Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Finding Hidden Links among Variables in a Large-Scale 4G Mobile Traffic Network Dataset with Deep Learning

Ndolane Diouf (ndolane.diouf89@gmail.com)
Cheikh Anta Diop University of Dakar https://orcid.org/0000-0003-4966-9648
Massa Ndong
Senegal Virtual University https://orcid.org/0000-0001-5773-7589
Dialo Diop
Cheikh Anta Diop University of Dakar
Kharouna Talla
Cheikh Anta Diop University of Dakar
Aboubaker Chedikh Beye
Cheikh Anta Diop University of Dakar
Ibrahima Gueye
Polytechnic Institute (ESP), Cheikh Anta Diop University

Research Article

Keywords: dataset, machine learning, throughput, DL_perceived_throughput, mobile networks, 4G, small cell networks, deep learning

Posted Date: April 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1530463/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Using a mobile dataset from Orange Senegal small cell 4G network, we study the effect of variables such as data traffic at the downlink level, data traffic at the uplink level, total data traffic, maximum number of active users, signaling protocol, uplink user rate, downlink user rate, physical resource block rate for downlink, block rate physical resources for uplink, load data logging, channel quality indicator, downlink radio delay average, over the perceived rate at the downlink. We are looking to find the variable that most affects the perceived flow. We do this by using machine learning to find the variable that closely explains the variation in perceived flow and helps predict flow with greater accuracy. We observe that correlation analysis is unable to find a hidden relationship between throughput and other variables. With models such as linear regression, decision tree, random forest, multi-layered perceptron and deep neural network, the channel quality indicator (CQI_Avg) turns out to be the variable that more closely explains the variation in perceived flow and more accurately contributes to the prediction compared to others variables.

1 Introduction

Today we are witnessing an exponential growth in mobile traffic following the deployment of very high speed mobile networks such as 3G, 4G / LTE and 5G [1]. A key factor in the success of a mobile telephone operator is therefore its ability to meet subscribers' expectations in terms of coverage. Therefore, small cells are widely deployed and will continue to play a very important role in the presence of next generation cellular networks like the fifth generation (5G). Small cells may experience a problem related to the quality of the channel.

Considering the very high speed applications to come, the effect of the quality of the channel on the variation of the rate perceived at the level of the downlink is of significant interest and the CQI_Avg is one of the variables whose relation with the perceived rate of downlink (DL_perceived_throughput) is discussed in this article. Channel quality indicator (CQI) reporting is an analysis tool for LTE systems to evaluate user quality of experience at the MAC-level throughput [2]. Such an indicator contributes to the QoS assessment. By using the CQI values and other protocols, the work in [2] developed an analytical method to calculate the MAC-level downlink throughput of LTE systems in the presence of Rayleigh fading. In our work, we assess the average CQI impact as a variable influencer of the user perceived-throughput.

The spectrum scarcity has led to a comprehensive work on license assisted access such as LTE systems accessing WiFi spectrum. The impact of unlicensed shared in-band signaling is presented in [3] by assessing the influence that such signaling has on the performance of both the LTE and WiFi systems. Thus, an assessment of the impact of downlink shared channel control signaling for the indoor coexistence scenario is presented [3]. In [4] the authors model the downlink packet scheduling problem for cellular networks, which optimizes throughput, fairness, and packet drop rate. They proposed a Deep Reinforcement Learning (DRL) framework with the Actor-Critical Advantage (A2C) algorithm for the

optimization problem. The authors in [5] have developed practical user scheduling algorithms for downstream burst traffic with an emphasis on user fairness. They propose to use the user rate of 5% tile (5TUDR) as a metric to assess user fairness.

The transmission rate is a very important metric which allows to optimize the quality of service for the users of mobile networks. Prior knowledge of throughput on different channels can help in choosing the best channel in an overlapping Basic Service Set (BSS) deployment to mitigate inter-BSS interference [6]. Artificial intelligence (AI) has become the revolutionary technology of the 20th century and has found multiple applications in fields ranging from weather forecasting, astronomical exploration to autonomous systems [7].

