 https://doi.org/10.24191/jeesr.v21i1.013

Machine Learning Model for Performance Prediction in Mobile Network Management

Hazim Wahid, Nur Idora Abdul Razak* and Syahrul Afzal Che Abdullah

***Abstract*—Nowadays, mobile network management involves human interaction and human work ranging from conducting drive tests that can evaluate network performance and coverage to diagnosing customer complaints. Predictive network analytics related to network management involves predicting network performance at locations or times in which no direct measurement data is available. One of the major challenges when applying machine learning is to choose the best model from a variety of models to solve a problem. Therefore, the research gap is proposing the best machine learning model for predicting mobile network performance in the identified dimensions. The methodology includes drive test measurement for data collection, exploratory data analysis, data preparation, and applying machine learning models in predicting mobile network performance. Whereas throughput has a strong correlation with signal strength, throughput is the targeted parameter in network performance prediction. Three machine learning models were applied in this study which are Random Forest, Gaussian Process Regression, and K-Nearest Neighbor for throughput prediction. Based on the results and analysis of the evaluation metric comparison, it shows that the Random Forest model comes with the highest performance prediction with the R^2 score of 0.79 followed by KNN 0.66 and lastly Gaussian Process Regression 0.34. Random forest achieved the best result because of an additional layer on randomness that can lessen the variance thus increasing the model accuracy. Using the hyperparameter tuning the number of trees and the value for the depth of each tree in the forest will increase random forest model accuracy. Based on the important features, the location of the measurement and SNR value is important feature in affecting network performance. Thus, network operators need to improve the network coverage in a certain area to give a better experience to the user.**

***Index Terms*—Machine Learning, Mobile Network, LTE, Performance Prediction**

This manuscript is submitted on 24th February 2022 and accepted on 24th February 2022. This paper was part of works conducted under the Geran Penyelidikan Khas (600-RMC/GPK 5/3 (209/2020)). The authors would also like to acknowledge all support given by the Universiti Teknologi MARA through the grant.

Hazim Wahid is a post-graduate student from the College of Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor (email: 2021186713@isiswa.uitm.edu.my). Nur Idora Abdul Razak and Syahrul Afzal Che Abdullah are lecturers at the College of Engineering, Universiti Teknologi MARA, Selangor. (e-mail: nuridora@uitm.edu.my, bekabox181343@uitm.edu.my

*Corresponding author
Email address: nuridora@uitm.edu.my

I. INTRODUCTION

THERE is a growing body of literature that recognizes the importance of conducting drive tests. Mobile network management and drive tests provide services that give a clear insight into the quality of mobile network coverage and other wireless networks including identifying areas of poor signal quality and identification of black spots [1]. Conducting drive test also provide service to the user with information on the mobile network compliance with relevant technical requirements and regulations.

The mainstream key performance parameters when conducting LTE drive tests are Signal to Noise Ratio (SNR), Reference Signal Receive Power (RSRP), Reference Signal Receive Quality (RSRQ), Reference Signal Strength Indicator (RSSI), and throughput. Among these parameters, throughput can determine signal strength because of the strong correlation with signal strength parameters. It is also a dependent variable with other parameters such as geographical location. Mobile network management needs human interaction from conducting drive tests to diagnosing and analyzing customer complaints to meet user's network Quality of Service (QoS). This approach is also time-consuming when processing large and complex data to find the hidden fluctuation pattern in the network connectivity and to give valuable insights into the dataset [2]. The amount of data needed including cell neighboring information is very high in the process [3]. Some of the places are not being covered during the drive test due to the location-based query was not supported by the radio frequency tools and the capability of the drive test in which there are no direct measurement is available [1]. Also, in the measurement log, there will be a missing value due to the drive test tools' incapability.

Nowadays Predictive Network Analytics (PNA) has become a mainstream network management tool and diagnostic when processing a large volume of data [4]. PNA also plays an important role in network performance monitoring. Machine learning in predictive analytics used historical data to predict future events and the identifiable patterns are then used to facilitate network changes to solve network performance issues. Machine learning is also capable to help the network operations team to identify the performance issues and potential network failures with better accuracy and lower mean time to repair [5]. Machine learning in network performance prediction can improve network analytics by reducing human interaction and predicting value where there is no direct data measurement

available and the missing data in the measurement. The main challenge is to identify the best machine learning model for predicting network performance [6]. Because different datasets used for prediction will yield a different machine learning model performance.

