Review of Computer Engineering Research
2022 Vol. 9, No. 2, pp. 96-108
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v9i2.3065

MACHINE LEARNING BASED NETWORK TRAFFIC PREDICTIVE ANALYSIS

Sridevi Ponmalar P[1]
Ayushi Kamboj[2+]
Vikram Shete[3]
Harikrishnan R[4]

[1]*Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Chennai. India.*
[1]*Email: sridevip@srmist.edu.in*
[2,3,4]*Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India.*
[2]*Email: ayushikamboj3322@gmail.com*
[3]*Email: dr.vikram.shete@gmail.com*
[4]*Email: dr.rhareish@gmail.com*

(+ Corresponding Author)

Article History
Received: 28 April 2022
Revised: 10 June 2022
Accepted: 30 June 2022
Published: 18 July 2022

Keywords
Datacenter
Machine learning
Network traffic prediction
Networking
Prediction
Scopus.

ABSTRACT

Digital enterprises that use various Internet of Things implementation prototypes, such as cloud, mobile, and edge equipment, are experiencing unprecedented traffic volume and dynamicity. Data center networks (DCN) have faced various issues due to the transient and random nature of traffic created by services and apps. The primary objective of this paper is to predict the network traffic using the machine learning (ML) models before the performance of the network start degrading. Because in the last decade, ML has had a tremendous impact on handling the massive amount of data. With the increase in complexity and traffic, we tried to implement four ML models such as K – Nearest Neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), and Decision Tree (Tree), with tuned sub-parameters to predict the network traffic. We create a matching ML environment based on a sequential database and provide a comparison table of mean square error (MSE), root mean square error (RMSE), mean absolute error (MSE), and coefficient of determination (R2) for each prototype. The simulation results show that the GB with different types is the best-suited model for predicting the network traffic with performance matrix parameters such as MSE 0.001 and RMSE of 0.030. Therefore, the Orange tool is used to stimulate the predictive models.

Contribution/Originality: This research is focused on predicting the network traffic by using the different prediction regression models such as KNN, RF, GB, and DT with different sub-parameters.

1. INTRODUCTION

Internet traffic has expanded rapidly in the recent decade due to the advent of new technologies, industries, and applications. Wireless network traffic is growing due to increased smartphone subscriptions and a steady increase in average data volume per subscription. Due to the rapid growth in sensing data and quick reaction requirements, high-speed transmission has become a critical concern in the Internet of Things (IoT) delivery network. Assigning optimal channels in the wireless IoT delivery network ensures high-speed transmission. Many cloud resource companies have invested in larger data centers (DCs) to handle the computational equipment required to provide various services. As a result, academics and industry practitioners have focused on developing network fabrics that can seamlessly interconnect and manage traffic throughout these DC whiles being performant and efficient. Unfortunately, DC operators are often hesitant to share specific application specifications, making it impossible to evaluate the feasibility of any particular architecture. Because of the vastly different engineering constraints that arise when interconnecting many extremely interdependent homogeneous endpoints in a comparatively tiny physical space versus weakly paired heterogeneous endpoints dispersed throughout the world, DC is revolutionizing how we customize networks.

Cloud computing services are the solutions for a substantial percentage of the growth in Internet data traffic in DCs. Because of the unknown aspects and uncertainty of the traffic produced by such applications, traffic increase presents difficulties for inter and intra-center networks. While these physical characteristics are critical in many parts of network and protocol architecture, others necessitate a thorough understanding of the network's demands from end hosts. Therefore, the main objective of this paper is to forecast the network traffic before it affects the system's performance. We tried implementing different ML models such as GB, KNN, DT, and RF to predict the network traffic.

For this paper, we used one of the appliance data from the data center of PTC software India Private Limited. Therefore, we are using real-time, sequential data for this work. In general, obtaining real-time data from DCs for analysis is extremely challenging.

The following are the paper's main contributions:

1. Pre-processing and analysis of time series data collected from the DC.
2. Implement the different prediction models with different sub-parameters.
3. A performance evaluation of different machine learning prediction models and their comparative analysis.

The remaining paper is organized in the following manner. The related work for predicting network traffic are detailed in Section 2. Section 3 explains the proposed methodology for this paper, and in section 4, the performance analysis and prediction models' evaluation. Section 5 discusses the results. Finally, section 6 gives a significant conclusion as well as future work.

2. REVIEW OF LITERATURE

DC utilization and energy consumption will continue to climb as mobile data traffic increases, and the cloud computing environment evolves. Compared to state-of-the-art approaches, empirical results suggest PATGEN boosts productivity and decreases average latency. The algorithm will be improved in the future by simultaneously optimizing more targets, according to the author Amir Iranmanesh. In addition, the proposed approach will be used with additional intelligence technologies [1]. Based on historical statistics from prior measurement periods, author Aggelos Lazaris proposes the DeepFlow, a forecast evaluation approach for Software Defined Network (SDN) that utilizes the current Ternary Content-Addressable Memory (TCAM) to configure quantification guidelines for substantial flows and a productive ML algorithm to forecast the size of the remainder of the flows that cannot be tracked with accurate match norms. In the future, author Aggelos Lazaris hopes to improve DeepFlow by incorporating more sophisticated flow interactions and network signals, reducing the number of precise flow measurements required [2].

ML has lately been offered as a way to process that data [3]. Unfortunately, DC operators are often unwilling to discuss the specific needs of their applications, making it difficult to assess the feasibility of any given architecture. Arjun Roy and Hongyi Zeng, the authors, examine network traffic in certain of Facebook's DCs. The authors discuss how network traffic in Facebook's DCs differs in terms of localization, stability, and predictability, as well as the consequences of network layout, traffic engineering, and switch design [4]. LSTM learners beat state-of-the-art conventional machine learning forecasting prototypes, according to the author Desta Haileselassie Hagos, the adaptability and reliability of the suggested methodology and its potential for analyzing TCP transmission statuses associated with network congestion from passive measures through a detailed empirical evaluation of numerous scenarios. Deep Learning appears to be a feasible method for analyzing system-wide TCP states from passive measurement, relying on findings from simulated and actual conditions [5].