In this work, we use machine learning (ML) and deep learning (DL) to find a hidden relationship between variables in Orange Senegal's small cell 4G mobile network, by performing analyzes on datasets large-scale motives collected from small cells. In this study, our objective is to use the techniques of ML (Random Forest (RF), Decision Tree (DT), Linear Regression (LR)) and DL (Deep Neural Network (DNN), Multi-Layer Perceptron (MLP)), on a mobile dataset to find the variable in the 4G network that can more accurately explain the variation in perceived throughput at the downlink level and predict the perceived throughput with greater accuracy. A comparison between a deep learning technique such as a Deep Neural Network (DNN) and those used in our previous work [8] is made using the evaluation metrics. We perform correlation analysis and prediction analysis. We use the root mean square error to evaluate the performance of a variable used to predict the perceived flow on the downstream. In addition, the performance of a prediction model is evaluated in terms of prediction accuracy, root mean square error (RMSE) and mean absolute error (MAE).

From the correlation analysis, one can imagine that the user downlink throughput is more likely to act as a reasonable predictor of the perceived throughput while the other variables should not reasonably predict the perceived throughput at the level of the user. downlink. The meaning of each of these variables is given in Table 1.

Predictive analysis gives an RMSE of *6.04* and *5.43* when CQI_Avg is used to generate the DNN and RF model respectively. When used to generate the DT model, the CQI_Avg achieves an RMSE of *5.78*. We get an RMSE of *5.65* and *5.72* for LR and MLP respectively when this same variable is used to generate each of these models. Comparison between the ML and DL algorithms using real 4G data sets shows that DNN offers the highest accuracy and best RMSE (i.e. *accuracy = 96.1%, RMSE = 0.73*), followed by RF (*93.31%, 0.83*), LR (*91.75%, 0.92*), DT (*81.01%, 1.40*) and MLP (*75.73%, 1.58*) respectively. Thus the DNN outperforms other algorithms.

The contributions of this article are:

- After deploying 7 small Star Pole type cells to ensure indoor coverage and increased capacity, we collected a set of raw data exchanged between these 7 small cells for a period of more than three months, from August 3, 2020 to November 10, 2020. By definition, small cells of the Pôle Star type

are a specific dedicated solution for indoor coverage which has been set up by Huawei and which calls it Pôle Star.

- Mobile data in raw form is not suitable for use as input in ML and DL models. Therefore, the data preprocessing and the choice of interesting features for our study are carried out.
- This work compared different ML and DL algorithms available for data processing and identified DNN as the best model to predict the perceived throughput at the downlink level of Orange Senegal's 4G small cell network. Additionally, this work applied ML and DL to discover CQI_Avg, as the variable among others that more accurately predicts and therefore most significantly affects perceived throughput on 4G.

The remainder of the article is organized as follows: Section 2 presents previous studies focused on the use of machine learning techniques for prediction. Section 3 details the materials, dataset description, data preprocessing, and data analysis of the regression algorithms used. The stages in performing the various analyzes are presented in section 4. Section 5 presents the results of these analyzes. Section 6 summarizes the results.