The key contribution of this study is to propose the best machine learning model for predicting mobile network performance in the identified dimensions. This can be done by benchmarking various types of machine learning models for downlink throughput prediction. The evaluation metric will be used to evaluate the machine learning model performance and then the comparative analysis were made. By using the random forest model specialty that is capable to identify the features importance list that help to determine which feature is affecting the downlink throughput prediction the most compared to the other network parameter. The hyperparameter tuning method is carried out to enhance the best model performance and accuracy for throughput prediction.

II. Related Research

There is a large volume of published studies describing the role of artificial intelligence in the field of mobile networks. One study by [3] examined the usage of Deep Neural Network (DNN) in the mobile network to reduce the number required to conduct drive tests for LTE network management. Due to the drive test being time-consuming for a dense area that has many obstacles will cause non-line of sight on the measurement field. By utilizing deep learning techniques, a neural network is able to predict signal quality at unseen locations by using satellite information and position-based measurements. The results show that the deep learning model could predict LTE signal quality parameters SNR, RSRQ, and RSRP. The author also highlighted that in predicting signal strength on the mobile network, the general problem in the field of machine learning is to obtain a model that is capable of performing well on unseen data. The model might be good at predicting in the selected dataset but when the dataset is changing, the performance of the model might drop off due to the unseen data and the hidden pattern in the dataset. They also point out that to reduce the final error the training set should be unlimited by extending the geographical region with an additional propagation scenario. Therefore, in order to increase the machine learning model accuracy, it requires a large volume of training datasets for the model the learn the data behavior.

A follow-up study [7] found that since throughput is a main indicator of the network performance, data from several studies have listed that the parameter used to predict throughput is Signal-to-Noise Ratio (SNR), Reference Signal Receive Quality (RSRQ), Reference Signal Receive Power (RSRP) and Received Signal Strength Indicator (RSSI). This view is supported by [8] who stated that throughput parameters depend on various features which are context information and network quality parameters. The writer uses two approaches that are well established used for prediction. The first approach is the classical regression machine learning predictive techniques. The model is K-Nearest Neighbor (KNN), Support Vector Machine Regression (SVR), Ridge Regression, and Random Forest. The second approach in their study is the time series forecasting which is the ARIMA model and Long Short-Term Memory (LSTM). The result shows that the random forest model performs the best among all the models with the highest score value for the coefficient of determination. By referring to the random forest theories, random forest performs the best because the model adds a layer of randomness to the features. This differs from another study [9] that performs comparative analysis on three commonly machine learning predictive models that are Support Vector Machine, Deep Neural Networks, and Extreme Gradient Boosting Trees. Their results show that XGBoost achieved the highest accuracy with 90%. The explanation behind this is the model easy to use, easy to parallelize, and has incredible predictive accuracy. This model is also good at handling missing values.

The dataset used in the mobile network performance prediction also plays an important role to determine the accuracy of the model prediction. Previous research in [10] has stated that the accuracy of the results depends on how big the dataset is. The more data the machine learning able to learn and train, the accuracy of the model will be increased. In their research, the highest accuracy will be achieved when splitting the dataset 90% train and 10% test. Not to be forgotten that the number of features and parameters in the prediction is also a crucial part of determining the accuracy of the model. Not all the features in the dataset are used in predicting the performance of the mobile network. Therefore, significant analysis and discussion were presented by [11] have explored that the features need to be carefully chosen in the data acquisition phase. The feature selection process should be adapted to the current condition in which the process is.

Similarly, a comprehensive study on machine learning models in mobile networks [12] stated that the classical machine learning model such as exponential smoothing time series, Gaussian Process Regression, and random forest can provide a very good result for drive test data. Gaussian regression can predict better because the model enables essentially all the parameters involved to be predicted directly from the dataset while exponential smoothing time series can apply at different time scales at estimated linear and periodic trends as well as the final point predictions. For the random forest model, one of the major advantages of using the tress-based algorithm is the final regression model is understandable and has a sense of transparency especially after being compared to deep learning machine learning algorithms.

In this paper, the overview of the dataset used in this study will be presented. Then for methodology, machine learning will be applied for throughput prediction. The experimental result will be analyzed and discussed. Lastly, the comparative analysis will be made to determine which model performs the best in predicting throughput network performance.

III. Dataset

The collection and measurement of this dataset were conducted in EUP Subang Jaya (USJ) area covering from USJ 1 until USJ 13. The drive test was taken in the USJ area because the area was an urban area. NEMO Handy was used to

collecting network quality information and NEMO Analyzer was used to analyze and convert the measurement logfiles into a suitable data type. Figure 1 shows the drive test route.