The network's throughput and usage can be improved and optimized using ML approaches. In particular, Neural Networks (NN) and Reinforcement Learning (RL) have depicted assurance in collaborating with complicated network processes and monitoring challenges. By implementing a Deep Q-Network (DQN) agent, the author EL Hocine

Bouzidi postulated an SDN-based standards placement methodology to adaptively anticipate traffic congestion by primarily NN and learning optimal path and reroute traffic to enhance network utilization. The suggested approach can significantly increase network performance by lowering link utilization, packet loss, and E2E (End-to-End) delay, according to numerical data obtained through emulation utilizing the ONOS controller and Mininet [6]. Fuyou Li looks into the temporal features of Twitter traffic and develops a Twitter traffic forecasting paradigm that combines statistical analytic and machine learning methodologies. According to empirical results based on real Twitter traffic datasets gathered in central London, the proposed paradigm delivers substantial forecasting precision with minimum processing complexity and little necessity for dataset size [7].

SDN-IoT (Software-Defined Networking-based Internet of Things) has recently been suggested as a way to optimize broadcast efficiency. As a result, Fengxiao Tang presents a new deep learning-based traffic load forecasting methodology for predicting future network traffic load and congestion. The channel is then intelligently assigned to every link in the SDN-IoT network using a Deep Learning-based Partially Channel Assignment Algorithm (DLPOCA). TPDLPOCA, a unique channel assignment protocol that combines learning-based forecasting and Partially Overlapping Channel Assignment (POCA) to proactively minimize future congestion and quickly allot relevant channels in SDN-IoT, is also introduced. The simulation outcomes show that the proposed method outperforms traditional channel assignment methods [8]. The software-defined networking (SDN) methodology is used to achieve global knowledge acquisition. After that, the SAGIN segment's resource utilization is anticipated using a traffic forecasting methodology relying on an autoregressive moving average (ARMA) model. After analyzing connection efficiency, the Adaboost methodology is implemented to locate network nodes for DTL deployment based on their data transmission capability. According to simulation results, the postulated Adaboost-based link management methodology is practical and efficient [9]. Using genuine DCN traces, author Jeandro Bezerra evaluates the efficacy of various elephant flow predictions. The sequential traffic from a Facebook DC is utilized to design elephant flow estimators and offer descriptive statistics and verdicts about the flows on a short-term basis. The research represents a hybrid forecasting prototype by mixing parts of the FARIMA and Recurrent Neural Network (FARIMA-RNN) models. A framework premised on the ranking of forecasting precision metrics is utilized to analyze the effectiveness of the hybrid prototype with the Auto-Regressive Integrated Moving Averages (ARIMA), GARCH, RBF, MLP, and LSTM models. According to the results [10], the FARIMA-RNN model has lower error rates than the other forecasting models [10].

Knowledge-defined networking (KDN) promises to enable flexible and self-driven optical networks.

Masoud Vejdannik, the author, focuses on approximating the quality of service (QoS) for unassociated light paths, which is an essential feature of the ecosystem. KDN is a machine learning-based network control plane remedy that addresses the inescapable troubles that emerge as networks become more independent and efficient. For classification and regression operations, five machine learning prototypes are evaluated. Support-vector machine and probabilistic neural network (PNN) models are merely used for categorization. In contrast, multilayer perceptron, radial basis function, and generalized regression neural network (GRNN) models are utilized for both regression and classification [11]. The best-transmitted forecasting (categorization methodology) and optical signal-to-noise ratio assessment (regression strategy) precision are achieved by PNN (with an aggregate precision of 99.6 0.5 percent) and GRNN (with an R-squared value of 0.957) [11].

Muhammad Faisal Iqbal looked at a number of forecasts with the hopes of finding one with high precision, low processing complexity, and low power consumption. A variety of estimates from three categories, including traditional time series, artificial neural networks (ANN) and wavelet transform-based predictors were investigated. In terms of processing complexity and power consumption, author Muhammad Faisal Iqbal provides an evaluation of precision and cost. According to the findings, a double exponential smoothing estimator provides a decent ratio between efficiency and cost overhead [12].

Rishabh Madan presents the Discrete Wavelet Transform (DWT), ARIMA model, and the RNN-related methodology for predicting computer network traffic. The traffic data is decomposed into non-linear (approximate) and linear (detailed) constituents using the discrete wavelet transform. Inverse DWT is used to reconstitute detailed approximation components, and ARIMA and RNN are used to make predictions. Internet traffic is sequential data employed to forecast future computer network traffic trends. The suggested methodology is simple to deploy and computationally less expensive, making it suitable for use in DCs to improve QoS while lowering costs [13].

Forecasting in SDNs allows for proactive and global optimization. For trigonometric Fourier-based traffic prototypes in SDNs, Grzegorz Rzym provides the shrinkage and selection heuristic methodology. Significant constraints are resolved and considered because the method has been devised to work as part of a complicated routing-assistance system. The forecasting methodology must be versatile to networks with varying traffic loads because the framework is independent of traffic and topology (such as those seen within Intra-DCNs). The proposed heuristic is compared to the well-known Lasso approach, which has been demonstrated to be accurate. The outcomes suggest that the approach can maintain a comparable degree of forecasting precision. Furthermore, the author Grzegorz Rzym operates in a manner that is always substantially faster, in accordance with the primary goal. Computation times are decreased by up to 50 times in some circumstances [14]. A high-speed diesel pump's effectiveness is measured using the vibration attribute as a throughput variable. Other elements that affect the high-speed diesel pump's operation and cause the change in vibration value have been discovered. Statistical Auto-Regressive Integration and Moving Average, Poisson's regression, and a few ML and Deep Learning techniques such as DT Regressor, Multi-Layer Perceptron, Linear Regression, and Long Short-Term Memory are all used to achieve this goal. In addition to the comparison, Smita Mahajan describes a distinctive layout that uses Convolution Neural Network (CNN) and LSTM. The results and comparisons reveal that the proposed unique Convo-LSTM model improves efficiency and aids in the prediction of the effectiveness of the high-speed diesel pump [15].

Numerous forecasting applications have benefited from machine learning, including traffic classification based on flow data or attributes. On the other hand, such technologies endure a data asymmetry challenge, in which some apps have less data flow and are therefore more difficult to predict. To recognize the application traffic category, the author Shi Dong utilizes network flow-level features. Moreover, a new support vector machine (SVM) technique called cost-sensitive SVM (CMSVM) is proposed to overcome the imbalance dilemma in network traffic detection. CMSVM uses an active learning multi-class SVM algorithm to allocate weight to applications dynamically. CMSVM assigns a weight to applications proactively using an active learning multi-class SVM algorithm. The CMSVM approach minimizes processing costs, improves classification precision, and solves the imbalance problem when contrasted to existing ML methodologies [16].

