Ensemble Learning-based Algorithms for Traffic Flow Prediction in Smart Traffic Systems

Anas Saleh Alkarim
Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia | Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
aaalkarim@stu.kau.edu.sa (corresponding author)

Abdullah S. Al-Malaise Al-Ghamdi
Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia | Information Systems Department, School of Engineering, Computing and Design, Dar Al-Hekma University, Saudi Arabia
aalmalaise@kau.edu.sa

Mahmoud Ragab
Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia | Department of Mathematics, Faculty of Science, Al-Azhar University, Egypt
mragab@kau.edu.sa

Received: 18 December 2023 | Revised: 8 January 2024 | Accepted: 16 January 2024

 DOI: https://doi.org/10.48084/etasr.6767

ABSTRACT

Due to the tremendous growth of road traffic accidents, Intelligent Transportation Systems (ITSs) are becoming even more important. To prevent road traffic accidents in the long term, it is necessary to find new vehicle flow management techniques in order to optimize traffic flow. With the high growth of deep learning and machine learning, these methods are increasingly being used in ITSs. This research provides a novel conceptual ITS model that aims to predict vehicle movement through the collective learning usage to anticipate intersections. The proposed approach consists of three main stages: data collection through cameras and sensors, implementation of machine learning and deep learning algorithms, and result evaluation, utilizing the coefficient of determination (R-squared), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). To accomplish this, various machine learning and deep learning algorithms, such as Random Forest, LSTM, Linear Regression, and ensemble methods (bagging), were incorporated into the model. The findings revealed the enhancement due to the proposed method, which was observed through a significant performance improvement of 93.52%.

Keywords-intelligent transportation systems; smart traffic systems; traffic flow; prediction models; smart cities; bagging ensemble learning

I. INTRODUCTION

Road traffic accidents are dramatically augmented each year due to the massive increasing number of vehicles on the roads. This problem is considered a serious risk, a major source of trouble for individuals worldwide, and a significant global concern. [1]. Collecting and analyzing comprehensive data is essential for any initiative aiming to improve traffic safety [2]. With the rising number of vehicles on the roads and the resulting congestion issues, optimizing traffic flow has become a pressing challenge in modern cities. Intelligent Transportation Systems (ITSs) have emerged as a promising solution to alleviate traffic congestion and enhance overall transportation efficiency [3-4]. The Vehicle Ad-Hoc Network (VANET) serves as a fundamental infrastructure for ITSs, enabling wireless connectivity among vehicles [5-6]. Additionally, intelligent transport systems are increasingly focused on addressing traffic congestion. Researchers have employed machine learning algorithms to predict traffic flow and reduce congestion at intersections. These models were evaluated using

the national road traffic dataset for the UK. An adaptive traffic light system was implemented, which adjusts green and red lights based on road width, traffic density, and vehicle categories. Simulations demonstrated a 30.8% decrease in traffic congestion [7].

Accurate traffic prediction is crucial for ameliorating the effectiveness of traffic systems and reducing energy consumption. Machine learning-based methods have become commonplace, but they often rely on historical data [8-9]. Furthermore, ML-based models are gaining popularity due to their ability to accurately forecast traffic conditions, thereby improving safety and infotainment applications. However, the efficacy of these models in predicting real-time traffic remains a subject of investigation [10].

Several research studies have focused on developing methods and models for traffic flow prediction and management. In [11], a framework is presented that utilizes Vector Auto Regression (VAR) and a CNN-LSTM hybrid neural network to predict short-term traffic flow. The CNN-LSTM model outperforms other models in forecasting short-term traffic flow and demonstrates predictive accuracy associated with spatial correlation in traffic flow. In [12], three proposed solutions are discussed to address the issue of missing data in traffic management. These solutions include a live-traffic simulation, a neural network traffic prediction and rerouting system based on pheromone principles, as well as a Weighted Missing Data Imputation (WEMDI) approach. The integration of WEMDI into the systems yields notable improvements in various traffic factors and demonstrates efficient routing to alternative destinations.