2 Related Works

ML and DL techniques like LR, DT, RF, MLP and DNN have been widely used in the literature for prediction [9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19]. Linear regression is used to predict the values of a dependent variable by observing an independent variable. We use the LinearRegression [20] function in the Sklearn package [21] in python to generate the model from the training dataset using LR. A DT is a nonparametric algorithm and can model arbitrarily complex relationships between inputs and outputs, without any a priori assumptions. DT are capable of handling heterogeneous data. DT inherently implement feature selection, making them robust to irrelevant or noisy variables. Further details on decision tree algorithms can be found in [22]. RF is a statistical or ML algorithm for prediction. RF has the ability to model highly nonlinear relationships. This is a very powerful ensemble method [23] which combines the results of many DT. The end result of this RF algorithm is the average of all developed trees. We use the RandomForestRegressor class from the Sklearn.ensemble library in python to solve regression issues through a random forest. We also use the DecisionTreeRegressor and RandomForestRegressor functions in the Sklearn package in python to generate the models from the set of training data using DT and RF, respectively. ANNs (or MLPs) are supervised learning algorithms used to model the complex nonlinear relationship between one or more inputs (characteristics) and a true-valued output (target). The goal of neural networks is therefore to train a network with the available data in order to have the best possible correlation between the measured output data and those estimated [24]. MLPs are layered architectures typically arranged as an input layer, one or more hidden layers, and an output layer. In this work we set the size of the input layer to 13 which corresponds to the number of features in our dataset. The MLPRegressor function in the Sklearn package in python is used to generate the model from the training dataset using MLP. DNN is a class of machine learning algorithms similar to the artificial neural network and aims to mimic the information processing of the brain. The DNN models are recently becoming very

popular due to their excellent performance to learn not only the nonlinear input–output mapping but also the underlying structure of the input data vectors [25]. Details on DNNs can be found in the following works, [26, 27, 28, 29, 30, 31, 32, 33 and 34].

3 Materials, Dataset And Machine Learning Algorithms

3.1 Materials

The materials used in this study include the following packages and libraries: Numpy, Pandas, SciPy, Scikit-Learn, Tensorflow, Keras and Matplotlib. We used an HP laptop with an Intel (R) Celeron (R) N3350 processor, 4 GB of memory.

3.2 Dataset

The dataset used for our analysis is provided by Orange Senegal. This dataset is collected from the 4G radio access network called LTE (Long Term Evolution). The Small Cell solution of the type is a specific solution dedicated to indoor coverage and capacity building which has been set up by Huawei. In total, 7 sites (MASSALIK_P1, MASSALIK_P2, MASSALIK_P3, MASSALIK_P4, MASSALIK_P5, MASSALIK_P6, MASSALIK_P8) of the type have been deployed at the zone level. Thus, we collected a set of data on Orange Senegal's 4G mobile network on these 7 Pôles Stars sites. However, choosing all the available variables in a dataset is not always a good option as it can lead to poor predictions. The descript the variables is given in Table 1. After preprocessing we get a dataset with 583 rows and 13 columns, DL_perceived_throughput is the target variable. An overview of the statistical distribution of this data set is presented in “Fig. 13”.

Table 1
Description of the dataset

Variables	Description
DL_Traffic (GB)	Represents downlink data traffic
Traffic_UL (GB)	Represents uplink data traffic
Total_Traffic (GB)	Represents total data traffic
Active_user_Max	Represents the maximum number of active users
RRC_user_Max	Defines the protocol allowing the UE and the l'(e)NB to exchange signaling
throughput_UL_user (Kbit/s)	Represents the user throughput of the uplink
Throughput_DL_user (Kbit/s)	Represents the user throughput of the downlink
DL_PRB_Rate	Represents the physical resource block rate for the downlink
UL_PRB_Rate	Represents the physical resource block rate for the uplink
CDR	Represents charging data recording
CQI_Avg	Represents the channel quality indicator
Radio_DL_Delay_avg	Defines the average of the downlink radio delay
DL_perceived_throughput (Mbps)	Represents the perceived throughput of the downlink
MASSALIK_P1, MASSALIK_P2, MASSALIK_P3, MASSALIK_P4, MASSALIK_P5, MASSALIK_P6, MASSALIK_P8	Represents small Pole Star type cells (or Pole Star sites)

3.3 Machine learning algorithms

LR, DT, RF, and ANN or MLP are the machine learning algorithms used in this article. LR is an ML technique that allows we to model linear relationships between variables. DT is a supervised, non parametric ML method capable of finding complex nonlinear relationships in data. DT can perform classification and regression tasks. In this article, we focus on DT with a regression task. The significant hyperparameter that can improve the performance of the algorithm is the maximum tree depth that we set to 4 (i.e. max_depth = 4). Further details on decision tree algorithms can be found in [35]. The decision tree built from the DT model is presented in "Fig. 1".