Because it can cope with many complicated surroundings and jobs in the upcoming communication environment, Software Defined Space Information Networks (SDSIN) are becoming more widely used in life and production. Some advanced technology relying on SDN are increasingly being used in satellite networks due to the efficiency advantages of SDN. First, Sheng Qi, the author, examines the obstacles and challenges of traffic engineering in SDSIN. Following that, a better Hidden Markov Model (HMM) for E2E traffic estimation is suggested. Additionally, modeling findings suggest that an upgraded HMM may be effectively used in SDSIN to anticipate E2E traffic [17]. SDPredictNet, an RNN framework developed by Sowmya Sanagavarapu and implemented on the SDN Controller, can forecast network traffic and alter flow tables of higher layer switches to execute routing depending on anticipated network constraints. SDPredictNet received an RMSE score of 0.07 and a 99.88 percent precision for traffic forecasting and route selection [18].

The author, Tieu Long Mai, presents enhancements to the underlying (Graph neural network) GNN paradigm in terms of training set generation and model structure. Tieu Long Mai, for example, will provide a new training set with a diverse span of architectures and traffic behaviors, focusing on difficult-to-predict combinations. Then, using more powerful activation functions, numerous channels, and a technique called global pooling improves the GNN model. On the same 13 testing sets, the precision of forecasting an ensemble of GNNs using a combination of the proposed enhancements increases dramatically, up to 11.9 percent [19].

The fifth-generation (5G) network will employ DC to help transition the wireless communication industry from specialized hardware to a more software-oriented environment. The author, Udita Paul, uses the K-means clustering technique to partition the city of Milan, Italy, into separate zones. Analyze the traffic characteristics of these zones in the city using a public big set of data from a network provider. As a result, perceive the best location for DC as a facility localization dilemma and propose utilizing Weiszfeld's technique to solve it. Furthermore, use cutting-edge RNN technologies to forecast future traffic requirements premised on previous request patterns for each location to help operations and promote dynamic resource usage [20]. Packet loss is one of the most common transmission problems, and it has a negative impact on the system's performance. If the bottleneck is identified ahead of time and appropriate steps are taken to reduce the packet generation rate at the source, packet loss can be avoided. Existing algorithms are predetermined linkages among observable states and corresponding actions. Consequently, these methods cannot adjust to new situations or learn from past errors in terms of improving their efficacy. In networking, machine learning is employed to learn from previous events and analyses the present network state to make decisions. The ability of ML to handle vast amounts of complex data is one of the reasons for its use in networking [21].

The author Yuqing Wang proposes a simple traffic forecasting method for SDN applications. Unlike conventional network traffic measurements, the proposed methodology obtains traffic statistics from the data plane using the flow-based routing concept in SDN. Second, the time-correlation hypothesis has been used to describe flow traffic because traffic is dynamic. In contrast, the time-series evaluation concept and regressive modeling methodology characterize network traffic in SDN. For traffic prediction, a completely new method is presented. Finally, the author proposes the flow-based forwarding traffic forecast approach for estimating SDN traffic. The exact method of the algorithm is shown and analyzed. Finally, the proposed method is supported by a sufficient number of well-designed and published experiments [22].

The author, Yuanqi Yang, recommends combining the Short-Time Fourier Transform (STFT) with traffic modeling to predict network traffic. Yuanqi Yang divides network traffic into high-frequency and low-frequency elements using STFT. Researchers relate the low-frequency element of network traffic as an Auto-regression (AR) prototype because it exhibits network traffic's consistency and long-range similarity. Meanwhile, a high portion of network traffic fluctuates greatly, denoting network traffic's unpredictability and thus showing it as an exponential distribution. However, because the network traffic model's forecasting error is high, an optimization function to improve network traffic forecasts and reduce errors is suggested. In simulations, the proposed forecasting method outperforms WABR and PCA [23].

By clustering the data using k-means clustering and Pearson correlation distance and then training a Deep-Belief Network (DBN) for each cluster, the author, Zahra Movahedi Nia, hopes to enhance the precision of earlier prominence forecasts. As a result, Zahra Movahedi Nia evaluates the strategy using Youtube clips as a data source, demonstrating that forecasting with a compositional prototype like DBN increases productivity significantly. The proposed methodology outperforms the previous state-of-the-art paradigm by minimizing Mean Absolute Percentage Error (MAPE) and mean RMSE (mRSE) by up to 47.86 percent and 25.18 percent, respectively, according to statistical data [24].

3. PROPOSED METHODOLOGY

In this paper, we tried four different ML algorithms to predict the network traffic. The main goals of this paper are to predict the DC traffic in advance to avoid congestion, reduce the transmission time between source and destination, reduce or mitigate the latency, and avoid packet loss. Figure 1 shows the basic block diagram for the network traffic prediction. Therefore, each step of the block diagram is discussed in the coming sections.

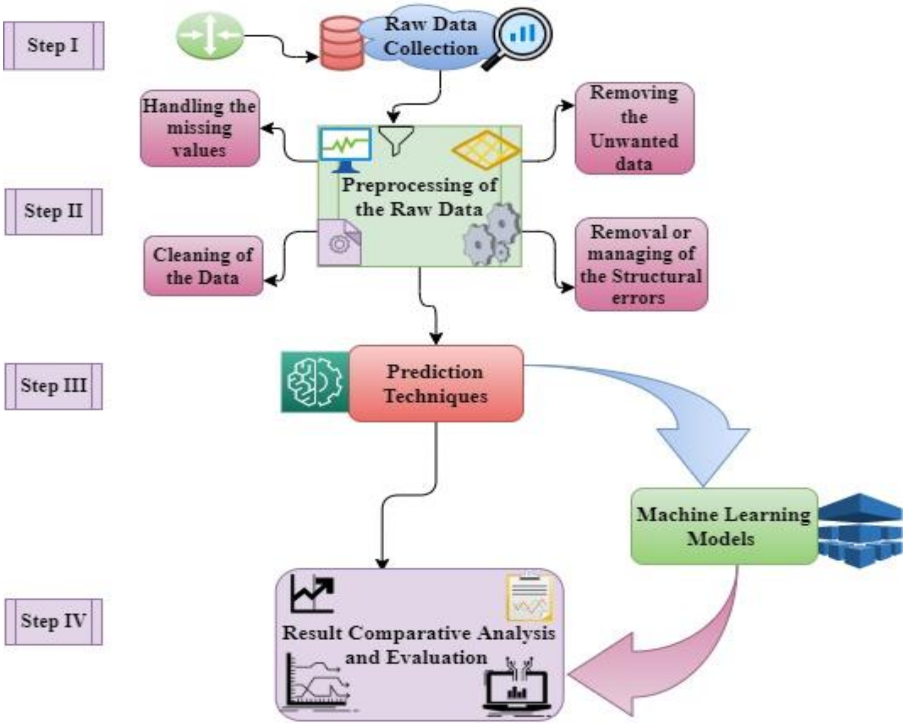

Figure 1. Depicts the basic block diagram for network traffic prediction.