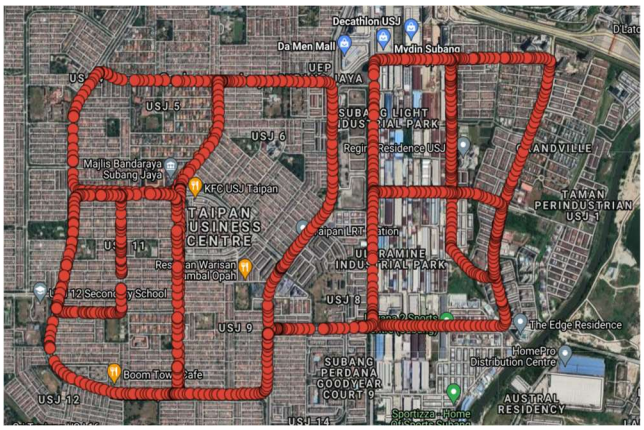

Fig. 1. Drive test measurement route

The total measurement route is 24km distance and 1.5km radius. The drive test speed is averaging from 60km/h. The measurement was taken in the morning, and it takes about 1 hour to finish. The telco service provider used on this drive test measurement and collection is Celcom. The measurement was taken several times, and this is the best measurement because the downlink throughput is varying relatively smoothly over the drive test region and the dataset has a lower number of missing data points. The dataset from the drive test contains multiple network parameters such as timestamp, longitude, latitude, RSRP, RSRQ, RSSI, SNR, and throughput. For the downlink throughput, 1Gbit of dummy file was used as the download file and will automatically repeat when the 1Gbit file is finished downloaded. The raw dataset contains 9886 rows and 8 columns. But there are many missing value gaps in the raw dataset because the drive test tools are unable to capture all the parameters for each row. To overcome this, the mean imputation method was used to solve the missing value gap. For every 9 rows, the data will be grouped by and replaced with the mean value. This method still maintains the variability of the dataset. After data cleaning, the dataset contains 1099 rows representing the network information and 8 columns representing the network parameters. Below is the listed log of measurement parameters and their description:

- Timestamp: The precise time when the measurement is taken.
- Longitude: One of the GPS coordinates of the mobile device.
- Latitude: One of the GPS coordinates of the mobile device.
- RSSI: Received Signal Strength Indicator. A measure of the strength in the presence of received radio signal only on 4G.
- RSRP: Reference Signal Received Power. This is the measure of the power of the reference signal.
- RSRQ: Reference Signal Received Quality. This indicates the quality of the received reference signal.
- SNR: Signal-to-noise ratio. Which indicates the ratio of signal power to the noise power in the decibels.
- Downlink throughput: Downlink bitrate of the measurement expressed in kbps.

A. Exploratory Data Analysis

The analysis and study were conducted in Python 3.7.0 using Jupyter Notebook where users can create an interactive environment that includes visualizations, live code, and text notes. For predicting mobile network performance, the targeted parameter is the downlink throughput. Therefore, timestamps are removed because this parameter will not affect the prediction and make the dataset more complicated thus will affect machine learning accuracy. By visualizing the data, some graphs and plots can be created which will easier to analyze and identify the hidden pattern in the dataset.

Figure 2 shows the downlink throughput in each geometry coordinate. From the figure, the downlink throughput value is varying relatively smoothly over the drive test region. It can be noticed that the majority of the downlink throughput value are between 20000kbps and 40000kbps. The area with a high downlink throughput value is a hotspot location such as petrol stations and restaurant areas. Theoretically, downlink throughput is related to SNR because the achievable data rate is determined by signal-to-noise ratio [12]. Figure 3 illustrates the relationship between downlink throughput with SNR and RSSI. The higher the SNR value, the higher the value for downlink throughput.

