A Comparative Study of Artificial Intelligence Algorithms for Network Traffic Prediction in VANET

1st Sanaz Shaker Sepasgozar
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer and Software Engineering
Polytechnique Montreal
Montreal, Canada
sanaz.shaker-sepasgozar@polymtl.ca

2nd Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM)
Department of Computer and Software Engineering
Polytechnique Montreal
Montreal, Canada
samuel.pierre@polymtl.ca

***Abstract*—Increasing the number of vehicles and their communications in smart cities is a critical issue that leads to road and network traffic. Traffic prediction with high accuracy and less complexity is a challenge in Intelligent Transportation System (ITS). In Artificial Intelligence (AI), Machine learning (ML) algorithms are promising solutions to prediction problems, and Deep Learning (DL) algorithms are used for more complicated issues. In this paper, we propose a comparative analysis of the prediction performance of the most five common AI algorithms used to solve classification problems. Different evaluation metrics are employed to analyze algorithms and get the most accurate one for selecting such problems. Simulation results on the Vehicular Ad-Hoc Network (VANET) dataset revealed that Random Forest (RF) as a traditional ML algorithm performed better than other algorithms in terms of accuracy (96%) and execution time (0.73 minute) for traffic prediction.**

***Keywords*—Vehicular Ad-Hoc Network, network traffic prediction, Artificial Intelligence, machine learning method, neural network, Intelligent Transportation System**

I. Introduction

One of the most common technologies deployed in smart cities is ITS, which VANET is a crucial part. This latter consists of wireless communication between Vehicle-to-Road Side Unit (V2R) and Vehicle-to-Vehicle (V2V). Different types of AI algorithms are applied for discovering the pattern and extracting valuable knowledge for prediction [1]. Several studies address network traffic prediction but finding the best algorithm with the highest prediction accuracy and minimum error is still a challenge. Consequently, this challenge can affect the network performance in terms of delay and accuracy. Regarding the importance of accuracy of prediction and execution time based on increasing the volume of data, we need to implement the most precise algorithm considering its execution time.

In this paper, we implement four different machine learning algorithms including RF [2], K-Nearest Neighbors (KNN) [3], Support Vector Machine (SVM) [4] and Naive Bayes (NB) [5]. Moreover, we try a Multi-Layer Perceptron (MLP) [6], which is a class of feed-forward Artificial Neural Network (ANN). MLP algorithm is used for solving classification problems, and it relies on a neural network. It uses the back-propagation technique for training, which can produce a single output from a set of inputs.

Moreover, we use a real dataset collected from Global Positioning System (GPS) and is taken from V2R communication based on packets that are sent by vehicles to the Road Side Units (RSUs). Our main target is finding the best classification algorithm to predict the network traffic based on the ”packet receiving” parameter. We define traffic situations as when the packet has not been received and in non-traffic situations, the vehicles will receive the packet. We compare the prediction results of the five above listed AI algorithms for classification problems. The main contribution of this paper is to show that RF is the best AI algorithm in terms of accuracy, execution time and error.

The rest of this paper is organized as follows. Section II presents related work on different machine learning and deep learning algorithms for traffic prediction. Section III explains the proposed method for implementing KNN, RF, SVM, NB and MLP algorithms. The evaluation results are illustrated in Section IV and the conclusion is drawn in Section V.

II. RELATED WORK

With the growth of the number of vehicles and their communications in smart cities, network traffic prediction becomes a major issue. A wide range of related work has focused on deploying various AI algorithms, in order to propose an accurate and accelerated traffic prediction model. A large number of these works adapted ML algorithms [7], [8], plenty of them presented DL algorithms [9], [10] and some works combined DL and ML algorithms together to propose a prediction model [11], [12].

A novel prediction model was proposed in [13] with the aim of predicting network traffic accurately. They combined three different algorithms: (1) Local Mean Decomposition (LMD) for decomposing the network traffic time series to get components of product function, (2) Bidirectional Long Short-Term Memory (BiLSTM) algorithm for predicting the network traffic data and (3) the Bayesian optimization algorithm to optimize the hyperparameters in BiLSTM. They

978-1-6654-6975-3/22/$31.00 ©2022 IEEE

implemented their proposed model on two actual datasets and used some evaluation metrics including Root Mean Square Error (RMSE), Relative Root Mean Square Error (RRMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Square Sum Error (SSE), Theil Inequality Coefficient (TIC), Index of Agreement (IA), and Coefficient of Determination (R2). The prediction results showed that their model outperformed Auto Regressive Integrated Moving Average (ARIMA), Empirical Mode Decomposition (EMD), BiLSTM, SVM and MLP algorithms.