A. Raw Database Collection

One of the main challenging steps is to collect real-time data from the DC. The database for the network traffic prediction is taken from one of the appliances of the DC of the PTC software India Private Ltd. Therefore, data is real-time based and is sequential.

B. Pre-processing of the Raw Dataset

The data pre-processing is the next step after the data collection that we implemented. As we collected the data from one of the appliances of the DC, our data had some missing values. To make the data into a more defined form, we used two tools two pre-process the data, an Excel tool and a tool from Orange (https://orangedatamining.com).

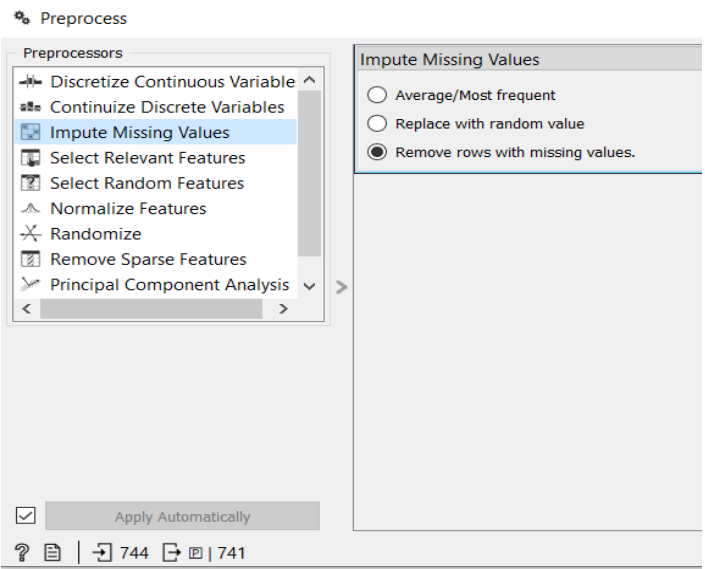

Figure 2. Shows the sub-parameter selected to remove the missing values.

Using the Excel tool, we make our data into a more defined form by removing the unwanted column from the database. Therefore, the Orange tool handled the missing values by imputing the removal of the rows with missing values, as shown in Figure 2. By selecting the parameter assigning missing values, we removed the rows with the missing values in the data. Therefore, Figure 3 shows the basic workflow to pre-process the data using the Orange tool. The raw data was uploaded in the form of a CSV file by using CSV import widgets. After that preprocess a widget was used to remove the missing values in the data and the data table widget was used to visualize the data after preprocessing. To save the preprocessed data, a save data widget was used.

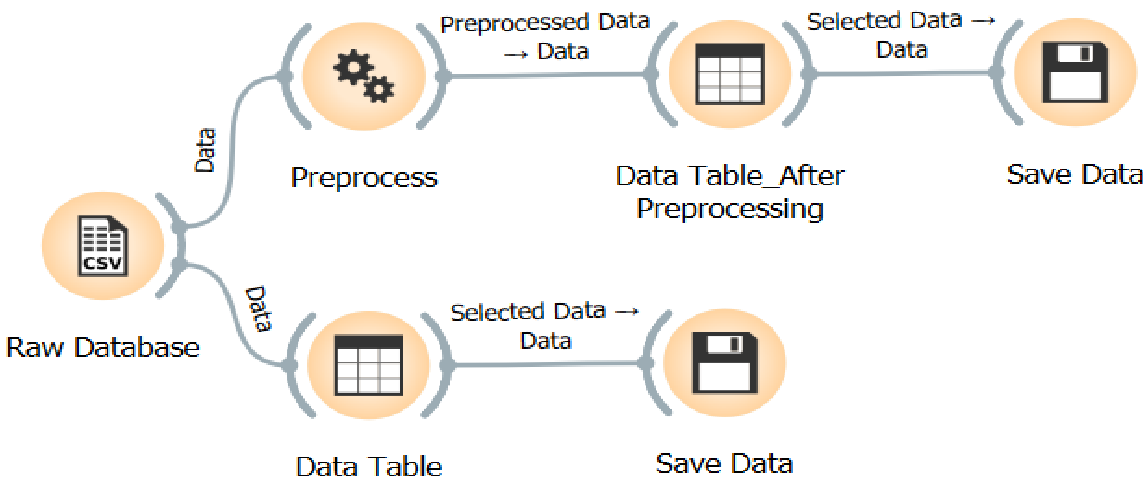

Figure 3. Depicts the workflow for pre-processing of the data.

C. Prediction Model Used

In this paper, we tried to implement the four ML models to forecast the network traffic by using the real-time data. Figure 4 shows the ML models we implemented to predict the network traffic in this paper by using real-time DC data. The models used for the network traffic prediction are GB, RF, KNN, and DT with different sub-parameters.

Despite the classification technique, the statistics index of precision cannot be used to assess prototype productivity because the accuracy of predictions characterizes the efficiency of models. The RMSE, R2, and MAE parameters were also employed to evaluate the models. The RMSE and MAE are statistical metrics that distinguish between true and false, whereas R2 depicts the difference between the actual and estimated reaction variance and how well the projected response variance explains the genuine response variance. As a result, a measurement near one is required, showing that the estimated response closely reflects the real reaction and is accurate [11].

Figure 4. Depicts the machine learning model used for predicting the network traffic.