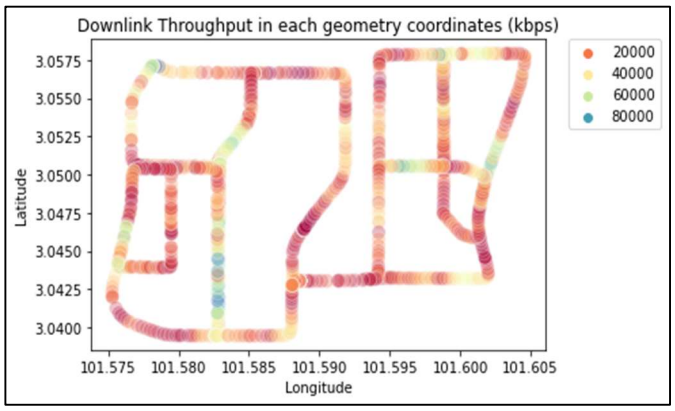

Fig. 2. Downlink throughput in each geometry coordinates

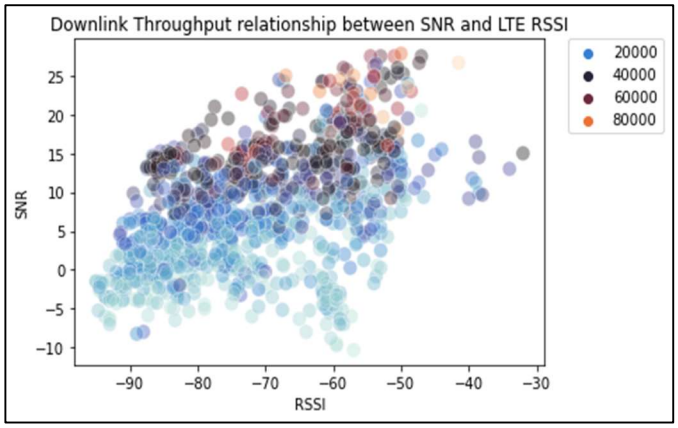

Fig. 3. Downlink throughput with SNR and RSSI

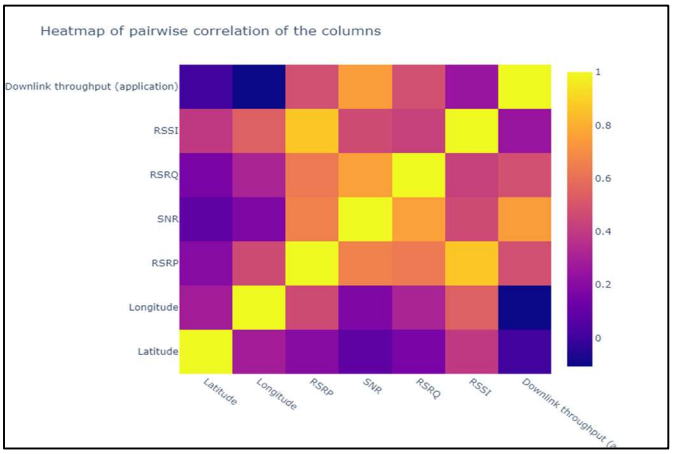

Fig. 4. Heatmap pairwise correlation between the columns

Next is the heatmap plot. It is useful when cross-examining multivariate data variables. Heatmaps plot are good for revealing any patterns to put on view whether the variable in the data is similar to each other and detecting correlations between the variable. Figure 4 shows the heatmap plot of the pairwise correlation between the columns. There is a high correlation between SNR and downlink throughput. This is because the value of SNR will affect downlink throughput the most comparable to the other network quality parameter.

IV. Performance Prediction

A. Machine Learning models

In this study, the commonly used classical machine learning models were used for prediction. These machine learning models were trained and tested for making the prediction. The performance of the models will be evaluated to determine the best model for predicting network performance.

1. Random Forest

One of the well-known techniques in machine learning algorithms is the random forest. It is a supervised machine learning algorithm. Random forest is a tree-based algorithm that uses multiple decision tree qualities features for making predictions or decisions. The term "forest" can be referred to as the combination of multiple decision trees. In the tree, the node stands for a feature, the branch stands for a decision and the leaf represents the final output. It uses the method of strategy which will random selection of a subset to grow on each tree. The tree will grow on a bootstrap sample of the dataset. It uses the majority vote in the inference stage overall tress to obtain the prediction [13]. The average value of all the output from the forest will be taken for making the prediction.

In addition, this machine learning algorithm is robust and fast compared to the other regression models. This predictive model can be used for both classification and regression tasks because of its versatility and it is a handy algorithm because of the hyperparameters used that always produce a good prediction output. To sum up, the machine learning random forest algorithm will merge the output from multiple decision trees and generate the final output.

2. Gaussian Process Regression

In recent years, many researchers have utilized the Gaussian Process Regression (GPR) machine learning model in predicting mobile network performance [14]. These machine learning algorithms are a Bayesian approach and non-parametric regression that make waves in the machine learning area. This model works well and has several benefits on a small dataset and the ability that can provide uncertainty measurements to the prediction [15]. By using the Bayesian approach, it infers the distribution probability over all the possible values. When doing prediction on unseen data points, the predictive distribution is calculated by way of including all the possible predictions on their calculated posterior distribution. Predictive distribution is calculated using equation (1) where $p(w|y, X)$ is the posterior distribution, w is the parameter, and $x*$ is the prediction distribution that can be calculated by weighing all the possibilities.

$$p(f*|x*,y,X) = \int_w p(f*|x*,w)p(w|y,X)dw \quad (1)$$

The likelihood and prior are the gaussian for the integration part and to be traceable. By using the assumption and solving the predictive distribution, Gaussian distribution is created and by that can obtain a prediction point using the mean and uncertainty quantification using the variance. Because it is a nonparametric algorithm, instead of calculating the probability distribution of parameters from a specific function, this model calculates the probability distribution over all the admissible functions that will fit the data. The algorithm will calculate the posterior predictive distribution on the points of interest.