ML and neural networks play a significant role in solving traffic congestion issues. In this context, authors in [13] propose ML and DL algorithms for predicting intersection traffic flow. The models were trained, validated, and tested using public datasets, and the Multilayer Perceptron Neural Network (MLP-NN) produced the best results. Gradient Boosting, Recurrent Neural Networks, RF, LR, and Stochastic Gradient also showed promising performance.

ITSs require traffic flow monitoring for effective management and optimization. Conventional methods of data collection and analysis are being augmented with AI techniques, such as ensemble learning [14]. The IAROEL-TFMS methodology utilizes feature subset selection and optimal ensemble learning to predict traffic flow, outperforming other approaches with its low RMSE. Authors in [14] used Hybrid-LSSVM, AST2FP-OHDBN, and IAROEL-TFMS models for evaluation purposes, considering their respective performance indicators. Among the several models evaluated, it was observed that IAROEL-TFMS had the most superior predictive performance. In close succession, the AST2FP-OHDBN model exhibited robust performance, whereas, in contrast, the Hybrid-LSSVM model demonstrated a somewhat reduced level of prediction accuracy. Regarding predictive performance, the IAROEL-TFMS model had the best precision and accuracy in forecasting the target variable. The AST2FP-OHDBN model closely followed it. On the other hand, the Hybrid-LSSVM model exhibited slightly inferior predictive skills.

This paper utilizes four Machine Learning (ML) and Deep Learning (DL) models: Random Forest (RF), Linear Regression (LR), Long Short-Term Memory (LSTM), and ensemble bagging (RF). The objective is to utilize these predictions to enhance the efficiency of traffic light controllers in the context of traffic flow prediction at intersections. Experimental results demonstrate that all models exhibit a strong predictive capacity for estimating vehicular flow, highlighting their potential utility in smart traffic systems.

II. THE PROPOSED MODEL

This study has developed a model for monitoring traffic flow. The primary objective of this model is to predict traffic movement. To achieve this objective, the model operates in three distinct stages. Firstly, data collection can be accomplished using cameras or sensors. Secondly, ML and DL technologies are applied. Thirdly, the outcomes are evaluated using MAE, RMSE, and the coefficient of determination (R-squared). The workflow of the suggested approach is illustrated in Figure 1. Overall, the proposed model aims to monitor traffic flow by predicting its movement. It involves data collection through cameras or sensors, the application of machine learning and deep learning technologies, and the evaluation of outcomes using specific error metrics. Figure 1 provides a visual representation of the workflow.

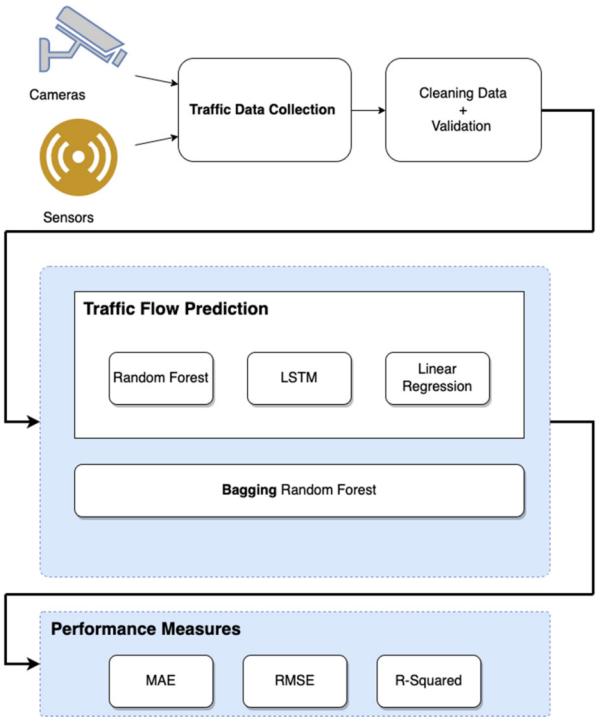

Fig. 1. The workflow of the proposed model.