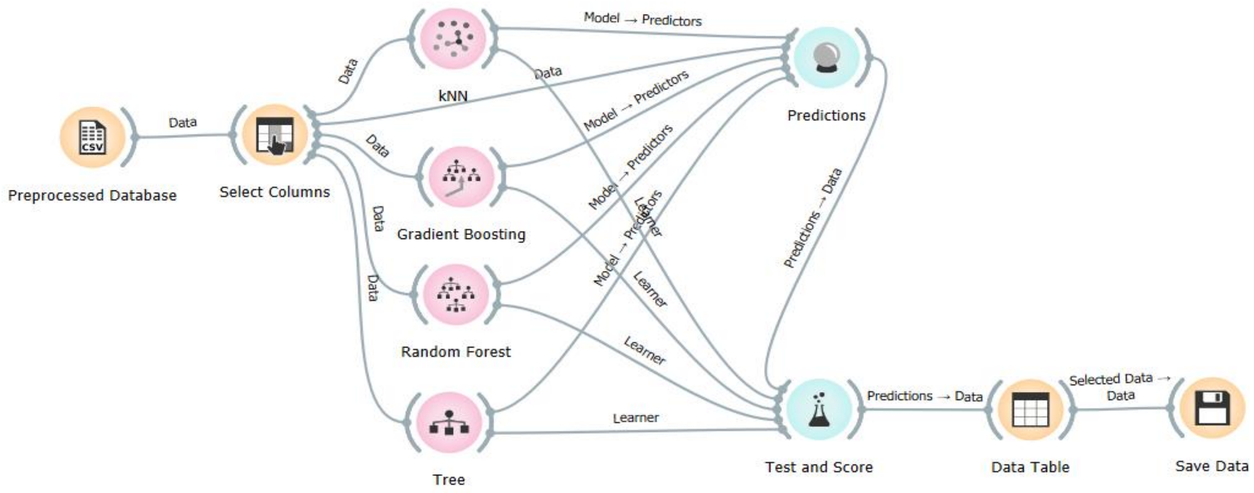

Figure 5. Depicts the workflow of the prediction model used.

Figure 5 demonstrates the workflow for predicting the network traffic. The workflow, however, consists of main parts such as training and testing of the raw database, implementation of different ML models to predict the network traffic with different sub-parameters, prediction results, and evaluation of prediction results using tests and scores. The models are tuned with different sub-parameters to predict the network traffic and to give the best outcome. Therefore, Figures 6,7,8, and 9 show the parameters selected for KNN with different metrics, for GB while varying the value of the Lambda, and different types of GB for DT and RF. At the same time, the training and test size of the data is in the proportion of 80 percent of the set for training and 20 percent for the test.

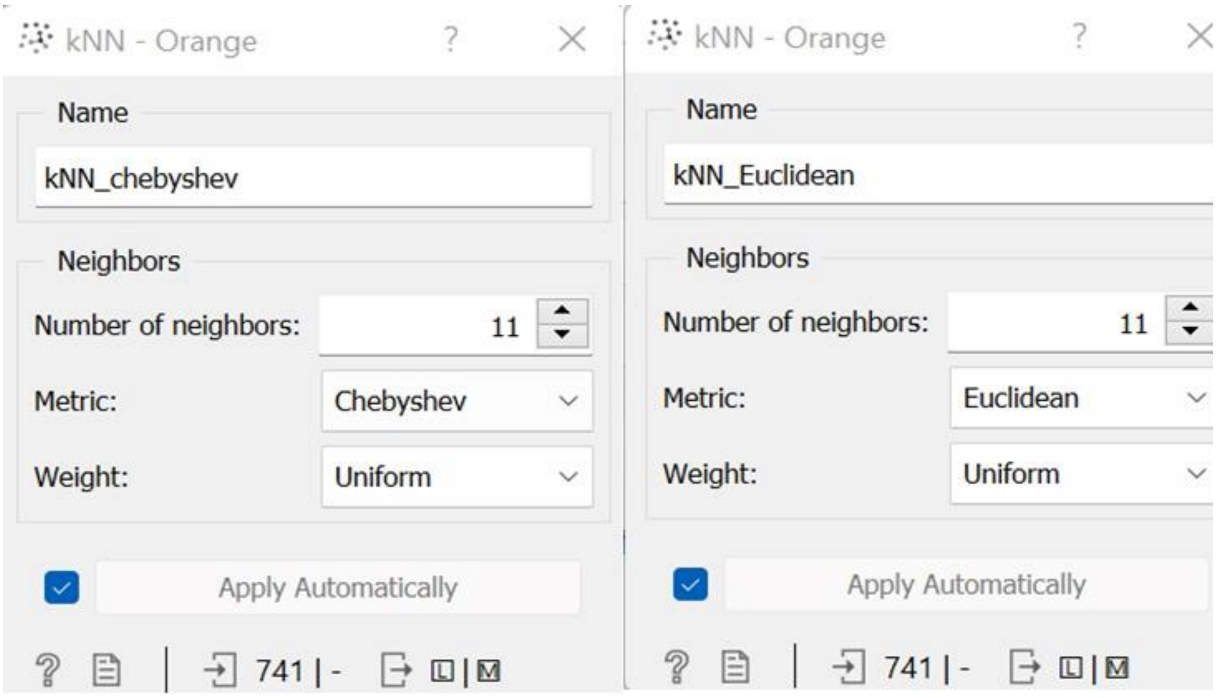

Figure 6. Depicts the subparameter selected for the KNN model with different metrics.