Lim *et al.* [14] performed network traffic classification by implementing Convolutional Neural Network (CNN) and Residual Network (ResNet) for classifying network traffic. Their main objective was to fit unseen packets by enabling a deep learning model. They implemented the proposed model on the generated packet-based dataset and analyzed the performance of the network traffic classification model using the $F1-score$.

A framework on a Long Short Term (LSTM) neural network algorithm to perform the network traffic prediction at a short time scale (i.e., <30 seconds) was used in [15]. The authors compared their results with RNN and ARIMA algorithms by using Normalized Root Mean Square Error (NRMSE) as an evaluation metric. Therefore, the results showed that the proposed model performed better than compared algorithms.

An improved Support Vector Machine algorithm called Cost-Sensitive SVM (CMSVM) has been proposed in [16]. This proposed algorithm is based on multi-class SVM algorithm. The authors used active learning for every single class of traffic and considered weight values, in order to solve the imbalance problem in network traffic recognition. They did the implementation by using two imbalanced datasets. The experimental results showed that the proposed CMSVM algorithm was able to overcome the imbalance problem, decrease the computation cost and enhance the accuracy of classification.

A method for predicting the Long-Term Evaluation (LTE) network edge traffic was proposed in [17]. They implemented three algorithms including the SVM, RF and Bagging on the public cellular traffic dataset. They used the RMSE, MAE and R2 as evaluation metrics. Based on these metrics, the Bagging algorithm has the lowest error, but its training time is 116s, while the SVM takes 6s for training. Therefore, SVM is significantly better than the Bagging and the RF which have both the worst error and training time. Thus, the algorithm that has the best performance in case of error evaluation is the most training time-consuming.

In [18], the authors attempted to solve the network traffic prediction problem by implementing various Recurrent Neural Network (RNN) architectures like simple RNN, LSTM and Gated Recurrent Unit (GRU) algorithms. They focused on three problems including packet distribution prediction, volume prediction and packet protocol prediction. To evaluate the performance, they used the Mean Squared Error (MSE), metric and compared the various architectures of the RNN algorithm with ARIMA, Naive Method (NM) and Moving Average (MA). The experiment results showed that the simple RNN, LSTM and GRU outperformed other algorithms in terms of accuracy error evaluation.

III. PROPOSED MODEL FOR NETWORK TRAFFIC PREDICTION

In this section, we present a comparison between various traditional ML algorithms (i.e., KNN, RF, SVM and NB) and one of the most common AI algorithms based on the neural network called MLP for network traffic prediction. Fig. 1 depicts the architecture of our proposed prediction model to compare the five selected algorithms.

A. *Dataset*

We used a real dataset based on VANET communications [19], where the type of network is 802.11 Ad-Hoc. It contains two kinds of communication data based on V2V and V2R for measuring short-range communications. The authors installed a GPS antenna on each vehicle and collected the longitude, latitude, speed and heading every two seconds. In our comparison work, we focused on V2R communication where 1470 bytes of packets had been broadcasted via the senders at an average rate of around 150 packets/s.

In the preprocessing step, we did data cleaning and scaling using the MinMaxScaler method for normalizing and scaling the data.

Moreover, for network traffic prediction based on V2R communication, we labeled data into class 0 for not receiving packets in a traffic situation and class 1 for receiving packets in a non-traffic situation. Then, we split data into 75% for training and 25% for tests. We implemented five different AI algorithms (i.e., KNN, RF, SVM, NB and MLP). Finally, we evaluated their performance based on various evaluation metrics, in order to find the most suitable algorithm for network traffic prediction in VANET.

B. *Implementation*

To implement the KNN, RF, SVM, NB and MLP algorithms, we used Python version 3.7 [20] and some libraries including Sklearn, Pandas, NumPy, Matplotlib and Mlxtend. For implementing the MLP algorithm, Adam was utilized as an optimizer and we chose ReLU as an activation function including 300 iterations. All the implementations took place in Google colab using GPU accelerator to have better performance and improve the processing time.

C. *Selected AI Algorithms*

This section presents the five selected AI algorithms for network traffic prediction: KNN, RF, SVM, NB, and MLP.

