Iraqi Journal of Science, 2023, Vol. 64, No. 8, pp: 4160-4168
DOI: 10.24996/ijs.2023.64.8.36

ISSN: 0067-2904

Network Traffic Prediction Based on Time Series Modeling

Naors Y. Anad AlSaleem
Department of Computer Science, College of Education, University of Al-Hamdaniya, Mosul, Iraq

Received: 3/4/2022 Accepted: 8/10/2022 Published: 30/8/2023

Abstract

Predicting the network traffic of web pages is one of the areas that has increased focus in recent years. Modeling traffic helps find strategies for distributing network loads, identifying user behaviors and malicious traffic, and predicting future trends. Many statistical and intelligent methods have been studied to predict web traffic using time series of network traffic. In this paper, the use of machine learning algorithms to model Wikipedia traffic using Google's time series dataset is studied. Two data sets were used for time series, data generalization, building a set of machine learning models (XGboost, Logistic Regression, Linear Regression, and Random Forest), and comparing the performance of the models using (SMAPE) and (MAPE). The results showed the possibility of modeling the network traffic time series and that the performance of the linear regression model is the best compared to the rest of the models for both series.

Keywords: Computer networks**,** network traffic modeling, time series, machine learning algorithms, XGboost.

التنبؤ بمرور الشبكة باعتماد نمذجة السلاسل الزمنية

نورس يونس عناد

قسم علوم الحاسوب, كلية التربية, جامعة الحمدانية, نينوى, العراق

الخلاصة

ان التنبؤ بحركة مرور الشبكة لصفحات الويب احد المجالات التي ازداد عليه التركيز في السنوات الاخيرة. ان نمذجة تلك الحركة تساعد في ايجاد استراتيجيات لتوزيع احمال الشبكة بالإضافة الى تحديد سلوكيات المستخدم والمرور الضار وتوقع الاتجاهات المستقبلية. ولنمذجة حركة مرور الويب تم دراسة العديد من الطرق الاحصائية والذكائية للتنبؤ بها باستعمال السلاسل الزمنية لمرور الشبكة. في هذه الورقة تم دراسة استعمال خوارزميات تعلم الالة لنمذجة حركة مرور موقع ويكيبيديا باستعمال مجموعة بيانات السلاسل الزمنية من google. تم استعمال مجموعتين بيانات لسلاسل زمنية وتعميم البيانات وبناء مجموعة من نماذج تعلم الالة (و xgboost و Logistic Regression و Linear Regression و Random Forest) والتحقق من اداء النماذج باستخدام (SMAPE) و (MAPE) . اظهرت النتائج امكانية نمذجة السلاسل الزمنية لمرور الشبكة وان اداء نموذج الانحدار الخطي هو الافضل بالمقارنة مع بقية النماذج لكلا السلسلتين.

* Email: nawrasyounis@uohamdaniya.edu.iq

1. Introduction

With the advancement of information technology, network resource allocation has become an important research topic in recent years [1]. An optimal resource allocation mechanism can ensure that important or high-priority traffic is not delayed or ignored when the network is overloaded or congested. At the same time, it ensures the efficient operation of the network [2,3]. The development and maturity of network traffic prediction technology makes it possible to create dynamic resource allocation based on accurate traffic forecasts [4]. According to different applications, network traffic forecasting can usually be divided into short-term and long-term forecasting. In general, long-term forecasting usually relies on historical data with relatively large details, a long period of months and days, to analyze, model, and predict future traffic flow [5]. Changes usually pay more attention to the accuracy of the trend rather than the absolute accuracy of the expected value. Short-term forecasting requires real-time performance and predicting future network traffic in seconds or even smaller ranges. Modern methods of network management can control network traffic dynamically. Analyzing network traffic in real-time and making quick decisions may be futile, so long-term network traffic modeling is the solution for better network management [6,7]. To predict network traffic systematically, machine learning models can be trained on network traffic and predict future network traffic. In this paper, machine learning algorithms (XGboost, Logistic Regression, Linear Regression, and Random Forest) were applied, and the performance of each model was evaluated using (SMAPE) and (MAPE). Specific and relevant features are used for Wikipedia's network traffic. Then it is used to generalize the data, apply models, and compare the performance of the models.