RF [36] is a combination of several DT, i.e. a forest, where each tree is generated from a new training data set, which is a randomly sampled subset from the original training dataset. The hyper parameters of the RF algorithm include the number of trees defined as 200 in this article, the maximum characteristics in an individual tree, and the minimum number of leaf nodes needed to split an internal node.

ANNs or MLPs are ML techniques that allow the modeling of nonlinear relationships between variables. In this paper, the MLP consists of a system of single interconnected neurons, or nodes, which is a model representing a non-linear mapping between an input vector and an output vector. The nodes are linked by weights and output signals which are a function of the sum of the inputs of the node modified by a simple nonlinear transfer or activation function. It is the superposition of many simple nonlinear transfer functions that allows the MLP to approximate extremely nonlinear functions. The output of a node is scaled by the connection weight and forwarded to be input to nodes in the next layer of the network. It involves a direction of information processing, hence the multi-layered perceptron is known as the anticipatory neural network. The input layer does not play any computational role but simply serves to transmit the input vector to the network. The terms input and output vectors refer to the inputs and outputs of the multilayer perceptron and can be represented as single vectors. The input is made up of 13 neurons and we have 3 hidden layers each containing 13 neurons. A single layer consisting of a neuron represents the output.

A little detail on DNN

Deep learning is a machine learning technique that uses multiple-layer architectures to extract features at different levels of abstraction from raw data. These different layers are generally organized in the form of a neural network. A basic DNN is a composition of several layers consisting of weights. Each layer takes the output of its previous layer as input and typically applies a nonlinear activation function to calculate its own output. In our study, the Rectified Linear Unit (ReLU) activation function is used. ReLU is the most commonly used activation function in deep learning models. The function returns 0 if it receives a negative input, but for any positive x value it returns that value. At the final layer, this results in rate prediction and this process is known as feedforward propagation. A visualization of an DNN is shown in “Fig. 2”. During the training phase, the prediction loss of the model is calculated via a pre-defined loss function. The loss is then back-propagated through the network via gradient descent to update the weights of each layer. This cycle of forward and backward propagation is repeated which reduces the loss until convergence. Models with this basic architecture are known as MLP and have already been used to model data both in the network and in other application domains [37, 25]. The difference between DNN and MLP models is that we have forward propagation for the former while at the latter we have a repeated cycle of forward and backward propagation.

The classical supervised learning approach needs features and labels. For this purpose, data traffic at the downlink level (I_1), data traffic at the uplink level (I_2), total data traffic (I_3), maximum number of active users (I_4), signaling protocol (I_5), uplink user rate (I_6), downlink user rate (I_7), physical resource block rate for downlink (I_8), block rate physical resources for uplink (I_9), load data logging (I_{10}), channel quality

indicator (I_{11}), downlink radio delay average (I_{12}), are used as features. The perceived rate at the downlink is the label for the model. The DNN was designed as shown in “Fig. 2” with trees hidden layers. The O1 output of the DNN is the predicted value of the perceived throughput of the downlink.

As shown in this “Fig. 2”, the network architecture has sequential, densely connected layers with rectified linear unit activation functions as neurons. The mean square error between predicted DL_perceived_throughput and measured DL_perceived_throughput was minimized during training process with Adam optimizer. The ReLU activation function is used at each of the hidden layers. The activation function creates a non-linear relationship between the input and the output [37].

The operator of Orange Senegal could implement these ML algorithms through deployments in a cloud environment, on edge (in devices or sensors), or in an on-premises environment.

4 Analysis Methodology

The impact of the independent variables on the dependent variable or the target variable is the subject of this analysis. In our case, the dependent variable is DL_perceived_throughput and all other variables are the independent variables. We use five ML techniques, LR, DT, RF, MLP and DNN to find any hidden relationship between other variables and the perceived rate of the downlink, more precisely, we want to discover the variable that most significantly affects the perceived throughput.