- KNN: one of the straightest algorithms in ML is the nearest neighbor used for both regression and classification problems. The KNN classification has two stages: (1) searching the nearest neighbors and (2) defining the classes that used those neighbors. Selecting the k nearest neighbor is done by distance metric [3]. Although the KNN is simple and easy to implement, finding sample take a long time to perform.

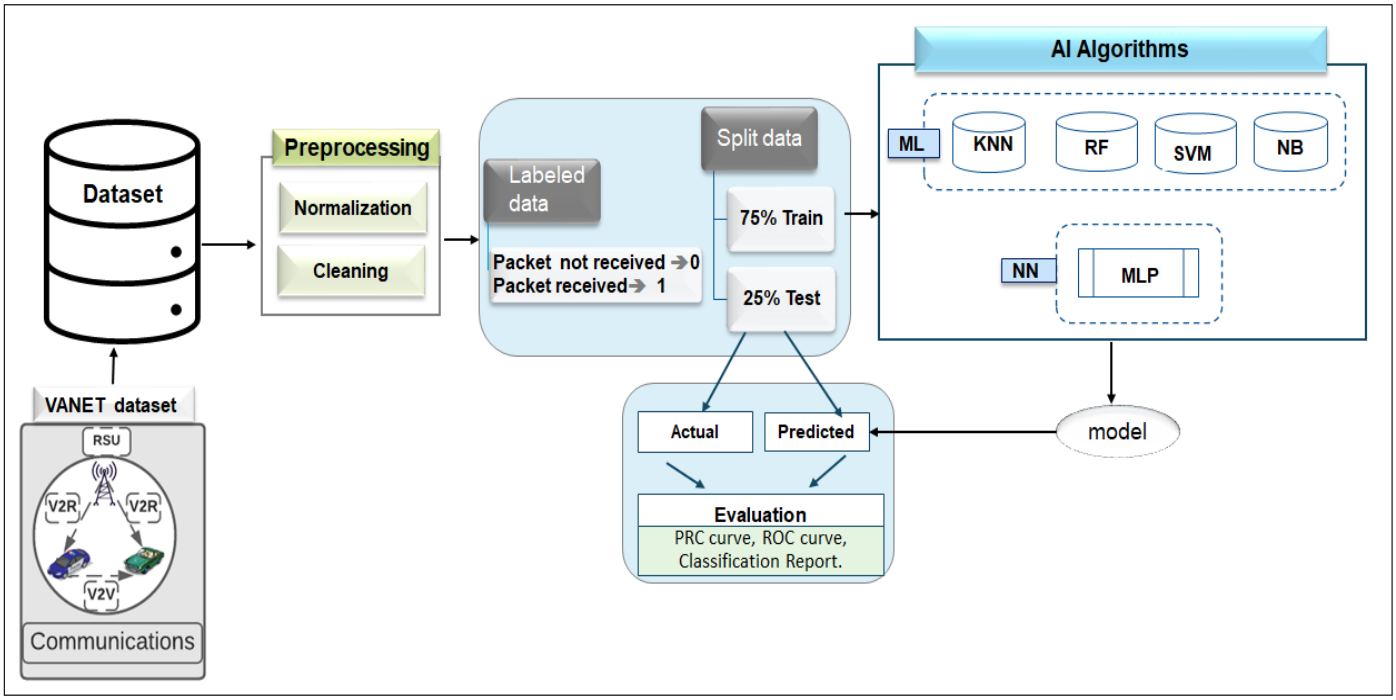

Fig. 1. The architecture of proposed comparison model.

- RF: includes a set of decision trees selected randomly from the training set, then for finding the final class, the votes will aggregate to make the final decision. The RF can decrease the overfitting in decision tree algorithms and enhance the accuracy [2].
- SVM: can be used for both binary and multiclass classifications or regression problems with high accuracy. It is based on separating data by a hyperplane in an optimal way to make classes by considering training samples, which exist at the edge of each class [4].
- NB: is a powerful algorithm for binary and multiclass classification problems and works based on Bayes theory that connects the different conditional probabilities of random events. The NB is a scalable and fast algorithm but it cannot learn the feature's relation.
- MLP: is an appropriate prediction algorithm in labeled data and regression problems. The MLP is a feed forward class of ANN algorithm that adjusts suitable connection between input and output data. It consists of an input layer, a hidden layer including a certain number of neurons called perceptron and an output layer, where each of these layers has a full connection with the weighted connection to the units in the previous layer and the output will go through a nonlinear activation function designated by (1) [6]:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(W^T X + b) \quad (1)$$