A. Data Collection

The dataset used for traffic prediction was obtained from various traffic sensors provided by the Huawei Munich Research Center. The dataset plays a crucial role in predicting traffic patterns and making necessary adjustments to stop-light control settings, including cycle length, offset, and split

timings. The dataset consists of recorded data from six intersections located within an urban area, collected over a period of 56 days (Table I). The data are presented as a flow time series, which indicates the number of vehicles passing through each intersection every 5 minutes, spanning 24 hours. This results in 12 readings per hour, 288 readings per day, and a total of 16,128 readings over the course of the 56 days. For this study, 4 out of the 6 intersections were selected to replicate a 4-lane intersection scenario [15].

TABLE I. THE 6 INTERSECTIONS OF THE DATASET USING COLAB

	Cross 1	Cross 2	Cross 3	Cross 4	Cross 5	Cross 6
0	105.0	48.0	30	62.0	31	110.0
1	97.0	41.0	32	55.0	42	103.0
2	76.0	47.0	44	58.0	40	100.0
3	98.0	40.0	39	59.0	43	104.0
4	87.0	41.0	47	49.0	35	112.0

B. Data Preparation

Data cleaning is a critical step in the preprocessing phase, where incorrect, incomplete, duplicate, or erroneous data within a dataset are rectified. Fortunately, the collected data for this study do not contain any missing values. The dataset has been divided into two parts: 70% for training the model and the remaining portion for testing. To ensure consistency and optimal performance during training, the data were scaled using the MinMaxScaler from the scikit-learn library. This scaler transforms the data, making them range between zero and one [16].

C. Proposed Techniques

In this study, four regression models from the scikit-learn module in the Python programming language are employed. The scikit-learn module is a comprehensive Python library that offers a wide range of state-of-the-art machine learning algorithms designed to tackle various supervised and unsupervised challenges [16]. The authors applied four ML/DL techniques to the dataset: RF, LSTM, LR, and ensemble method (bagging). The following section provides an overview of the traditional ML and ensemble methods utilized in the experiment.

III. OVERVIEW OF TRADITIONAL MACHINE LEARNING AND ENSEMBLE METHODS

A. Random Forest

RF is a learning method that combines multiple tree predictors. Each tree in the forest is constructed based on the values of a random vector, sampled independently from the same distribution for all trees. Tree-based models form the core components of the random forest algorithm. A tree-based model involves iteratively dividing a given dataset into two distinct groups, guided by a specific criterion, until a predetermined stopping condition is met. The terminal nodes of decision trees are commonly known as leaf nodes or leaves [17].

B. Long Short-Term Memory (LSTM)

LSTM networks have found extensive applications in various domains, including image processing, speech recognition, manufacturing, autonomous systems, communication, and energy consumption, for dynamic system modelling purposes. LSTM has gained significant attention in recent years due to its effectiveness in modeling and predicting the dynamics of nonlinear time-variant systems. It incorporates the characteristics of short-term and long-term memory, the ability to make predictions several steps ahead, and the propagation of errors. Sequence-to-sequence networks with partial conditioning have been shown to outperform other techniques such as bidirectional or associative networks, making them well-suited for achieving the specified objectives [18].

C. Linear Regression

LR is a widely used and straightforward ML algorithm. The technique is a mathematical methodology employed to do predictive analysis. Moreover, LR is a statistical technique that enables the prediction of continuous or numerical variables. LR is a statistical technique employed to assess and quantify the association between variables under consideration [19].

D. Ensemble Method (Bagging)

Bagging, short for bootstrap aggregating, is a technique that involves creating multiple iterations of a predictor and combining them to form an aggregated predictor. In the aggregation process, the mean is calculated across the iterations when predicting a numerical outcome, while a majority vote is used when predicting a class. To generate multiple versions, bootstrap copies of the original learning set are created, and these replicates are then used as new learning sets [20].

IV. EVALUATION MEASURES

In model evaluation, the coefficient of determination (R-squared), RMSE, and MAE are standard metrics [21].