3. K-Nearest Neighbor (KNN)

This machine learning model also is a supervised machine learning algorithm. This model is commonly used in regression prediction because of its applicability and simplicity in solving numerous real-world problems. This model will store all the available data and will predict based on the chosen similarity metric [16]. It will consider the K closest examples training point of interest for predicting. It can be done by a simple majority vote over the K closest point. The algorithm uses a distance measure to find similarities. Three aspects primarily affect the performance of the classifier such as the number of closest neighbors, Euclidean distance, and the decision rule. This model is easy to implement and versatile. In this study, the value K is 13 because the root mean squared error for that value of K is the lowest.

B. Evaluation Metrics

The evaluation metric was used to evaluate the performance of the machine learning model. Differ from machine learning classifications, error metrics has been developed to evaluate the quality of regression in machine learning model and enable to comparison of regressions among other machine learning model [17]. These metrics are useful and short summaries of the quality of the machine learning model and data for prediction. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2) are used to evaluate the regression problem's accuracy [18]. The metric error range is depending

on the targeted variable range. For this study, the downlink throughput is range between 20000 Kbps and 80000 Kbps. In the following subsections, the evaluation metrics are described.

1. Mean Absolute Error

Generally, an error is an absolute difference between the actual values and the predicted values. Mean absolute error will take the average of the error from every actual and predicted value in the dataset and give the output. This evaluation metric is usually used when then the performance of the machine learning model is measured on continuous data variables. The lower the mean absolute error value, the better the model's performance. Mean Absolute Error can be calculated using equation (2). Where n is the number of data points, y is the actual value and ŷ is the predicted output.

$$MAE = \frac{1}{n} \sum |y - \hat{y}| \quad (2)$$

2. Mean Squared Error and Root Mean Squared Error

For Mean Squared Error (MSE) is calculated by including the average of the square difference between the predicted value and actual values of the data. As for Root Mean Squared Error (RMSE), it is the standard deviation of the error that turns out when the prediction is made on the dataset. This metric same as Mean Square Error but the root value is considered while evaluating the accuracy of the model. The error is squared before averaged. MSE is the most useful metric when the dataset contains unexpected values or outlier data. But for RMSE it is more useful when large errors happened and affect the machine learning model's performance. This metric also uses the same concept. The lower the value, the better the performance of the model. MSE can be calculated using equation (3) where N is the total number of data points in the dataset and value i ranges from 1 to n.

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y - \hat{y})^2 \quad (3)$$

3. Coefficient of Determination

The coefficient of determination is also known as the R squared. This evaluation metric indicates how well the model fits in the dataset given [19]. It also indicates the machine learning model on how close the regression line between the predicted plotted values and the actual data values is. R^2 can be calculated in equation (4). Keep in mind that the coefficient regression values lie between 0 to 1, where 1 indicates that the machine learning model fits perfectly into the dataset which is quite impossible to achieve while 0 indicates that the model did not fit the dataset provided.

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares} \quad (4)$$

V. Results And Discussion

The data will be divided into two groups that are 80% for training and 20% for testing using train test split Scikit-Learn in Python library [20]. 824 data was randomly selected for training and the remaining 275 data is for testing.

The random forest algorithm is one of the machine learning algorithms that was used to predict network performance. The downlink throughput is the targeted parameter for predicting network performance because network throughput is a dependent variable with others quality parameters and context information. Also, the fact that throughput feature has a strong correlation with signal strength. Instead of only predicting, the random forest model is also capable determine the variable importance top list that influences the tree's impurity. The random forest algorithm has built-in feature importance that can be computed. By that, it can identify the variable importance list.

Table 1 shows the variable importance list. The variable importance list is based on the node on the random forest tree. It measures how many nodes use those features to reduce node impurity. The higher the score, the more that features contribute to the target variable prediction. As refers to the theory that the achievable data rate is determined by the Signal to Noise Ratio. This explains why SNR is top listed in the variable importance list. While for the longitude and latitude are also top listed in the variable importance list because of the location of the data measurement that has a strong or weak mobile network connection. If the location is at the hotspot area, that area has a strong mobile network connection.