2. Related work

Several models have been proposed for network traffic modeling and prediction that implement supervised machine learning algorithms used dynamically or statically. Most attempts to model network traffic are based on real-time, typically using memory-use prediction algorithms. This section summarizes the work related to network traffic modeling. To predict network traffic, the researchers [8] used the time-series prediction data set of web traffic for Wikipedia articles. Using the RNN seq2seq model, they created a time-series model. It next looks at using the Symmetric Mean Absolute Percentage Error (SMAPE) to evaluate the produced model's overall performance and correctness. [9] predicted encrypted user traffic. They created a representative traffic data set, including video and web traffic, and used two comparison models (ARIMA and LSTM). The results showed the superiority of (LSTM) in accuracy and time.

In [10], the researchers presented the NetScrapper classifier, a flow-based network traffic classifier for online applications NetScrapper classifies 53 web apps, including Amazon, YouTube, Google, Twitter, and many others, using three machine learning models: K-Nearest Neighbors (KNN), Random Forest (RF), and Artificial Neural Network (ANN). The network traffic dataset contains 35,77296 stream packets with 87 different features. In [11] long-term memory (LSTM) and online sequential extreme learning machine (OS-ELM), the real traffic of the Chilean ISP was used to predict the network traffic. The results concluded that OS-ELM is superior to LSTM in computational cost. To classify the network traffic in terms of applications used [12], the researchers proposed an intelligent traffic management model using deep learning that includes multiple decision tree-based models.

The proposed model deploys a blending set-learning method for merging tree-based classifiers to increase generalization accuracy. In [13], multivariate time series models were studied using time series analysis such as clustering and sequencing to create sequence models

using long-term memory architecture (LSTM). The dataset to which the study application is applied is Wikipedia web page traffic. The dataset contains about 145,000 web pages and corresponding web page traffic from July 2015 to December 2016. In [14], an updated version of the 2018-2020 Wikipedia page views dataset is used, and an LSTM Neural Network with Distributed Asynchronous Training is used. A predictive model was built using the heavy rain strategy in training to achieve parallel training.

3. Machine learning model

This study used four models to model time series to predict network traffic. Two modeling concepts were used: regression-based modeling (logistic regression and linear regression) and modeling based on ensemble learning with both bagging and boosting using two algorithms (XGBoost and RandomForest).

3.1. Logistic Regression

Logistic regression was chosen because it is simple, easy to implement and comprehend, and has low computational requirements. It's also one of the most widely used supervised learning algorithms nowadays. Because of these properties, logistic regression is a strong option for problems involving vast volumes of data and quick, possibly automated decisions, such as network traffic [15,16]. The logistic regression model, as shown in the following equation, links the likelihood of an outcome with a series of potential predictor variables:

$$l = log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_i X_i \quad (1)$$

where p is the probability of the visited page, β_0 is an intercept term, $\beta_0, \beta_1 \ldots . \beta_i$ are coefficients associated with each variable $X_0, X_1 \ldots X_i$.

3.2. Linear Regression

A collection of independent variables is utilized to estimate a dependent response variable using the linear regression approach. The method attempts to find regressor coefficients β that best represent a linear relationship between the response variable Y and the regression variable X. For n observations, each consisting of k regression variables, the model can be described in the following equation[17]:

$$Y = X\beta + \varepsilon \quad (2)$$

3.3. Random Forest:

The Random Forest constructs decision trees using a random approach: each tree is trained on randomly picked objects and randomly selected features, a technique known as the random subspace method. After that, a forecast may be made based on the results acquired for each tree [18]. The outcome can be determined in various ways, including utilizing a rapid majority vote or an average. Using such a random strategy can lower the model error, i.e., the spread of model predictions [19].

3.4. XGboost

Chen and Guestrin presented extreme gradient boosting (XGBoost) in 2016. This approach enhances the gradient boosting-based calculation method for the objective function and saves computation time. Parallel computing is automatically achieved throughout the training phase to address large data science issues rapidly and precisely [20]. XGBoost main premise is to learn new features by including a tree structure, fitting the residuals of the final prediction, and calculating the sample score. The sample ultimate prediction score may be calculated by

aggregating the scores of each tree [21]. The framework of XGBoost is briefly described in the following paragraphs.

The estimated output of the gradient boosting tree model can be expressed as the sum of the prediction scores $\hat{y}$ of all trees:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i),\ f_k \in T \quad (3).$$

where T is the space of regression trees, and K is the number of regression trees, xi represents the features corresponding to sample i. For a given dataset, there is a prediction score $f_k(x_i)$, also known as leaf weight, for each leaf node j.