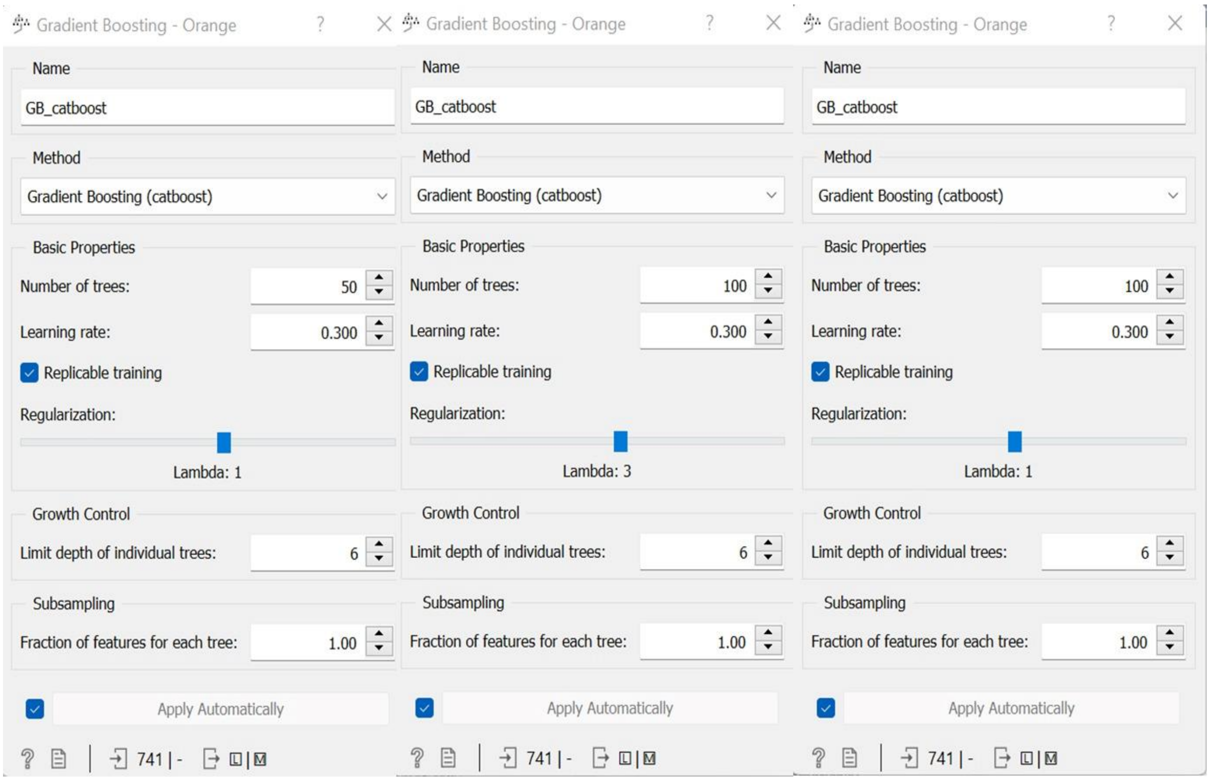

Figure 7. Demonstrate the parameters selected for the Gradient Boosting model.

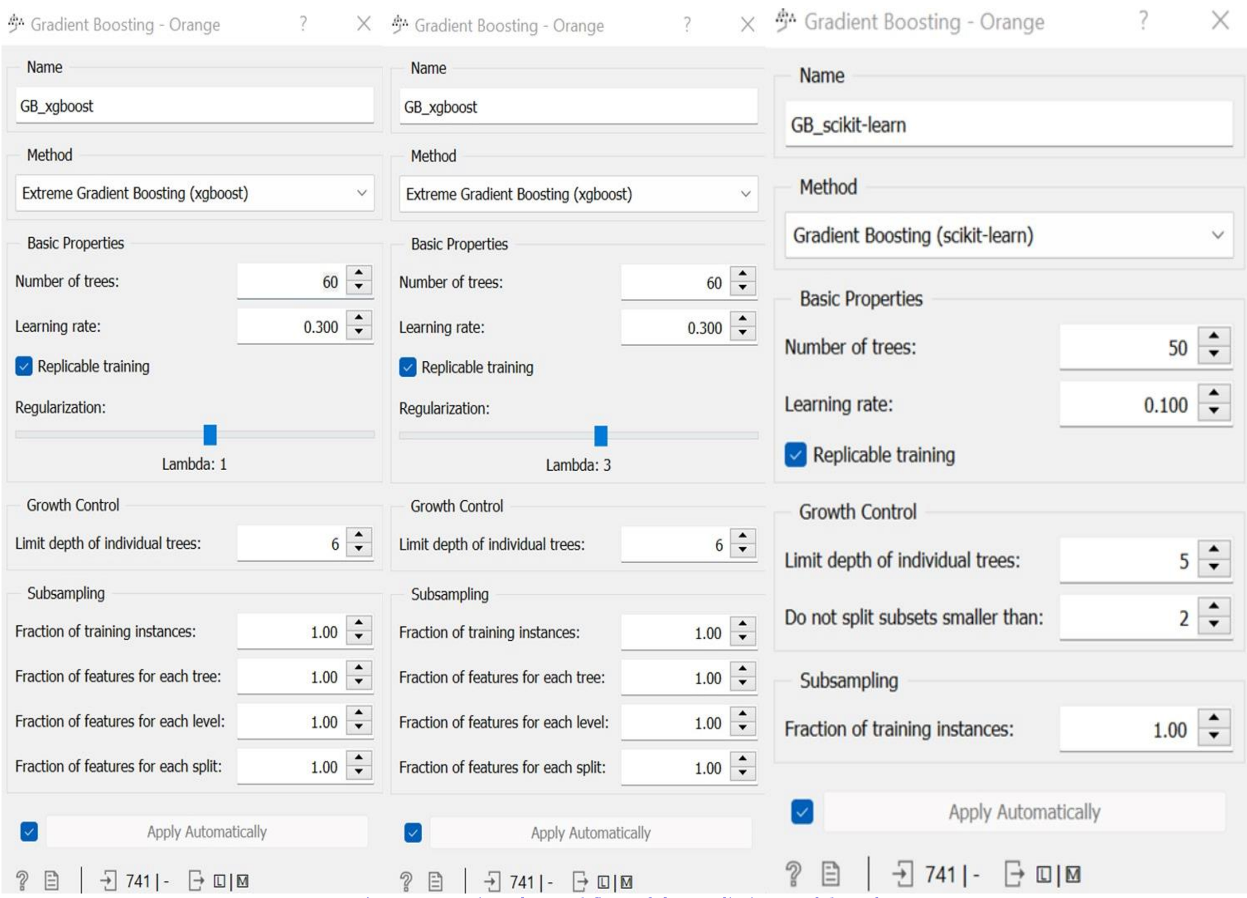

Figure 8. Depicts the workflow of the prediction model used.