We begin our analysis by examining the visible correlation between other variables and DL_perceived_throughput. Then these five techniques are used to try to find a hidden relationship between other variables and DL_perceived_throughput. Under these specific conditions, 80% of the data in the randomly selected data set constitutes the training set, while the remaining 20% of the data comprises the test set. “Fig. 3” visualizes all training and testing data We use the variables from the training set as input to LR, DT, RF, MLP or DNN to train the model. Finally, we use the model trained with variables from the test set to predict the perceived throughput, i.e. DL_perceived_throughput in the test set.

5 Results And Discussions

5.1 Correlation analysis

The correlation between the characteristics of the dataset provides crucial information about the characteristics and the degree of influence they have on the dependent variable. Pearson's heat map showing the correlation between features is shown in “Fig. 4 ”.

This “Fig. 4” reveals a strong positive correlation between variables such as Throughput_DL_user, CQI_Avg and DL_perceived_throughput. A relatively weak positive correlation is observed between CDR and DL_perceived_throughput. We have a weak negative correlation between DL_perceived_throughput and the other variables.

5.2 Predictive analysis

During this analysis, we compare, using ML techniques, the performance of different RMSE variables to predict the perceived throughput of the downlink. RMSE determines the error between the predicted and actual values of the flow in the test set and its expression is given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(A|i - P_i\right)^2}$$

1

Where A_i is the ith value of DL_perceived_throughput in the test set, P_i is the ith corresponding predicted value of DL_perceived_throughput, and N is the number of DL_perceived_throughput observations in the test set. Using the LR technique gives us in Table 2 the RMSE between the actual values of DL_perceived_throughput in the test set and the predicted values.

Table 2
RMSE between actual values of DL_perceived_throughput in test set and predicted values when using LR model

Method	Variable used in prediction for DL_perceived_throughput	RMSE
LR	DL_Traffic	10.17
	Traffic_UL	11.91
	Total_Traffic	11.33
	Active_user_Max	7.60
	RRC_user_Max	33.51
	throughput_UL_user	3331.72
	Throughput_DL_user	11937.21
	DL_PRB_Rate	8.54
	UL_PRB_Rate	8.54
	CDR	13.12
	CQI_Avg	**5.65**
	Radio_DL_Delay_avg	52.50

An RMSE of 5.65 is obtained when CQI_Avg in the test set is used in combination with the LR model to generate predicted values for DL_perceived_throughput. With the DT, the RMSE between the actual values of DL_perceived_throughput in the test set and the predicted values is given in Table 3.

Table 3
RMSE between actual values of DL_perceived_throughput in test set and predicted values when using DT model

Method	Variable used in prediction for DL_perceived_throughput	RMSE
DT	DL_Traffic	9.96
	Traffic_UL	12.00
	Total_Traffic	11.11
	Active_user_Max	7.67
	RRC_user_Max	33.49
	throughput_UL_user	3331.70
	Throughput_DL_user	11937.14
	DL_PRB_Rate	8.52
	UL_PRB_Rate	8.61
	CDR	13.23
	CQI_Avg	**5.78**
	Radio_DL_Delay_avg	52.37

We get an RMSE of 5.78 when CQI_Avg in the test set is used in combination with the fitted DT model to generate the predicted values for DL_perceived_throughput. Using the RF technique, the RMSE between the actual values of DL_perceived_throughput in the test set and the predicted values is given in Table 4.

Table 4
RMSE between actual values of DL_perceived_throughput in test set and predicted values when using RF model

Method	Variable used in prediction for DL_perceived_throughput	RMSE
RF	DL_Traffic	9.96
	Traffic_UL	11.71
	Total_Traffic	11.15
	Active_user_Max	7.37
	RRC_user_Max	33.54
	throughput_UL_user	3331.85
	Throughput_DL_user	11937.38
	DL_PRB_Rate	8.28
	UL_PRB_Rate	8.31
	CDR	12.94
	CQI_Avg	**5.43**
	Radio_DL_Delay_avg	52.54

From this Table 4, an RMSE of 5.43 is obtained when CQI_Avg in the test set is used in combination with the RF model fitted to give the predicted values of DL_perceived_throughput. With ANN or MLP, the RMSE between the actual values of DL_perceived_throughput in the test set and the predicted values is given in Table 5.