4. Methods

The data set was obtained, initialized, and divided into a training set and a test set to predict network traffic by adopting time series. The selected models were applied, and the performance of each module was measured using mean absolute percentage error and symmetric mean absolute percentage error.

4.1. Dataset

The data set used in this paper is a time-series forecast of Google web traffic consisting of 141,385 Wikipedia articles. The dataset includes a chronological order field representing the time series or multiple points. Each time series represents some daily views of a different Wikipedia article, from 7/1/2015 to 9/10/2017 [13,14,22,23]. Figure 1 shows samples of daily views of a different Wikipedia article.

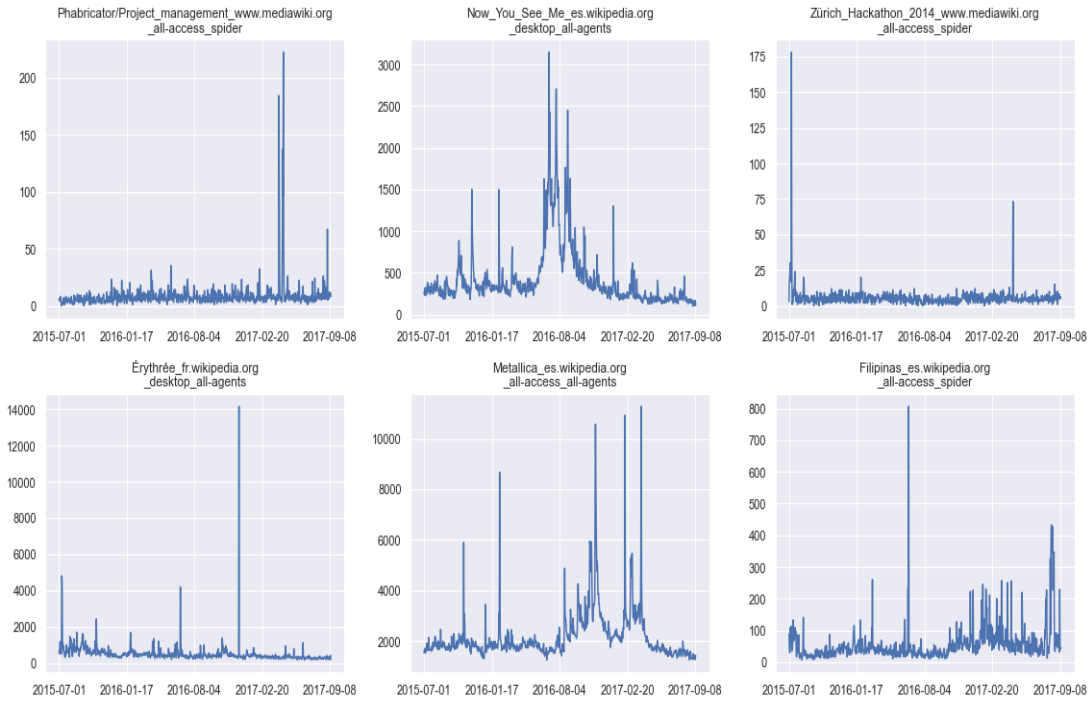

Figure 1: Samples of daily views of a different Wikipedia article

4.2. Proposed framework

The framework includes data recall, setting target values, and data preprocessing that includes data generalization to reduce the effect of outliers and overfitting. Then the data set was divided into training data and test data in a ratio (33:67), with the selected models being trained and the performance of each model being measured as shown in Figure (1).