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(x_i - y_i)^2}{\sum_{i=1}^{m}(\bar{y} - y_i)^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

V. EXPERIMENTAL RESULTS AND DISCUSSION

Table II presents the results of the model using various ML and DL algorithms. It can be observed that RF achieved an MAE of 13.76, while LSTM and LR achieved 14.74 and 17.80, respectively. When Bagging (RF) was applied, the minimum MAE obtained was 13.69. In terms of RMSE, the models achieved values of 22.39, 23.50, and 27.04, while the Bagging model achieved a lower RMSE of 22.21. In terms of R^2, the experimental results for the models were 0.9341, 0.9275, and 0.9040, respectively. The best R^2 value was obtained by the Bagging model, which achieved a value of 0.9352. The results show that the RF model and the Bagging model (using RF as the base model) outperformed the LSTM and LR models in terms of both MAE and RMSE. Additionally, the Bagging

model showed the highest R^2 value, suggesting a better fit to the data. Overall, these findings demonstrate the effectiveness of the RF algorithm and the potential benefits of using ensemble methods like Bagging for traffic flow prediction.

TABLE II. RESULTS

ML/DL Model	MAE	RMSE	R^2
RF	13.76	22.39	0.9341
LSTM	14.74	23.50	0.9275
LR	17.80	27.04	0.9040
Bagging (RF)	13.69	22.21	0.9352

Figure 2 illustrates the numerical values of MAE measurements of the considered models. It can be observed that the LSTM model has a slightly higher MAE (14.74) compared to the RF (13.76) and Bagging (RF) (13.69) models. This suggests that the LSTM model may not perform optimally in this particular scenario. On the other hand, the LR model has the highest MAE score (17.80). This indicates that it may not excel at accurately predicting the target variable. These results suggest that the RF and Bagging models (using RF as base) perform better than the LSTM and LR models in terms of MAE. It is important to consider these findings when selecting the most suitable model for traffic prediction in this context.

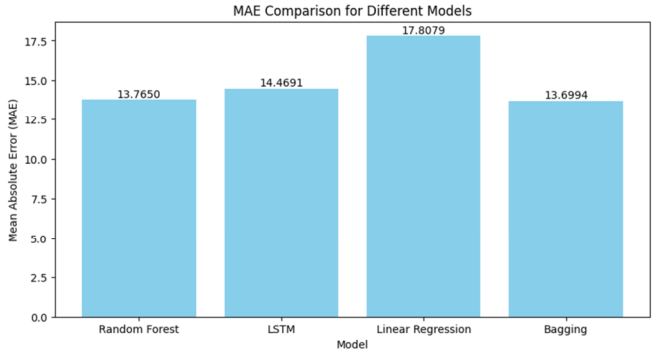

Fig. 2. MAE of the different models.

The figures presented in Figure 3 illustrate the RMSE values of the considered models. It can be observed that the Bagging (RF) model has the lowest RMSE score (22.21), indicating that, on average, its predicted values deviate the least from the actual values. This suggests that the model exhibits strong predictive accuracy. The RF model also performs well, although it has a slightly higher RMSE (22.39) compared to the Bagging model. On the other hand, the LSTM model shows a larger RMSE (23.50), indicating potentially inferior performance in terms of predictive accuracy. The LR model has the largest RMSE value (27.04), suggesting a potentially lower level of accuracy in predicting the target variable. These findings again suggest that both the Bagging (RF) and RF models perform well in terms of RMSE, indicating their ability to provide accurate predictions. However, the LSTM and LR models may have limitations in accurately predicting the target variable based on their higher RMSE values.

Figure 4 shows the R^2 values of the considered models. R^2 ranges from 0 to 1, with a value of 1 indicating a perfect fit. Among the models presented, it is evident that the Bagging (RF) model shows the highest R^2 value (0.9352), designating its superior ability to fit the data accurately. The RF (0.9341) and LSTM (0.9275) models also demonstrate high R^2 values, suggesting their effectiveness in explaining a significant proportion of the observed variability in the dependent variable, while LR performs satisfactorily (0.9040), its R^2 value is slightly lower compared to the alternative models. Overall, it can be noticed that all of the models exhibit strong performance in elucidating the variability in the dependent variable. However, it is noteworthy that Bagging (RF) emerges as the most prominent performer among them.