TABLE I
VARIABLE IMPORTANCE LIST

Variable	Importance
SNR	0.63
Longitude	0.12
Latitude	0.07
RSRP	0.06
RSRQ	0.06
RSSI	0.05

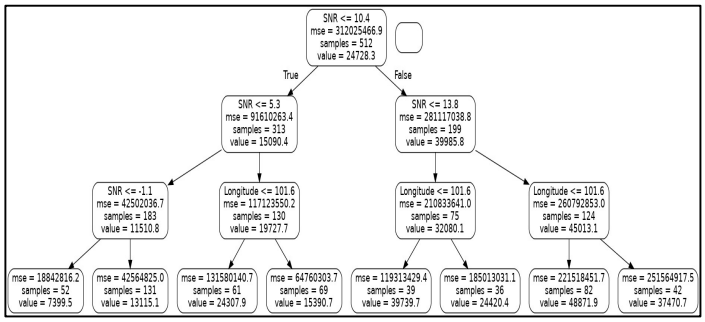

Fig. 5. Random Forest Regression Tree

Figure 5 illustrates one of the regression trees from the random forest model that is trained to obtain the prediction results. MSE in the leaf node shows the Mean Squared Error of the node, the sample is the number of data points in the node and the value is the prediction value (in Kbps) for all data points in the node. The decision tree structure makes immediate intuitive and physical sense to the network engineer that low SNR values will be resulting in a low predicted downlink throughput data rate.

For comparison between machine learning models, the model performance will be evaluated by the evaluation metric.

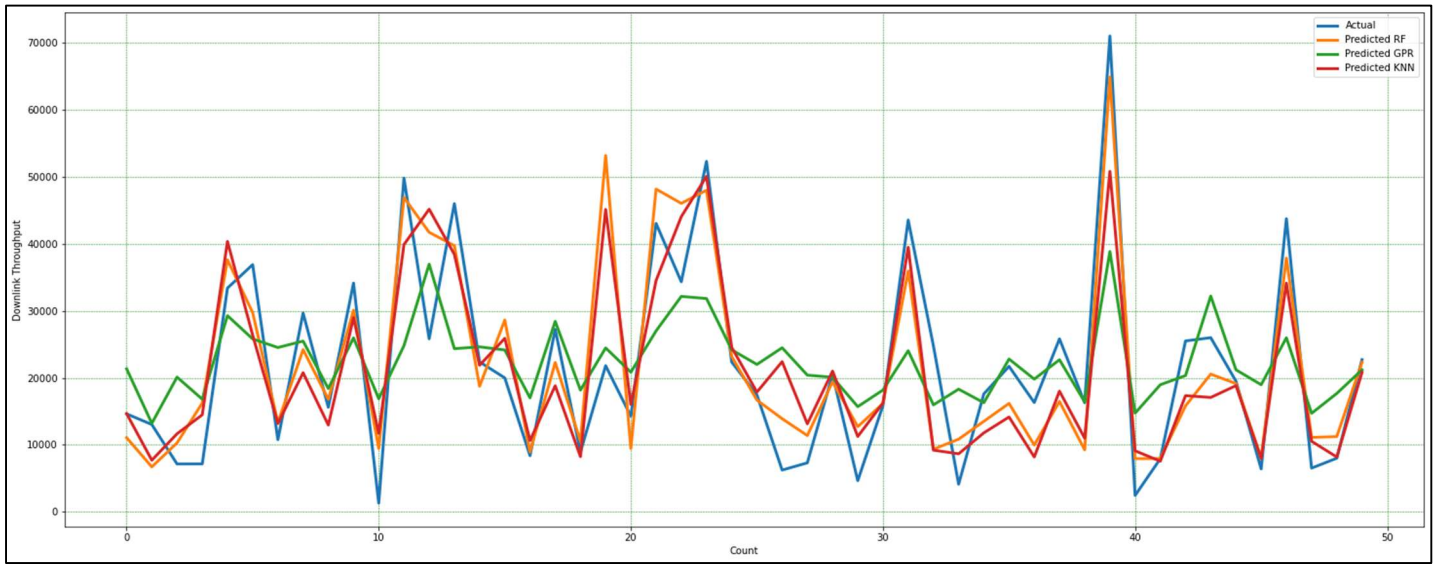

Fig. 6. Actual and Predicted Values

TABLE II
MACHINE LEARNING MODEL PERFORMANCE COMPARISON

Model	MAE	MSE	RMSE	R^2
RF	6754.16	83737811.45	9117.99	0.79
KNN	8114.77	133215994.3	11541.92	0.66
GPR	12606.98	260949579.7	16153.93	0.34