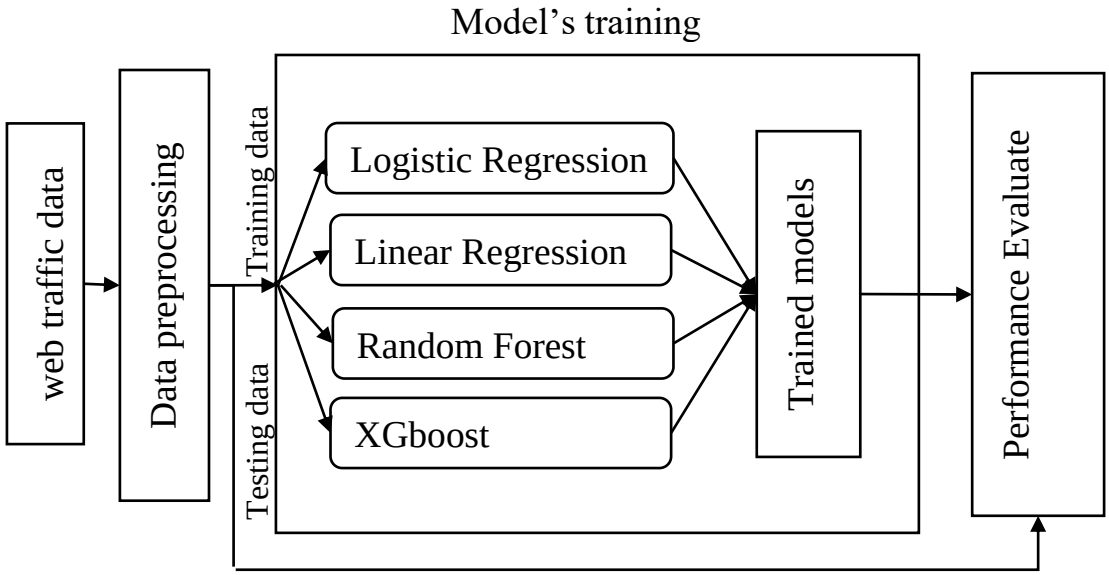

Figure 2 : Proposed framework.

4.3. Evaluation criteria and experimental results

The results have been analyzed and the proposed models trained in the open-source Python tool. To evaluate the models, comparisons, and proposed results, the data has been divided into two groups: the first for training and the second for testing. To evaluate the performance of the selected models, two scales, mean absolute percentage error (MAPE) and symmetric mean absolute percentage error (SMAPE) [24,25], were used.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \quad (4)$$

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|P_t - A_t|}{(A_t + P_t)/2} \quad (5)$$

Where:
n: is the number of instance testing,
At: is the actual value,
Pt: is the prediction value.

After training, the models were tested. Table (1) compares the performance of the selected models.

Table 1 Comparison of the performance of the models.

Models	MAPE	SMAPE
Logistic regression	**33.78**	**39.75**
Linear Regression	**19.88**	**20.06**
Random Forest	**39.22**	**50.63**
XGboost	**48.00**	**67.07**

To display the prediction results against the true values and the prediction periods for the time series and anomalies, Figures (2, 3, 4, and 5) show the performance of each model.