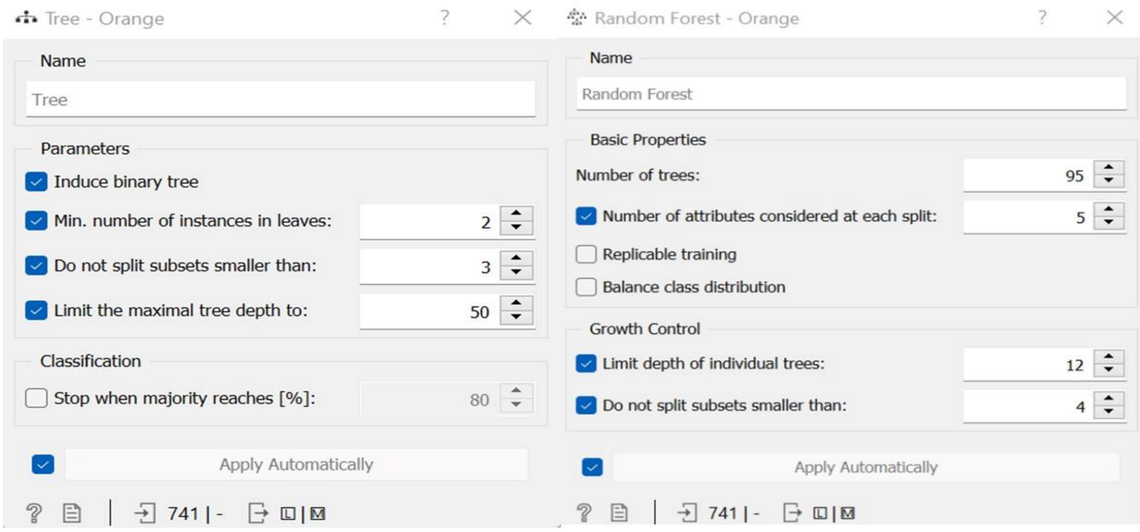

Figure 9. Depicts the workflow of the prediction model used.

4. SIMULATION RESULT ANALYSIS AND PERFORMANCE EVALUATION OF PREDICTIVE MODELS

In this paper, we implemented four different ML algorithms as KNN, Tree, RF, and GB with different sub-parameters. Table 1 depicts the results obtained after executing the different ML algorithms with tuned sub-parameters. The prediction results are analyzed by considering four parameters: MSE, RMSE, MAE, and R2. It seems that the Extreme GB (xgboost) with the Lambda (Regularization) value is one (1) that gives the best results with MSE is 0.001. If we increase the Lambda value from 1 to 3, the model's efficiency starts decreasing while the MSE increases from 0.001 to 0.002 and RMSE increases from 0.030 to 0.050. Therefore, the GB with the Scikit Learn method gives the MSE of 0.003 and RMSE of 0.051. Again, if we increase the Lambda value, which is the regularization value for GB with Catboost the method from 1 to 3, the error starts increasing; it shows the MSE increases from 0.017 to 0.036, while the RMSE from 0.129 to 0.189. In GB, if we kept the Lambda value the same but varied the value of depth of trees, the MSE, RMSE, MAE, and R2 vary. If the depth of trees is reduced from 100 to 50 the errors increase and it shows MSE 0.079, RMSE 0.280, MAE 0.204, and R2 0.999.

While the RF gives the RMSE of 0.088 and RMSE of 0.297. Followed by the KNN model with Chebyshev metric and uniform weight it gives the MSE of 1.082 and RMSE of 1.040 therefore the KNN with Euclidean metric and uniform weight gives the MSE of 0.687 and RMSE of 0.829.

Table 1. Prediction outcome after implementation of prediction models.

Model	Mean Square Error (MSE)	Root Mean Square Error (MSE)	Mean Absolute Error (MAE)	Coefficient of Determination (R2)
KNN_ Chebyshev	1.082	1.040	0.333	0.998
KNN_Euclidean	0.687	0.829	0.262	0.993
GB_Catboost (Lambda = 3)	0.036	0.189	0.135	1.000
GB_Catboost (Lambda = 1 and depth of tree is 100)	0.017	0.129	0.096	1.000
GB_Catboost (Lambda = 1 and depth of tree is 50)	0.079	0.280	0.204	0.999
GB_Scikit-learn	0.003	0.051	0.039	1.000
GB_Extreme Gradient Boosting (GB_xgboost) (Lambda = 3)	0.002	0.050	0.034	1.000
GB_xgboost (Lambda = 1)	0.001	0.030	0.021	1.000
Random Forest (RF)	0.088	0.297	0.060	0.999
Decision Tree (DT)	0.058	0.241	0.038	0.999

From Table 1 we can conclude that the GB gives us the best results with a minimum MSE of 0.001 while the KNN-Chebyshev gives the worst result with a maximum MSE of 1.082. Hence the KNN is more prone to errors while the GB is the least prone to errors followed by the Tree, and RF. In the case of RF as the number of trees increases the MSE and RMSE start decreasing. Therefore, we observed that the number of trees is inversely proportional to the RMSE and MSE. While in KNN with the Euclidean matrix as the number of the neighbors is increasing automatically RMSE start increasing.

5. CONCLUSION AND FUTURE SCOPE

The main goal of this paper is to predict the network traffic before it starts impacting the network performance. The authors of this paper proposed that by implementing the different ML algorithms such as GB, Tree, RF, and KNN with tuned sub-parameters, we are able to predict the network traffic. The simulation results show that the xgboost (Lambda (regularization) value one (1)) gives the best results with the MSE, RMSE of 0.001 and 0.030 respectively followed by the xgboost (Lambda (regularization) value three (3), MSE of 0.002 and RMSE of 0.050). The GB with the scikit learn method gives the MSE of 0.003 and RMSE of 0.051, therefore the Catboost ((Lambda = 1), MSE of 0.017, and RMSE of 0.129 but if decrease the depth of the tree the error start increasing therefore it shows MSE of 0.079, RMSE of 0.280), (Lambda = 3, gives the MSE of 0.036, and 0.189), while the DT gives the MSE of 0.058 and RMSE of 0.241 followed by the RF with MSE and RMSE of 0.088 and 0.297 respectively. Therefore, the KNN is more prone to errors and gives the worst result compared with those already implemented ML models. KNN with Chebyshev matrix gives the MSE of 1.082 and RMSE of 1.040, while KNN with Euclidean matrix gives the MSE of 0.687 and RMSE of 0.829.

So far, there has been little research on forecasting network traffic using data center data, despite the fact that collecting data from the DC is one of the most significant tasks. As a future scope, the research might be extended to estimate network traffic by using the data center data for various boosting models with customized parameters. It will give the new learner a way to analyze the efficiency of time-series algorithms to that of boosting and other machine learning techniques. As a result, the precision of the prediction can be assessed and evaluated. Till now not much research has been done in the field of networking to forecast the network traffic by using the data center data. This work has been carried out using the real-time data center data.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study.
Acknowledgement: PTC Software India Pvt. Ltd. contributed the data utilized in this paper.

