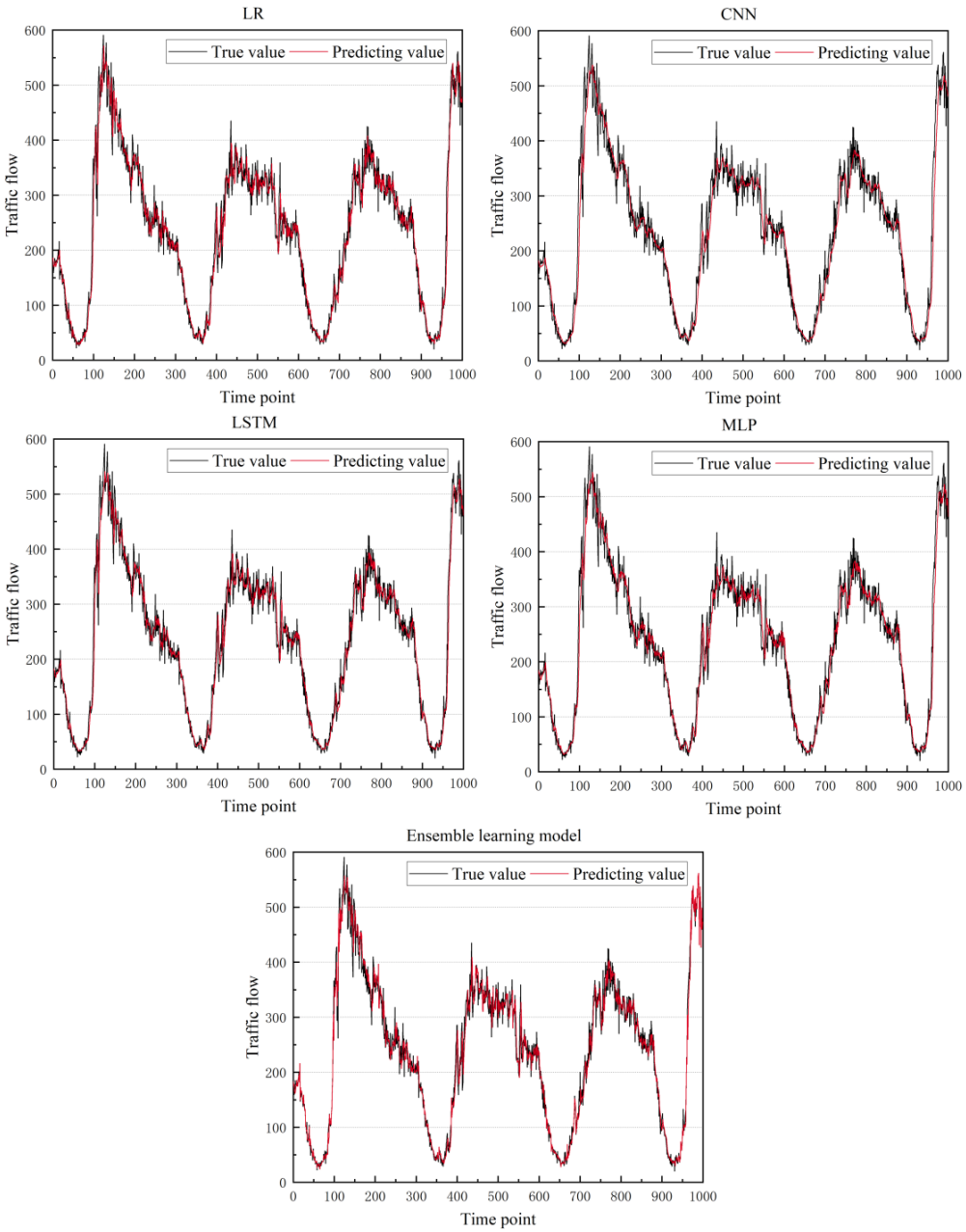


Figure 2. *The degree of fit between the prediction and true values of each model.*

5. Conclusion

The rapid urbanization of transportation has led to an increase in traffic flows. ITS is being developed to improve traffic management. Traffic flow prediction is interrupted by many factors and is a challenging problem. Based on the development of artificial intelligence techniques, including machine learning and deep learning, to predict traffic flow more effectively, a bagging-based ensemble learning model is proposed in this study. The superiority of the ensemble learning model in predicting traffic flow is verified by comparison experiments with the base prediction models.

However, this study has the potential to make further improvements. Firstly, the application of more complex and advanced ensemble learning methods, such as boosting and stacking, contributes to the improvement of the prediction accuracy and robustness of the ensemble learning model. Besides, more hyper-parameter optimization methods, such as swarm intelligence optimization algorithms and Bayesian optimization algorithms, should be explored to obtain superior prediction models. In addition, more evaluation indicators can be used to show and analyze the superiority of ensemble learning model more comprehensively. Finally, extending the application fields of ensemble learning models, i.e., to solve other prediction problems besides traffic flow prediction with ensemble learning models, is also a worthwhile exploration.