where W denotes the weights, X represents the vector of inputs, b is the bias and φ is the activation function. The training part of the MLP tries to adapt the weights and biases to their ideal value for minimizing the error, which is denoted by [6]:

$$\mathrm{E} = \frac{1}{l} \sum_{i=1}^{l} (T_i - Y_i)^2 \quad (2)$$

where T_i denotes the predicted value, Y_i is the actual value and l is the training set size. The architecture of MLP network is illustrated in Fig. 2 [6].

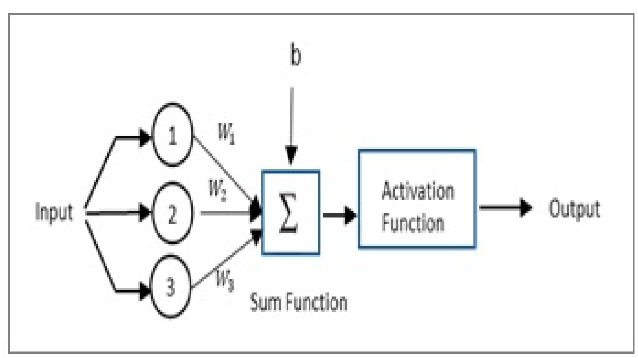

Fig. 2. The architecture of MLP network.

The MLP algorithm is an adaptive neural network algorithm that can learn to do the different tasks based on the training dataset. Moreover, it is a fully connected algorithm, which has too many parameters that might cause inefficiency.

In this paper, we implemented four traditional ML algorithms and one simple neural network algorithm for network traffic prediction to solve classification problems. We used some evaluation metrics, in order to evaluate the performance of these algorithms. The RF algorithm gives the best result in network traffic prediction in terms of accuracy and execution time.

After implementing these five algorithms considering traffic happening in the network based on packet receiving, we need to evaluate them using various evaluation metrics.

D. Evaluation Metrics

Intending to compare the performance of different algorithms, we used various evaluation metrics (e.g., Receiver Operating Curve (ROC) curve, Precision-Recall (PR) curve, calculating the execution time, etc.) [21]. Moreover, we tried four other classification metrics including $Precision$, $Recall$, $F1-score$ and $Accuracy$ which are calculated, respectively, by (3), (4), (5) and (6) [22].

$$Precision = \frac{T_P}{T_P + F_P} \tag{3}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4}$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Accuracy = \frac{T_P + T_N}{(T_P + T_N + F_P + F_N)} \times 100, \tag{6}$$

where T_p is True Positive (when actual = 1, predicted = 1), F_P is False Positive (when actual = 0, predicted = 1), T_N is True Negative (when actual = 0, predicted = 0) and F_N is False Negative (when actual =1, predicted = 0). Indeed, comparing various algorithms for network traffic prediction can get us more insight about the performance of them in classification problems.

Finally, we compared the execution time that each algorithm takes to perform the network traffic prediction. Based on all these above evaluation metrics, we can decide which algorithm has better performance in network traffic prediction.

IV. SIMULATION RESULTS

In the VANET network, vehicles have different wireless communications like V2V and V2R, where we consider the *sending packets* as the parameter for these communications. We assume traffic and non-traffic situations in the network based on packet receiving and compare the results of the five selected algorithms (i.e., KNN, RF, SVM, NB and MLP) for predicting it in V2R communication. Then, by investigating the performance of each algorithm, we find the most performed algorithm for network traffic prediction in classification problems.

To measure the performance of algorithms in classification problems, we used one of the most prevalent graphical tools called the Receiver Operating Characteristics (ROC) curve. Then, we calculated the Area Under the Curve (AUC), where the large size of this area is a good indication of the performance of each algorithm. Different probabilities of prediction are shown by the ROC curve, where the False Positive Rate (FPR) indicates on the X-axis and the True Positive Rate (TPR) on the Y-axis. The nearest curve to the top left corner, related to the biggest AUC, has the best performance in prediction. Fig. 3 depicts the performance of the five different algorithms based on the ROC curve and the calculated AUC.