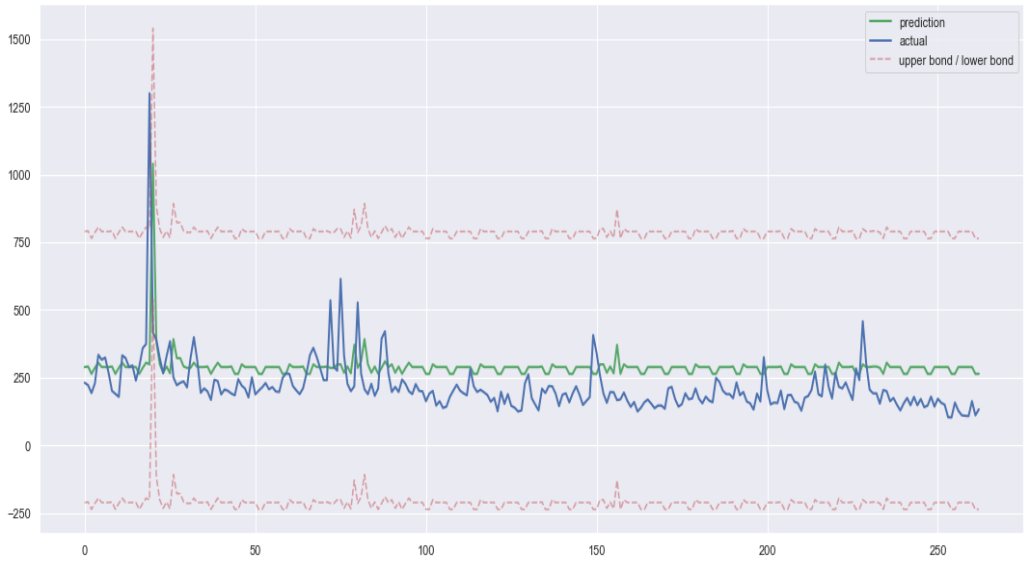

Figure 3: Logistic regression performance

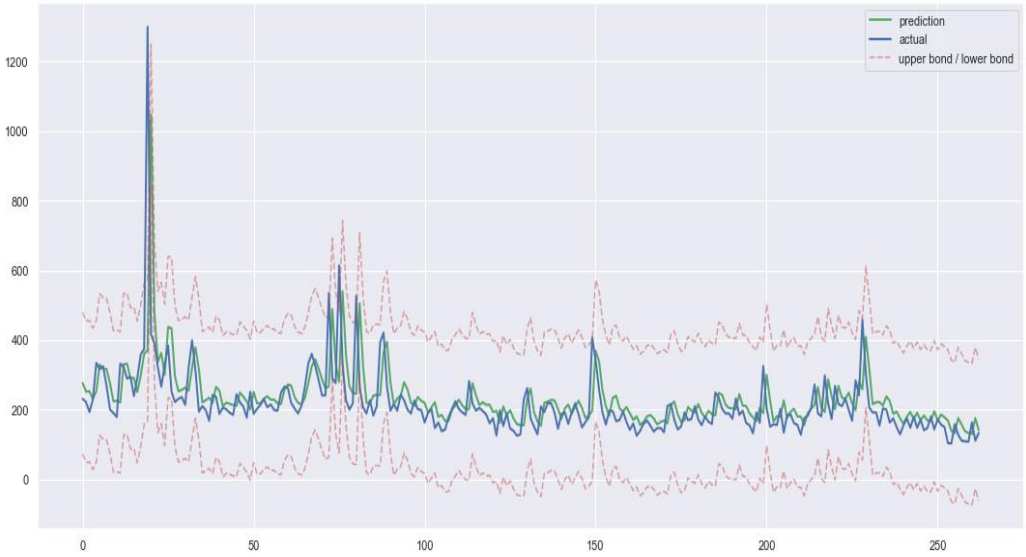

Figure 4 : The performance of linear regression

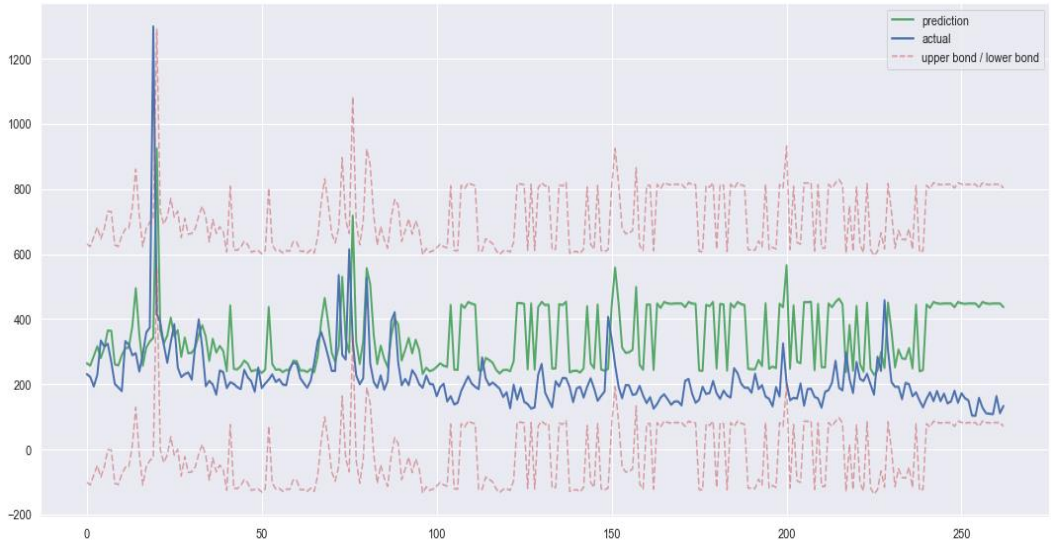

Figure 5: The performance of random forests

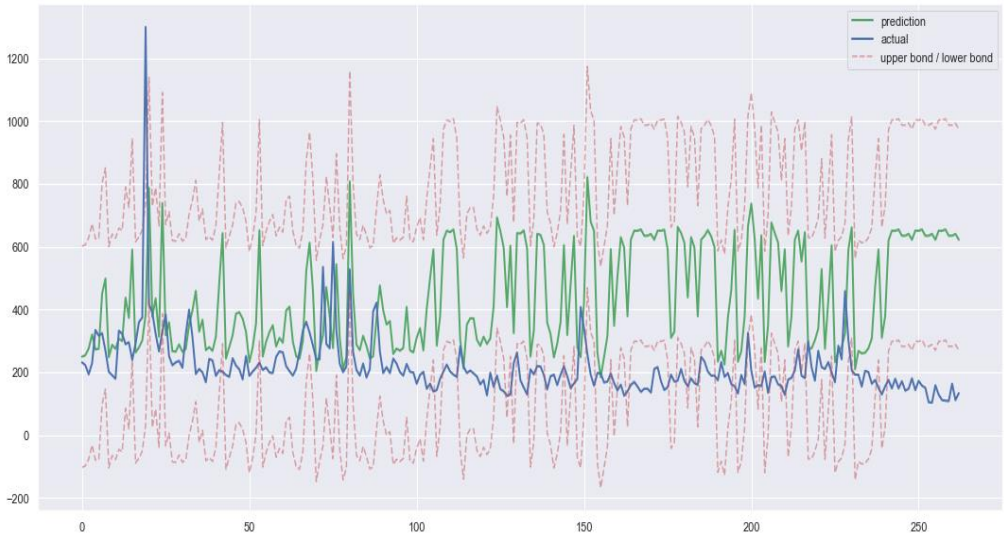

Figure 6: The performance of XGboost

5. Discuss the results

By comparing the performance results of the selected models in Table 1, it is noted that linear regression achieved the best performance on both scales, as it achieved 19.88 on the MAPE scale and 20.06 on the SMAPE scale. Both results show the possibility of modeling network traffic. At the same time, the logistic regression followed the linear regression, while the performance of random forests and XGboost was lower on both scales. The results indicate that network traffic can be modeled and that linear regression models are better for modeling this type of time series. In displaying the predictive values versus the real values, we find that the linear regression as in Figure 3 was the best compared to the rest of the models and over the length of the test examples.

After analyzing the results, the linear regression model was the best. To determine the effectiveness of this method, it is compared with previous work that used the same data set as in Table 4.

Table 2 A Comparison with the previous work.

Paper	year	Model	MAPE	SMAPE
Yang et al [26]	2020	Vector Auto Regression	36%	
Petluri et al [8]	2018	RNN		35%
Ours	**2022**	**linear regression**	**19%**	**20%**

Compared to the previous work, it can be concluded that linear regression was the best.

6. Conclusion

Traffic modeling is a critical task in network management and cybersecurity. Network traffic is modeled based on four machine learning algorithms: logistic regression, linear regression, RandomForest, and XGboost. The study found that supervised machine learning algorithms can model network traffic for time series. The results showed that these features could be modeled and categorized. The best performance was achieved by the linear regression algorithm, and it achieved 19.88 on the MAPE scale and 20.06 on the SMAPE scale.

References:

[1] W. Qin, S. Chen, and M. Peng, "Recent advances in Industrial Internet: insights and challenges," *Digit. Commun. Netw*., vol. 6, no. 1, pp. 1–13, 2020.