Table 2 shows the machine learning model comparison. These results show that the random forest model achieved the best performance with the lowest of all evaluation metric errors and the highest R^2 score. The error metric for the random forest is an MAE value of 6754.16, MSE value of 83737811.45, and RMSE value of 9117.99. This indicates that the difference between actual and predicted values is not too far. For the coefficient of determination score, the random forest model achieves the highest with 0.79. This designates that the random forest model fits in the dataset provided in predicting downlink throughput accurately. These results on the random forest are in accord with recent studies indicating that the random forest model obtains the highest performance prediction among the classical machine learning model [8]. The reason behind random forest's excellent performance is this model has an additional layer of randomness that can lessen the variance thus increasing the model accuracy. It also uses multiple decision trees to make predictions more accurate. This model also works well on a large dataset. On the other hand, the KNN model performance was the second-best with an MAE value of 8114.77, MSE 133215994.3, RMSE 11541.92, and an R^2 score of 0.66. The score is not too far from the random forest model. This model accuracy depends on the dataset quality and sensitivity to the irrelevant features. Therefore, feature selection needs to choose carefully to preserve the model performance. Because of that KNN model perform much better with fewer examples [21]. Lastly, gaussian process regression had the lowest performance of all. The metric error is very high with an MAE value of 12606.98, MSE value of 260949579.7, RMSE 16153.93, and R^2 score of 0.34 by far the lowest compared to the other model. These findings are rather disappointing because the model will lose efficiency in a high-dimensional space. By way of explanation, the major disadvantage of this model is bad at predicting when dealing with a large volume of datasets [22]. Together these results provide important insights that among those three machine learning models, the random forest model is the best machine learning model in predicting mobile network performance in the identified dimension.

Figure 6 shows illustrate the performance of all models on the test set that plot the predicted throughput values as well as the actual values. From the plot, we can see that the random forest predicted value are not too far from the actual values compared to the KNN and gaussian process regression.

The hyperparameter tuning in the random forest is the parameters of the algorithm that can be tuned to improve the model performance and prediction accuracy. It can be done by using the library function GridSearchCV or using the trial-and-error method. Since the random forest is the best model for throughput prediction, it will go through a hyperparameter tuning to enhance the performance. The hyperparameters are the number of decision trees in the forest and the number of max depths. Before tuning, the number of trees in the random forest model is 2000. Now the number of trees has increased to 5000. Past study has stated that higher numbers of trees in the random forest will be resulting in higher accuracy but slower learning time [23]. Model accuracy is often improved by adding more decision trees because the predictions are made based on the average sample from multiple trees. Next, the max depth parameter signifies the depth of the tree in the forest. This parameter is tuned by setting the max depth to 20. [24] has studied the relationship between the random forest accuracy and the max depth parameter. The resulting state that tuning the maximum depth parameter will improve the random forest accuracy. But it needs to tune to reach the optimal value to avoid overfitting. Usually, max depths range from 1 to 32. Table 3 below presents the comparison of random forest performance before and after hyperparameter tuning. Both MAE and RMSE have a slight improvement of about 100 Kbps. For MSE the difference is 2000000 Kbps. Lastly coefficient of

determination increases by 0.005. These results provide an important insight that hyperparameter tuning does improve the accuracy of the training model. Therefore, it is suggested that doing the hyperparameter tuning step to make a better prediction on downlink throughput.

TABLE III
MACHINE LEARNING MODEL PERFORMANCE COMPARISON

Model	MAE	MSE	RMSE	R^2
RF	6754.16	83737811.45	9117.99	0.7919
RF Optimize	6673.1	81301068.73	9016.71	0.7964

VI. CONCLUSION AND FUTURE WORK

The main goal of this paper is to propose a machine learning model for mobile network performance prediction in the identified dimensions. Then benchmarking various types of machine learning models to identify the best model for throughput prediction. The following conclusions can be drawn from this study. The random forest model is the best machine learning model for predicting downlink throughput. Random forest achieves the lowest evaluation metric error and the highest R^2 score. Since the model has an additional layer of randomness it can lessen the variance thus increasing the model accuracy. It is suggested that mobile network operators use a random forest model in predicting mobile network performance. Referring back to the variable importance list, the location of measurement and SNR feature play an important role in affecting network performance. To meet users' network Quality of Service (QoS), network operators need to improve network signal strength in a certain area. In addition, the results show that tuning the hyperparameter number of trees and depth of the trees slightly reduces the evaluation metric error and increases the R^2 score.

For future work, using an extensive real-world drive test dataset, a bigger dataset with the addition of variability features such as cell information and the number of users can be used in throughput prediction to make throughput feature dependent on many other features. Other machine learning models should be added for comparison in the interest of variety. Besides, the deep learning technique also should be applied in throughput prediction. For hyperparameter tuning, other parameters should be considered such as minimum samples split, minimum samples leaf, and max features [25].