Table 5
RMSE between actual values of DL_perceived_throughput in test set and predicted values when using MLP model

Method	Variable used in prediction for DL_perceived_throughput	RMSE
MLP	DL_Traffic	10.12
	Traffic_UL	11.90
	Total_Traffic	11.26
	Active_user_Max	7.49
	RRC_user_Max	33.16
	throughput_UL_user	3331.74
	Throughput_DL_user	11937.26
	DL_PRB_Rate	8.49
	UL_PRB_Rate	8.53
	CDR	13.13
	CQI_Avg	**5.72**
	Radio_DL_Delay_avg	52.45

This Table 5 gives an RMSE of 5.72 when CQI_Avg in the test set is used in combination with the MLP model fitted to give the predicted values for DL_perceived_throughput. With the DNN model, the RMSE between the actual values of DL_perceived_throughput in the test set and the predicted values is given in Table 6.

Table 6
RMSE between actual values of DL_perceived_throughput in test set and predicted values when using DNN model

Method	Variable used in prediction for DL_perceived_throughput	RMSE
DNN	DL_Traffic	11.96
	Traffic_UL	12.11
	Total_Traffic	13.30
	Active_user_Max	7.80
	RRC_user_Max	31.12
	throughput_UL_user	3159.79
	Throughput_DL_user	12292.75
	DL_PRB_Rate	9.01
	UL_PRB_Rate	8.77
	CDR	13.28
	CQI_Avg	**6.04**
	Radio_DL_Delay_avg	56.50

From this Table 6, an RMSE of 6.04 is obtained when CQI_Avg in the test set is used in combination with the DNN model fitted to give the predicted values of DL_perceived_throughput. Thus, the DNN model further confirms that the Channel Quality Indicator (CQI) represented here by the CQI_Avg parameter is a good predictor of throughput because it gives the lowest RMSE compared to the other input parameters of the DNN model.

5.3 Model prediction performance.

In this part, we use training and test datasets to train and evaluate each of our models.

“Fig. 5” illustrates actual values of DL_perceived_throughput vs. Predictions generated when using the LR, DT, RF and MLP models. The black colored line reffered to as test set in this figure represents the actual values of the DL_perceived_throughput in the test set.

We evaluate these five ML models (LR, DT, RF, MLP and DNN) on the dataset using two regression metrics: RMSE and R-Squared.

The RMSE is the square root of the root mean square errors and its expression is given in Eq. (1).

The R-Squared metric, also known as the coefficient of determination, provides an indication of how the model predicts invisible values. R-Squared is calculated using Eq. (2).

$$r^2 = 1 - \frac{SSE}{SST_0}$$

2

where SSE is the sum of the squared error and SST_0 is the total sum of the squared values.

$$SSE = \sum_{i=1}^{N} \left(P_j - \bar{A}\right)^2$$

3

$$SST_0 = \sum_{i=1}^{N} \left(P_i - \bar{A}\right)^2$$

4

where Pj, $\bar{A}$ are respectively the predicted and mean value of the target variable.

The MAE metric measures the average of the magnitude of errors over the test set with N data points. Its expression is given in Eq. (5).

$$MAE = \frac{1}{2N} \sum_{i=1}^{N} \left|A_i - P_i\right|$$

5

RMSE and MAE are negatively oriented scores, which means lower values are better.

In addition, RMSE is preferable to MAE when the behavior of outliers is important as is the case in our study.

The prediction performance of each ML algorithm is evaluated on a set of test data. The predicted values of DL_perceived_throughput versus actual values in the data set are plotted in “Fig. 6” for models (LR, DT, RF, MLP).

After implementing the DNN model, a set of training and testing or validation data to measure model performance. Thus, the MAE metric is used to see training and validation losses. These losses can be seen in “Fig. 7”.