Abbreviations

ITS	Intelligent Transportation Systems
DT	Decision Tree
RF	Random Forest
LR	Logistic Regression
CNN	Convolution Neural Network
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
PeMS	Performance Measurement System
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
R^2	R-squared
RT	Running Time

Author Contributions

Xinyue Cai: Writing – original draft, Methodology, Visualization

Qinyu Jin: Conceptualization, Software, Writing – review & editing

Wenyu Zhang: Writing – review & editing, Project administration

Data Availability Statement

The data that support the findings of this study can be found at: http://pems.dot.ca.gov.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Kim, E. Y. MRF model based real - time traffic flow prediction with support vector regression, Electronics Letters. 2017, 53(4) 243-245. https://doi.org/10.1049/el.2016.3472

[2] Chen, J. B., Li, D. M., Zhang, G. L., & Zhang, X. L. Localized space-time autoregressive parameters estimation for traffic flow prediction in urban road networks, Applied Sciences.2018, 8(2), 277. https://doi.org/10.3390/app8020277

[3] Chen, C., Li, K. L., Teo, S. G., Zou, X. F., Li, K. Q., & Zeng, Z. Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks, ACM Transactions on Knowledge Discovery from Data (TKDD). 2020, 14(4), 1-23. https://doi.org/10.1145/3385414

[4] Chen, K., Chen, F., Lai, B. S., Jin, Z. M., Liu, Y., Li, K., et al. Dynamic spatio-temporal graph-based CNNs for traffic flow prediction, IEEE Access. 2020, 8, 185136-185145. https://doi.org/10.1109/ACCESS.2020.3027375

[5] Wang, H., He, X. Y., Chen, L. Y., Yin, J. R., Han, L., Liang, H., et al. Cognition-driven traffic simulation for unstructured road networks, Journal of Computer Science and Technology. 2020, 35, 875-888. https://doi.org/10.1007/s11390-020-9598-y

[6] Ma, X. Y., Hu, X. W., Weber, T., & Schramm, D. Evaluation of accuracy of traffic flow generation in SUMO, Applied Sciences. 2021, 11(6), 2584. https://doi.org/10.3390/app11062584

[7] Aljuaydi, F., Wiwatanapataphee, B., & Wu, Y. H. Multivariate machine learning-based prediction models of freeway traffic flow under non-recurrent events, Alexandria engineering journal. 2023, 65, 151-162. https://doi.org/10.1016/j.aej.2022.10.015

[8] Redhu, P., Redhu, P., & Kumar, K. Short-term traffic flow prediction based on optimized deep learning neural network: PSO-Bi-LSTM, Physica A: Statistical Mechanics and its Applications. 2023, 625, 129001. https://doi.org/10.1016/j.physa.2023.129001

[9] Hunt, E. B., Marin, J., & Stone, P. J. Experiments in induction, The American Journal of Psychology. 1966, 80(4). https://doi.org/10.2307/1421207.

[10] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. 2015, 71, 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001

[11] Efron, B. Logistic regression, survival analysis, and the Kaplan-Meier curve, Journal of the American Statistical Association. 1988, 83(402), 414-425. https://doi.org/10.1080/01621459.1988.10478612

[12] Ferreira, A., and Giraldi, G. Convolutional neuralnetwork approaches to granite tiles classification, Expert Systems with Applications. 2017, 84, 1-11. https://doi.org/10.1016/j.eswa.2017.04.053

[13] Chen, J., Zeng, G. Q., Zhou, W. N., Du, W., and Lu, K. D. Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization, Energy Conversion and Management. 2018, 165, 681-695. https://doi.org/10.1016/j.enconman.2018.03.098

[14] Liu, D., Wang, J. L., and Wang, H. Short-term wind speed forecasting based on spectral clustering and optimized echo state networks, Renewable Energy. 2015, 78, 599-608. https://doi.org/10.1016/j.renene.2015.01.022

[15] Ma, F., Yu, L. M., Ye, L. S., Yao, D. D., & Zhuang, W. F. Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods, IEEE Journal of Biomedical and Health Informatics. 2020, 24(9), 2651-2662. https://doi.org/10.1109/JBHI.2020.2973285

[16] Milanović, S., Marković, N., Pamučar, D., Gigović, L., Kostić, P., & Milanović, S. D. Forest fire probability mapping in eastern Serbia: Logistic regression versus random forest method, Forests. 2020, 12(1), 5. https://doi.org/10.3390/f12010005

[17] Huang, F. M., Cao, Z. S., Guo, J. F., Jiang, S. H., Li, S., & Guo, Z. Z. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping, Catena. 2020, 191, 104580. https://doi.org/10.1016/j.catena.2020.104580

[18] Hu, X. J., Liu, T., Hao, X. T., & Lin, C. X. Attention-based Conv-LSTM and Bi-LSTM networks for large-scale traffic speed prediction, The Journal of Supercomputing. 2022, 78(10), 12686-12709. https://doi.org/10.1007/s11227-022-04386-7

[19] Alharkan, H., Habib, S., & Islam, M. Solar power prediction using dual stream CNN-LSTM architecture, Sensors. 2023, 23(2), 945. https://doi.org/10.3390/s23020945

[20] Saboor, A., Hussain, A., Agbley, B. L. Y., ul Haq, A., Li, J. P., & Kumar, R. Stock market index prediction using machine learning and deep learning techniques, Intelligent Automation & Soft Computing. 2023, 37(2), 1325-1344. https://doi.org/10.32604/iasc.2023.038849

[21] Chen, C., Petty, K., Skabardonis, A., Varaiya, P., & Jia, Z. F. Freeway performance measurement system: mining loop detector data, Transportation Research Record. 2001, 1748(1), 96-102. https://doi.org/10.3141/1748-12