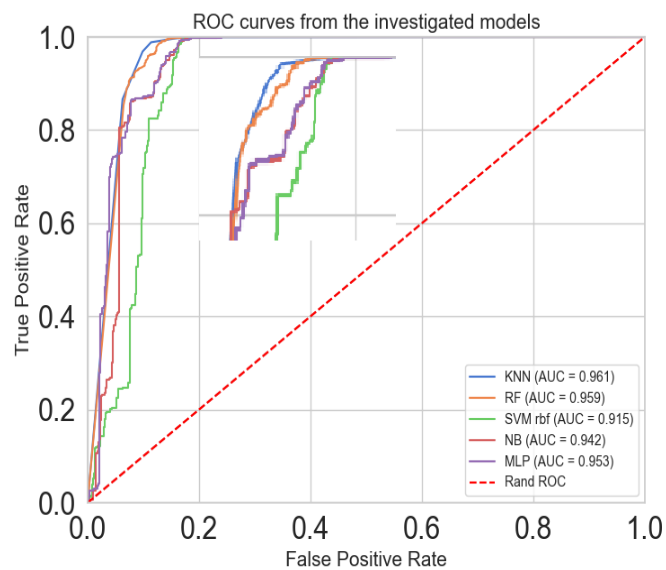

Fig. 3. ROC curve for KNN, RF, SVM, NB and MLP.

The red dotted line, which separates the whole area into two parts, is a classifier that acts on a random level. The curves located above it have better performance than the ones that are placed under it. As shown in Fig. 3, the blue and orange curves that belongs to KNN and RF, respectively, are the nearest curves to the top left corner and have the biggest AUC that shows their satisfying performance. The SVM, which is represented by the green curve, performs worse than all other algorithms. The NB and MLP, which are illustrated by the red and purple curves, respectively, have almost the same performance, specifically when TPR is ranging from 0.8 to 1.0.

There are other metrics like $accuracy$, $precision$, $recall$ and $F1-score$ that can get more insight into the performance of the different algorithms for network traffic prediction and we calculated them, as illustrated in Table I.

The KNN with 97% has the highest accuracy, the RF and MLP have the same accuracy with 96%, and the SVM and NB with 95% and 89% have the lowest accuracy accordingly.

Furthermore, we used another evaluation metric name precision-recall (PR) curve to show the balance between

TABLE I
CLASSIFICATION REPORTS

Algorithm	Accuracy (%)	Label (packet receiving)	Precision	Recall	F1-score
KNN	97	No	0.95	0.89	0.92
		Yes	0.97	0.99	0.98
RF	96	No	0.99	0.81	0.89
		Yes	0.95	0.99	0.97
MLP	96	No	0.99	0.81	0.89
		Yes	0.95	0.99	0.97
SVM	95	No	0.87	0.88	0.87
		Yes	0.97	0.96	0.97
NB	89	No	0.68	0.88	0.77
		Yes	0.97	0.89	0.93

precision and recall, which are pointed out at the Y-axis and X-axis, respectively. We compared the PR curves of the five selected algorithms, as illustrated in Fig. 4.

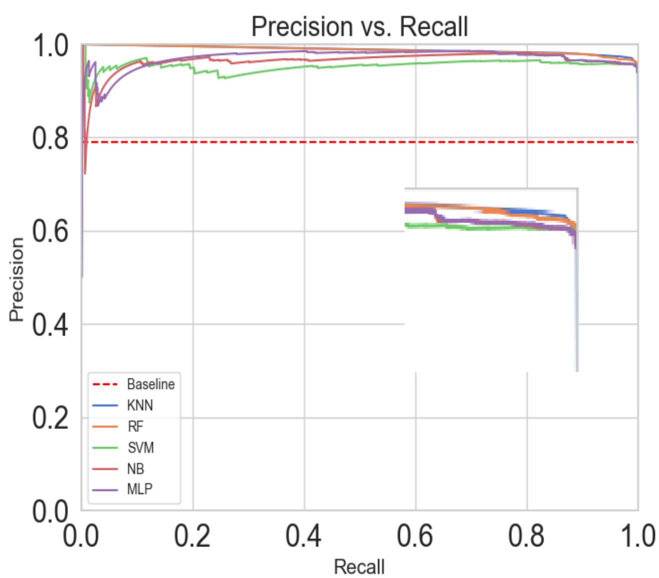

Fig. 4. PR curve for KNN, RF, SVM, NB and MLP.

The red dotted line, which is displayed horizontally, is a baseline classifier for more insight into the performance of algorithms in a way that the algorithms placed above this line and near the right top corner have better performance.