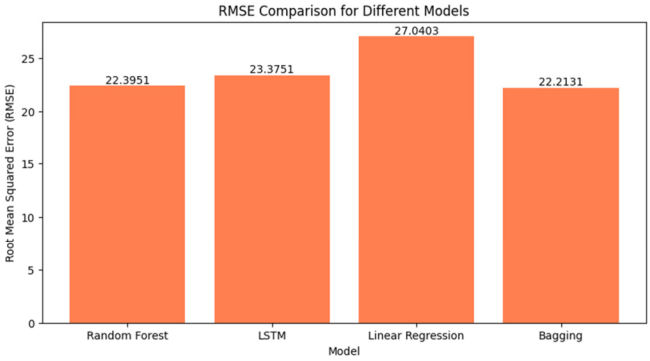

Fig. 3. RMSE of the different models.

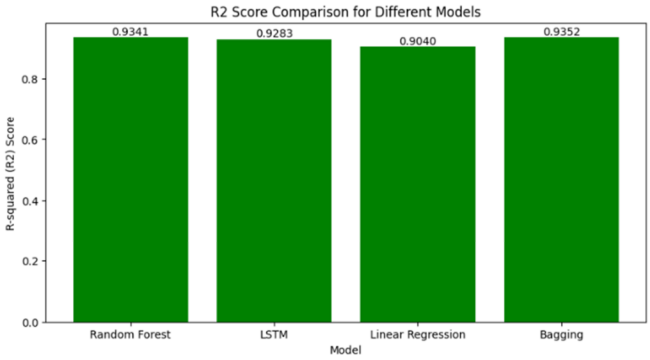

Fig. 4. R^2 of the different models.

Compared to [13], this research has improved the results by more than 0.5%. In [13], the researchers used the same dataset and applied 5 ML methods. Gradient boosting was the most succesful, with 93.05%. The proposed model gets 93.41% by utilizing an RF model. Bagging (RF) has the highest result of 93.52%. In the future, researchers in this field could use a combination of ML and DL models to improve model performance [22].

VI. CONCLUSION

In this article, we mainly presented a new model to enhance intelligent traffic systems. The main purpose of this method is to recognize the traffic flow prediction or vehicle movement at intersections applying an ensemble learning technique. The proposed framework consisted of three primary phases: data collection through cameras and sensors, implementation of

machine learning and deep learning algorithms, and evaluation of the outcomes using metrics such as R-squared, RMSE, and MAE. The model utilized Random Forest, LSTM, Linear Regression, and Bagging (Random Forest), to achieve its objectives. To safeguard the better performance of the proposed procedure, a series of tests was involved. The comprehensive results highlighted the improved performance of the proposed method by achieving the significant accuracy of 93.52%. Regarding future work, combining multiple machine learning and deep learning algorithms could be explored to further enhance the performance of this model for more effective intelligent traffic systems.

REFERENCES

[1] M. Angın and S. I. A. Ali, "Analysis of Factors Affecting Road Traffic Accidents in North Cyprus," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7938–7943, Dec. 2021, https://doi.org/10.48084/etasr.4547.

[2] N. K. Al-Shammari and S. M. H. Darwish, "In-depth Sampling Study of Charactersitics of Vehcile Crashes in Saudi Arabia," *Engineering, Technology & Applied Science Research*, vol. 9, no. 5, pp. 4724–4728, Oct. 2019, https://doi.org/10.48084/etasr.2939.

[3] P. Sen, "Optimization of Traffic Flow Using Intelligent Transportation Systems," *Mathematical Statistician and Engineering Applications*, vol. 70, no. 1, pp. 720–727, Jan. 2021, https://doi.org/10.17762/msea.v70i1.2530.

[4] Y. A. Hanafy, M. Mashaly, and M. A. Abd El Ghany, "An Efficient Hardware Design for a Low-Latency Traffic Flow Prediction System Using an Online Neural Network," *Electronics*, vol. 10, no. 16, Jan. 2021, Art. no. 1875, https://doi.org/10.3390/electronics10161875.