This figure above shows low training and validation losses that can justify the throughput prediction performance with this DNN model.

The prediction performance of the DNN model can be seen in “Fig. 8”.

As we can observe in this “Fig. 8” above, our DNN model predicts the throughput with great accuracy.

“Fig. 9” illustrates the relationship between the actual values of each predictor (independent variable) and the predicted values of DL_perceived_throughput by DNN model.

Performance Comparison of All Models

The comparison between the different models in this study is made based on evaluation metrics such as accuracy, MAE and RMSE. The static results of these metrics are grouped in the table below.

Table 7
Performance evaluation

Metrics	LR	DT	RF	MLP	DNN
MAE	0.72	0.90	0.63	1.10	**0.49**
RMSE	0.92	1.40	0.83	1.58	**0.73**
accuracy	91.75%	81.01%	93.31%	75.73%	**96.1%**

The results listed in Table 7 show that the five models can be used effectively for prediction. According to Table 7, DNN offers better performance (i.e. low RMSE value = 0.73, smaller MAE value = 0.49 and with higher accuracy = 96.1%).

For more details on the comparison between these different models, we have visualized in the form of a bar chart each of the evaluation metrics for each model. “Fig. 10, 11, 12” illustrate these bar charts.

As can be seen, “Fig. 10, 11, 12” show that the DNN is the best model because it offers the smallest RMSE and MAE. This DNN world also offers the highest accuracy up to 91.1%.

We present below the functional diagram of the model / system.

Random Forest – Decision tree

1. For k=1 to B

(a) Draw N sample points from the collected data from the MUEs and the neighboring small cell base stations (SBSs) to form a bootstrap at the designated SBS

(b) Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum size n_{min} is reached

i. Select m variables at random from the p variables
ii. Pick the best variable/split-point among the m variables
iii. Split the node into two daughter nodes

2. Output the ensemble of trees $\left\{T_b\right\}_1^B$.

The prediction of a new parameter value from the u = input data x is given by the regression

$$f_{RF}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

The classification is given by the majority vote as follows:

Let $C_b(x)$ be the class prediction of the both random forest tree, then

$$C_{RF}^{B}(x) = majorityvote\{C_b(x)\}_{1}^{B}$$

With the proposed RF algorithm, each SBS applies the RF locally using its own data and the data received from the neighboring SBSs to construct the bootstrap. The contribution to this scheme is the sharing of data by the SBSs which enables a dynamic cooperation clustering and effective parameter classification. The cluster of SBSs exchanging data is of a variable size too.

The optimization technique used for our model is cross validation. The method used is GridSearchCV from Scikit-Learn.

6 Conclusion

In this article, we investigated the relationship between perceived downlink throughput and other variables of Orange Senegal's 4G network. Our main objective was to find which variable significantly affected the perceived downlink speed. To achieve this, we used machine learning techniques, including linear regression, decision tree, random forest, and artificial neural networks or multilayer perceptron. A deep learning technique such as DNN is used in order to extend this study and obtain more precise results. The prediction performance on the dataset is studied in detail on the five ML and DL algorithms, where DNN showed the highest accuracy. During the predictive analysis, models like LR, DT, RF, MLP, and DNN managed to find the hidden relationship between CQI_Avg and DL_perceived_throughput. With these five models, CQI_Avg was able to closely capture the variation in perceived flow. The perceived speed of 4G downlink is generally perceived to be related to channel quality or load data recording. However, analysis of a mobile dataset from Orange Senegal's 4G network using machine learning and deep learning techniques revealed that CQI_Avg was the variable that most affected perceived network throughput. 4G. Such a relationship cannot be seen by a correlation analysis. Using machine learning and deep learning based on predictive analytics, we discovered a very strong relationship between CQI_Avg and perceived downlink throughput. This result may have important implications for the implementation of better very high speed mobile networks such as the fifth generation (5G). This work could be extended in the future to predict packet loss in very high-speed mobile networks made up of small cells. A detailed study on the state of the transmission channel in ultra-dense networks made up of small cells is also interesting.