The comparison results present that the KNN (i.e., the blue curve) and RF (i.e., the orange curve) algorithms have better performance. The NB algorithm, which is related to the red curve that even crosses the baseline, has the worst performance. The MLP algorithm (i.e., the purple curve) acts better than the NB and the SVM (i.e., the green curve) algorithms.

We tried different evaluation metrics for network traffic prediction in classification problems. However, there is another important parameter to find out the most performed algorithm which is the execution time. Based on the growth of the volume of data in wireless communications, the time that each algorithm takes to be executed could be a challenge. Though, we calculated the execution time for each of the five selected algorithms, as shown in Fig. 5.

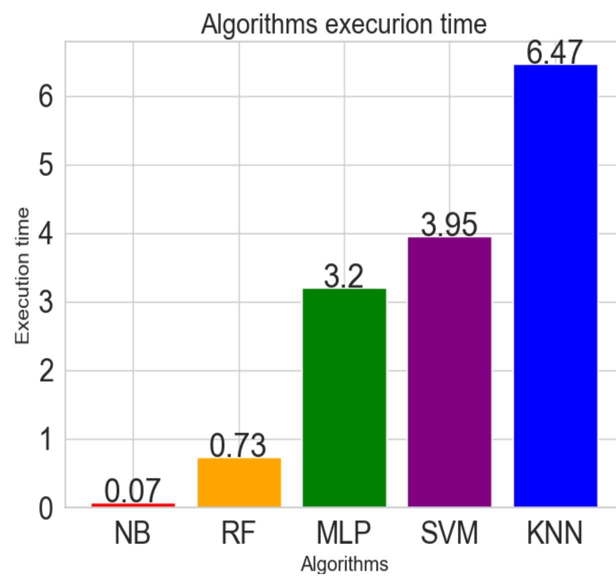

Fig. 5. Execution time for network traffic prediction using KNN, RF, SVM, NB and MLP.

The KNN algorithm, which has the best performance in other evaluation metrics, is considered the most time-consuming algorithm for network traffic prediction. As depicted in Fig. 5, the KNN takes 6.47 minutes, which is the highest time compared with other algorithms. Therefore, it could not be the most performed algorithm for our prediction. The SVM and MLP take 3.95 and 3.2 minutes, respectively. The NB is the fastest algorithm with 0.07 minute, but in other evaluation metrics, it has poor results. Finally, the RF has satisfying results as the KNN in other evaluation metrics. It takes 0.73 minute as execution time, which is an acceptable result for prediction.

Therefore, based on all the results obtained from different evaluation metrics to compare the performance of different traditional machine learning algorithms and one simple neural network algorithm for network traffic prediction, the RF is the most performed algorithm in terms of accuracy, execution time and prediction error.

V. CONCLUSION

In this paper, we investigated the performance of four traditional machine learning algorithms (i.e., KNN, RF, SVM and NB) and tried MLP as a neural network algorithm for network traffic prediction in VANET. We used a real dataset based on V2R communication and we implemented these five algorithms to find the most performed one for classification problems. We considered packet receiving as a network traffic parameter and we assumed in a non-receiving packet situation that traffic is happening, and when the packet is received, there is no traffic in the network. We analyzed the performance of various evaluation metrics, such as ROC curve, PR curve, classification report and execution time. The evaluation results show that the KNN with a slight difference from the RF, performs better than other algorithms, but its execution time for network traffic prediction remarkably takes higher time (6.47 minutes). It is enough that we do not consider the KNN as the best algorithm. The fastest algorithm is NB with 0.07 minute of execution time, but it performs weakly in other evaluation metrics. The MLP algorithm gives better results in evaluation metrics than SVM, even in execution time they take 3.2 minutes and 3.95 minutes, respectively. Based on all evaluation metrics, we can conclude that the RF is the most performed algorithm for network traffic prediction in VANET with a good execution time (0.73 minute) and reasonable results in other evaluation metrics.

REFERENCES

[1] L. Yong-Ju and M. Okgee, "Comparative analysis of machine learning algorithms to urban traffic prediction," *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Burlingame, Jeju, Korea (South), Oct. 2017, pp. 1034–1036.

[2] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, pp. 217–222, Jan. 2005.

[3] P. Cunningham, S.J. Delany, "k-Nearest neighbour classifiers-A Tutorial," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–25, Jul. 2021.