[5] A. K. Kazi and S. M. Khan, "DyTE: An Effective Routing Protocol for VANET in Urban Scenarios," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6979–6985, Apr. 2021, https://doi.org/10.48084/etasr.4076.

[6] S. S. Sepasgozar and S. Pierre, "Network Traffic Prediction Model Considering Road Traffic Parameters Using Artificial Intelligence Methods in VANET," *IEEE Access*, vol. 10, pp. 8227–8242, 2022, https://doi.org/10.1109/access.2022.3144112.

[7] I. Moumen, J. Abouchabaka, and N. Rafalia, "Adaptive traffic lights based on traffic flow prediction using machine learning models," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, Oct. 2023, Art. no. 5813, https://doi.org/10.11591/ijece.v13i5.pp5813-5823.

[8] Z. Wang, P. Sun, Y. Hu, and A. Boukerche, "A Novel Mixed Method of Machine Learning Based Models in Vehicular Traffic Flow Prediction," in *Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, New York, NY, USA, Jul. 2022, pp. 95–101, https://doi.org/10.1145/3551659.3559047.

[9] F. Sheriff, "ELMOPP: an application of graph theory and machine learning to traffic light coordination," *Applied Computing and Informatics*, Mar. 2021, https://doi.org/10.1108/ACI-07-2020-0035.

[10] P. Sun, N. Aljeri, and A. Boukerche, "Machine Learning-Based Models for Real-time Traffic Flow Prediction in Vehicular Networks," *IEEE Network*, vol. 34, no. 3, pp. 178–185, Feb. 2020, https://doi.org/10.1109/MNET.011.1900338.

[11] Z. Cheng, J. Lu, H. Zhou, Y. Zhang, and L. Zhang, "Short-Term Traffic Flow Prediction: An Integrated Method of Econometrics and Hybrid Deep Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5231–5244, Jun. 2022, https://doi.org/10.1109/tits.2021.3052796.

[12] R. K. C. Chan, J. M.-Y. Lim, and R. Parthiban, "A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system," *Expert Systems with Applications*, vol. 171, Jun. 2021, Art. no. 114573, https://doi.org/10.1016/j.eswa.2021.114573.

[13] A. Navarro-Espinoza et al., "Traffic Flow Prediction for Smart Traffic Lights Using Machine Learning Algorithms," *Technologies*, vol. 10, no. 1, Jan. 2022, Art. no. 5, https://doi.org/10.3390/technologies10010005.

[14] M. Ragab, H. A. Abdushkour, L. Maghrabi, D. Alsalman, A. G. Fayoumi, and A. A.-M. Al-Ghamdi, "Improved Artificial Rabbits Optimization with Ensemble Learning-Based Traffic Flow Monitoring on Intelligent Transportation System," *Sustainability*, vol. 15, no. 16, Aug. 2023, Art. no. 12601, https://doi.org/10.3390/su151612601.

[15] C. Axenie and S. Bortoli, "Road traffic prediction dataset." Zenodo, Oct. 07, 2020, https://doi.org/10.5281/zenodo.3653880.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, https://doi.org/10.1177/1536867X20909688.

[18] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, Jan. 2021, https://doi.org/10.1016/j.procir.2021.03.088.

[19] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, https://doi.org/10.38094/jastt1457.

[20] E. Hillebrand, M. Lukas, and W. Wei, "Bagging weak predictors," *International Journal of Forecasting*, vol. 37, no. 1, pp. 237–254, Jan. 2021, https://doi.org/10.1016/j.ijforecast.2020.05.002.

[21] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, Jul. 2021, Art. no. e623, https://doi.org/10.7717/peerj-cs.623.

[22] J. A. Fadhil and Q. I. Sarhan, "Internet of Vehicles (IoV): A Survey of Challenges and Solutions," in *2020 21st International Arab Conference on Information Technology (ACIT)*, Giza, Egypt, Aug. 2020, pp. 1–10, https://doi.org/10.1109/ACIT50332.2020.9300095.