[2] X. Jiang, F. R. Yu, T. Song, and V. C. M. Leung, "Resource allocation of video streaming over vehicular networks: A survey, some research issues and challenges," *IEEE Trans. Intell. Transp. Syst*., pp. 1–21, 2021.
[3] I. Nurcahyani and J. W. Lee, "Role of machine learning in resource allocation strategy over vehicular networks: A survey," *Sensors* (Basel), vol. 21, no. 19, p. 6542, 2021.
[4] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Trans. Wirel. Commun*., vol. 16, no. 6, pp. 3899–3912, 2017.
[5] C. Gijón, M. Toril, S. Luna-Ramírez, M. L. Marí-Altozano, and J. M. Ruiz-Avilés, "Long-term data traffic forecasting for network dimensioning in LTE with short time series," *Electronics* (Basel), vol. 10, no. 10, p. 1151, 2021.
[6] M. F. Iqbal, M. Zahid, D. Habib, and L. K. John, "Efficient prediction of network traffic for real-time applications," *J. Comput. Netw. Commun*., vol. 2019, pp. 1–11, 2019.
[7] M. Abbasi, A. Shahraki, and A. Taherkordi, "Deep learning for network traffic monitoring and analysis (NTMA): A survey," *Comput. Commun*., vol. 170, pp. 19–41, 2021.
[8] N. Petluri and E. Al-Masri, "Web traffic prediction of Wikipedia pages," in 2018 *IEEE International Conference on Big Data* (Big Data), 2018.
[9] Q. He, G. P. Koudouridis, and G. Dan, "A comparison of machine and statistical time series learning for encrypted traffic prediction," in 2020 *International Conference on Computing, Networking and Communications* (ICNC), 2020.
[10] A. A. Ahmed and G. Agunsoye, "A real-time network traffic classifier for online applications using machine learning," *Algorithms*, vol. 14, no. 8, p. 250, 2021.
[11] Ons Aouedi , Kandaraj Piamrat , Benoît Parrein, "Decision tree-based blending method using deep-learning for network management," in *IEEE/IFIP Network Operations and Management Symposium*, 2022.
[12] F. Rau, I. Soto, P. Adasme, D. Zabala-Blanco, and C. A. Azurdia-Meza, "Network traffic prediction using online-sequential extreme learning machine," in 2021 *Third South American Colloquium on Visible Light Communications* (SACVLC), 2021.
[13] V. P. Wunnava, "Exploration of Wikipedia traffic data to analyze the relationship between multiple pages," *North Carolina, Chapel Hill*, 2020.
[14] R. Casado-Vara, A. Martin del Rey, D. Pérez-Palau, L. de-la-Fuente-Valentín, and J. M. Corchado, "Web traffic time series forecasting using LSTM neural networks with distributed asynchronous training," *Mathematics*, vol. 9, no. 4, p. 421, 2021.
[15] S. Isak-Zatega, A. Lipovac, and V. Lipovac, "Logistic regression based in-service assessment of mobile web browsing service quality acceptability," *EURASIP J. Wirel. Commun. Netw*., vol. 2020, no. 1, 2020.
[16] X. Zhu, Y. Nie, S. Jin, A. Li, and Y. Jia, "Spammer detection on online social networks based on logistic regression," in *Web-Age Information Management, Cham*: Springer International Publishing, 2015, pp. 29–40.
[17] R. Gangurde and B. Kumar, "Web page prediction using genetic algorithm and logistic regression based on weblog and web content features," in 2020 *International Conference on Electronics and Sustainable Communication Systems* (ICESC), 2020.
[18] T.H. Lee, A. Ullah, and R. Wang, "Bootstrap aggregating and random forest," in *Macroeconomic Forecasting in the Era of Big Data, Cham*: Springer International Publishing, 2020, pp. 389–429.
[19] M. Alsaleem and S. Hasoon, "Predicting bank loan risks using machine learning algorithms," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 14, no. 1, pp. 159–168, 2020.
[20] M. Y. A. Alsaleem and S. O. Hasoon, "Comparison of dt& gbdt algorithms for predictive modeling of currency exchange rates," *EUREKA Phys. Eng*., vol. 1, pp. 56–61, 2020.
[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
[22] R. V. Shaptala and G. D. Kyselev, "Using graph embeddings for Wikipedia link prediction," Bulletin of National Technical University "KhPI". *Series: System Analysis, Control and Information Technologies*, vol. 0, no. 1, pp. 48–52, 2019.
[23] "web-traffic-time-series- forecasting," kaggle.com, 2022. [Online]. Available: https://www. kaggle. com /c/web-traffic-time-series-forecasting.

[24] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.

[25] A. S. Ahmar, "Forecast error calculation with mean squared error (MSE) and mean absolute percentage error (MAPE)," *jinav j. inf. vis*., vol. 1, no. 2, pp. 94–96, 2020.

[26] Yang, Y., Lu, S., Zhao, H., & Ju, X. (2020, September). Predicting Monthly Pageview of Wikipedia Pages by Neighbor Pages. In Proceedings of the 2020 *3rd International Conference on Big Data Technologies* (pp. 112-115).

[27] Paun, K. P., & Makwan, C. H. ,"A survey on web traffic forecasting on time series data," *J. Appl. Sci. Comput*, vol. 6, pp. 3588-3594, 2019.