REFERENCES

[1] A. Iranmanesh and H. R. Naji, "A protocol for cluster confirmations of SDN controllers against DDoS attacks," *Computers & Electrical Engineering*, vol. 93, p. 107265, 2021. Available at: https://doi.org/10.1016/j.compeleceng.2021.107265.

[2] A. Lazaris and V. K. Prasanna, "An LSTM framework for software-defined measurement," *IEEE Transactions on Network and Service Management*, vol. 18, pp. 855-869, 2020. Available at: https://doi.org/10.1109/tnsm.2020.3040157.

[3] A. R. Mohammed, S. A. Mohammed, and S. Shirmohammadi, "Machine learning and deep learning based traffic classification and prediction in software defined networking," in *2019 IEEE International Symp. Meas. Networking, M N - Proc*, 2019

[4] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (Datacenter) network," *Computer Communication Review*, vol. 45, pp. 123–137, 2015.

[5] D. H. Hagos, P. E. Engelstad, A. Yazidi, and O. Kure, "Recurrent neural network-based prediction of tcp transmission states from passive measurements," presented at the In 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), IEEE, 2018.

[6] E. H. Bouzidi, A. Outtagarts, R. Langar, and R. Boutaba, "Deep Q-Network and traffic prediction based routing optimization in software defined networks," *Journal of Network and Computer Applications*, vol. 192, p. 103181, 2021. Available at: https://doi.org/10.1016/j.jnca.2021.103181.

[7] F. Li, Z. Zhang, Y. Zhu, and J. Zhang, "Prediction of twitter traffic based on machine learning and data analytics," presented at the IEEE INFOCOM 2020 - IEEE Conf. Comput. Commun. Work. INFOCOM WKSHPS 2020, 2020.

[8] F. Tang, Z. M. Fadlullah, B. Mao, and N. Kato, "An intelligent traffic load prediction-based adaptive channel assignment algorithm in SDN-IoT: A deep learning approach," *IEEE Internet of Things Journal*, vol. 5, pp. 5141-5154, 2018. Available at: https://doi.org/10.1109/jiot.2018.2838574.

[9] F. Wang, D. Jiang, S. Qi, and C. Qiao, "An adaboost based link planning scheme in space-air-ground integrated networks," *Mobile Networks and Applications*, vol. 26, pp. 669-680, 2021. Available at: https://doi.org/10.1007/s11036-019-01422-4.

[10] J. d. M. Bezerra, A. J. Pinheiro, C. P. de Souza, and D. R. Campelo, "Performance evaluation of elephant flow predictors in data center networking," *Future Generation Computer Systems*, vol. 102, pp. 952-964, 2020. Available at: https://doi.org/10.1016/j.future.2019.09.031.

[11] M. Vejdannik and A. Sadr, "Machine learning-based QOT prediction for self-driven optical networks," *Neural Computing and Applications*, vol. 33, pp. 2919-2928, 2021. Available at: https://doi.org/10.1007/s00521-020-05123-y.

[12] M. F. Iqbal, M. Zahid, D. Habib, and L. K. John, "Efficient prediction of network traffic for real-time applications," *Journal of Computer Networks and Communications*, vol. 5, pp. 1-11, 2019. Available at: https://doi.org/10.1155/2019/4067135.

[13] M. Parashar, "Jaypee Institute of information technology university, University of Florida. College of Engineering, Institute of Electrical and Electronics Engineers. Delhi Section, and Institute of Electrical and Electronics Engineers," presented at the 2013 sixth International Conference on Contemporary Computing (IC3-2013): 8-10 August 2013, Jaypee Institute of Information Technology, Noida, India, 2013.

[14] P. Boryło, G. Rzym, and P. Chołda, "A time-efficient shrinkage algorithm for the Fourier-based prediction enabling proactive optimisation in software-defined networks," *International Journal of Communication Systems*, vol. 33, p. e4448, 2020. Available at: https://doi.org/10.1002/dac.4448.

[15] R. HariKrishnan, S. Mahajan, and K. Kotecha, "Prediction of network traffic in wireless mesh networks using hybrid deep learning model," *IEEE Access*, vol. 10, pp. 7003-7015, 2022. Available at: https://doi.org/10.1109/access.2022.3140646.

[16] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Expert Systems with Applications*, vol. 176, p. 114885, 2021. Available at: https://doi.org/10.1016/j.eswa.2021.114885.

[17] S. Qi, D. Jiang, and L. Huo, "A prediction approach to end-to-end traffic in space information networks," *Mobile Networks and Applications*, vol. 26, pp. 726-735, 2021. Available at: https://doi.org/10.1007/s11036-019-01424-2.

[18] S. Sanagavarapu and S. Sridhar, "SDPredictNet-A topology based SDN neural routing framework with traffic prediction analysis," presented at the in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference, CCWC 2021, 2021.

[19] L. T. Mai and N. Navet, "Improvements to deep-learning-based feasibility prediction of switched ethernet network configurations," presented at the In 29th International Conference on Real-Time Networks and Systems, 2021.

[20] U. Paul, J. Liu, S. Troia, O. Falowo, and G. Maier, "Traffic-profile and machine learning based regional data center design and operation for 5G network," *Journal of Communications and Networks*, vol. 21, pp. 569-583, 2019. Available at: https://doi.org/10.1109/jcn.2019.000055.

[21] Y. V. Sneha, V. Vishwasini, S. Boloor, and N. D. Adesh, "Prediction of network congestion at router using machine learning technique," in *in 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics, DISCOVER 2020 - Proceedings*, 2020, pp. 188–193.

[22] Y. Wang, D. Jiang, L. Huo, and Y. Zhao, "A new traffic prediction algorithm to software defined networking," *Mobile Networks and Applications*, vol. 26, pp. 716-725, 2021. Available at: https://doi.org/10.1007/s11036-019-01423-3.

[23] Y. Yang, "A new network traffic prediction approach in software defined networks," *Mobile Networks and Applications*, vol. 26, pp. 681-690, 2021. Available at: https://doi.org/10.1007/s11036-019-01413-5.

[24] Z. M. Nia and M. R. Khayyambashi, "Improving content popularity prediction with k-means clustering and deep-belief networks," *Multimedia Tools and Applications*, vol. 80, pp. 15745-15764, 2021. Available at: https://doi.org/10.1007/s11042-020-10463-x.