REFERENCES

[1] "QoE drive testing | rf drive test tool | Rantcell." https://rantcell.com/qoe-drive-testing.html (accessed Jul. 21, 2021).

[2] "Minimization of Drive Tests (MDT) – TTS Wireless." https://www.ttswireless.com/minimization-of-drive-tests/ (accessed Jul. 21, 2021).

[3] J. Thrane, M. Artuso, D. Zibar, and H. L. Christiansen, "Drive Test Minimization Using Deep Learning with Bayesian Approximation," *IEEE Veh. Technol. Conf.*, vol. 2018-Augus, pp. 1–5, 2018, doi: 10.1109/VTCFall.2018.8690911.

[4] "The Role of Predictive Analytics in Network Performance Monitoring | Kentik." https://www.kentik.com/blog/the-role-of-predictive-analytics-in-network-performance-monitoring/ (accessed Jul. 21, 2021).

[5] R. Kaur, "Machine Learning Technique for Wireless Sensor Networks," pp. 332–335, 2021.

[6] "A Gentle Introduction to Model Selection for Machine Learning." https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/ (accessed May 24, 2022).

[7] A. Samba *et al.*, "Throughput Prediction in Cellular Networks : Experiments and Preliminary Results To cite this version : HAL Id : hal-01311158 Throughput Prediction in Cellular Networks : Experiments and Preliminary Results," 2016.

[8] H. Elsherbiny, H. M. Abbas, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "4G LTE Network Throughput Modelling and Prediction," pp. 3–8, 2020.

[9] H. Gebrie, H. Farooq, and A. Imran, "What Machine Learning Predictor Performs Best for Mobility Prediction in Cellular Networks ?," no. i, 2019.

[10] M. Elmasry, "Predict Network , Application Performance Using Machine Learning and Predictive Analytics," 2019.

[11] B. Sliwa and S. Member, "Data-Driven Network Simulation for Performance Analysis of Anticipatory Vehicular Communication Systems," *IEEE Access*, vol. 7, pp. 172638–172653, 2019, doi: 10.1109/ACCESS.2019.2956211.

[12] J. Riihijarvi and P. Mahonen, "Machine Learning for Performance Prediction in Mobile Cellular Networks," *IEEE Comput. Intell. Mag.*, vol. 13, no. 1, pp. 51–60, 2018, doi: 10.1109/MCI.2017.2773824.

[13] "Random Forest in Python. A Practical End-to-End Machine Learning… | by Will Koehrsen | Towards Data Science." https://towardsdatascience.com/random-forest-in-python-24d0893d51c0 (accessed Jul. 16, 2021).

[14] M. Joshi and T. H. Hadi, "A Review of Network Traffic Analysis and Prediction" 2015, [Online]. Available: http://arxiv.org/abs/1507.022.

[15] "Quick Start to Gaussian Process Regression | by Hilarie Sit | Towards Data Science." https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319 (accessed Jul. 16, 2021).

[16] "Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science." https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Feb. 10, 2022).

[17] "Tutorial: Understanding Linear Regression and Regression Error Metrics." https://www.dataquest.io/blog/understanding-regression-error-metrics/ (accessed Jul. 16, 2021).

[18] "What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared? - Studytonight." https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared (accessed Jul. 16, 2021).

[19] "Numeracy, Maths and Statistics - Academic Skills Kit." https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html (accessed Feb. 10, 2022).

[20] F. Pedregosa, G. Varoquaux, and A. G. V. M. B. Thirion, "Scikit-learn: Machine Learning in Python Fabian," *J. ofMachine Learn. Res. 12 2825-2830*, vol. 127, no. 9, 2012, doi: 10.1289/EHP4713.

[21] K. Gayathri and A. Marimuthu, "Text document pre-processing with the KNN for classification using the SVM," *7th Int. Conf. Intell. Syst. Control. ISCO 2013*, pp. 453–457, 2013, doi: 10.1109/ISCO.281197.

[22] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, and S. Manzhos, "Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy," *J. Chem. Phys.*, vol. 148, no. 24, 2018, doi: 10.1063/1.5003074.

[23] "The Ultimate Guide to Random Forest Regression." https://www.keboola.com/blog/random-forest-regression (accessed May 20, 2022).

[24] N. R. Darji and S. A. Ajila, "Increasing Prediction Accuracy for Human Activity Recognition Using Optimized Hyperparameters," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, no. i, pp. 2472–2481, 2020, doi: 10.1109/BigData50022.2020.9378000.

[25] "Hyperparameters of Random Forest Classifier - GeeksforGeeks." https://www.geeksforgeeks.org/hyperparameters-of-random-forest-classifier/ (accessed May 22, 2022).