[4] G.M. Foody, A. Mathur, "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification," *Remote Sensing of Environment*, vol. 93, pp. 107–117, Oct. 2004.

[5] J. Ren, S.D. Lee, X. Chen, B. Kao, R. Cheng, D. Cheung, "Naive bayes classification of uncertain data," *2009 Ninth IEEE international conference on data mining*,Miami Beach, FL, USA, Dec. 2009, pp. 944–949.

[6] A.Y. Nikravesh, S.A. Ajila, C.H. Lung, W. Ding, "Mobile network traffic prediction using MLP, MLPWD, and SVM," *2016 IEEE International Congress on Big Data (BigData Congress)*,San Francisco, CA, USA, Oct. 2016, pp. 402–409.

[7] D. Xu, Y. Wang, P. Peng, S. Beilun, Z. Deng, H. Guo, "Real-time road traffic state prediction based on kernel-KNN," *Transportmetrica A: Transport Science*, vol. 16, pp. 104–118, Oct. 2020.

[8] W. Chen, Z. Shang, Y. Chen, "A novel hybrid network traffic prediction approach based on support vector machines," *Journal of Computer Networks and Communications*, vol. 2019, Feb. 2019.

[9] R. Vinayakumar; K.P. Soman; P. Poornachandran, "Applying convolutional neural network for network intrusion detection," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*,Udupi, India, Dec. 2017, pp. 1222–1228.

[10] L. Nie, X. Wang, L. Wan, S. Yu, H. Song, D. Jiang, "Network traffic prediction based on deep belief network and spatiotemporal compressive sensing in wireless mesh backbone networks," *Wireless Communications and Mobile Computing*, vol. 2018, Jan. 2018.

[11] S.S. Sepasgozar, S. Pierre, "An Intelligent Network Traffic Prediction Model Considering Road Traffic Parameters Using Artificial Intelligence Methods in VANET," *IEEE Access*, vol. 10, Jan. 2022, pp. 8227–8242.

[12] A.R. Mohammed, S.A. Mohammed,S. Shirmohammadi, "Machine learning and deep learning based traffic classification and prediction in software defined networking," *2019 IEEE International Symposium on Measurements & Networking (M&N)*,Catania, Italy, Jul. 2019, pp. 1–6.

[13] Z. Tian, P. Song, "A novel network traffic combination prediction model," *International Journal of Communication Systems*, vol. 35, May. 2022, pp. e5097.

[14] H.K. Lim, J.B. Kim, J.S. Heo, K. Kim, Y.G. Hong, Y.H. Han, "Packet-based network traffic classification using deep learning," *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*,Okinawa, Japan, Feb. 2019, pp. 046–051.

[15] S. Nihale, S. Sharma, L. Parashar, U. Singh, "Network traffic prediction using long short-term memory," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*,Coimbatore, India, Jul. 2020, pp. 338–343.

[16] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Expert Systems with Applications*, vol. 176, Aug. 2021, pp. 114885.

[17] N. Stepanov, D. Alekseeva, A. Ometov, E.S. Lohan, "Applying Machine Learning to LTE Traffic Prediction: Comparison of Bagging, Random Forest, and SVM," *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*,Brno, Czech Republic, Oct. 2020, pp. 119–123.

[18] N. Ramakrishnan, T. Soni, "Network traffic prediction using recurrent neural networks" *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*,Orlando, FL, USA, Dec. 2018, pp. 187–193.

[19] R.M. Fujimoto, R. Guensler, M.P. Hunter, H. Wu, M. Palekar, J. Lee, and J. Ko, "CRAWDAD dataset gatech/vehicular (v.2006-03-15)."https://crawdad.org/gatech/vehicular/20060315, https://doi.org/10.15783/C74S3Z / (accessed on Mar. 2006).

[20] G. Van Rossum and F.L. Drake "Python 3 Reference Manual CreateSpace," *Scotts Valley, CA*, 2009.

[21] H. Wu, M. Palekar, R. Fujimoto, R. Guensler, M. Hunter, J. Lee, and J. Ko, "An empirical study of short range communications for vehicles," *Proceedings of the 2nd ACM international workshop on Vehicular ad hoc networks*, pp. 83–84, Sep. 2005.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas- sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit- learn: Machine learning in Python," *Machine learning in Python," Journal of Machine Learning Re- search*, vol. 12, pp. 2825–2830, Nov. 2011.