Declarations

Availability of data and material The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. Nevertheless, we have presented an overview of this analyzed dataset in “Fig. 13”.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Y. Liu and J. Y. B. Lee, “An Empirical Study of Throughput Prediction in Mobile Data Networks,” 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, 2015, pp. 1–6, doi: 10.1109/GLOCOM.2015.7417858
2. A. Masaracchia, R. Bruno, A. Passarella and S. Mangione, "Analysis of MAC-level throughput in LTE systems with link rate adaptation and HARQ protocols," 2015 *IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2015, pp. 1–9
3. Baena, E., Fortes, S. & Barco, R. Assessing the impact of DRS signaling in unlicensed indoor coexistence scenarios. J Wireless Com Network 2020, 224 (2020)
4. C. Xu, J. Wang, T. Yu, C. Kong, Y. Huangfu, R. Li, Y. Ge, andJ. Wang, “Buffer-aware wireless scheduling based on deep reinforcementlearning,” in2020 IEEE Wireless Communications and NetworkingConference (WCNC), pp. 1–6, IEEE, 2020
5. YUAN, Mingqi, CAO, Qi, PUN, Man-on, *et al.* Fairness-Oriented User Scheduling for Bursty Downlink Transmission Using Multi-Agent Reinforcement Learning. *arXiv preprint arXiv*:2012.*15081*, 2020. https://doi.org/10.48550/arXiv.2012.15081
6. M. A. Khan, R. Hamila, N. A. Al-Emadi, M. S. Kiranyaz and M. Gabbouj, “Realtime throughput prediction for cognitive Wi-Fi networks,” Journal of Network and Computer Applications (2019), doi: https://doi.org/10.1016/j.jnca.2019.102499
7. S. Abolfazli, Z. Sanaei, S. Y. Wong, A. Tabassi and S. Rosen, “Throughput measurement in 4G wireless data networks: Performance evaluation and validation,” Computer Applications & Industrial Electronics (ISCAIE), 2015 IEEE Symposium on. 2015, doi: 10.1109/ISCAIE.2015.7298322
8. D. Ndolane, N. Massa, D. Dialo, T. Kharouna, B. C. Aboubaker and G. Ibrahima, "Finding Hidden Links among Variables in a Large-Scale 4G Mobile Traffic Network Dataset Using Machine Learning," 2021 *8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2021, pp. 1–8, doi: 10.1109/ISCMI53840.2021.9654806
9. S. Yin, D. Chen, Q. Zhang and S. Li, “Prediction-Based Throughput Optimization for Dynamic Spectrum Access,” in IEEE Transactions on Vehicular Technology, vol. 60, no. 3, pp. 1284–1289, March 2011, doi: 10.1109/TVT.2010.2101090

10. Nihal H. Mohammed, H. Nashaat, Salah M. Abdel-Mageid and Rawia Y. Rizk, “A Machine Learning-Based Framework for Efficient LTE Downlink Throughput.” Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications. Springer, Cham, 2021. 193–218
11. L. Xu et al. “Supervised Machine Learning in 5G NSA Networks for Coverage and Performance Analysis,” Signal and Information Processing, Networking and Computers. Springer, Singapore, 2021. 910–916. DOI: 10.1007/978-981-33-4102-9_109
12. T. Grace Shalini, S. Jenicka, “Weighted Greedy Approach for Low Latency Resource Allocation on V2X Network,” Wireless Pers Commun (2021), doi: 10.1007/s11277-021-08332-3
13. A. Kulkarni, A. Seetharam, A. Ramesh and J. D. Herath, “DeepChannel: Wireless Channel Quality Prediction Using Deep Learning” in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 443–456, Jan. 2020, doi: 10.1109/TVT.2019.2949954
14. A. Thantharate, R. Paropkari, V. Walunj and C. Beard, "DeepSlice: A Deep Learning Approach towards an Efficient and Reliable Network Slicing in 5G Networks," 2019 *IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2019, pp. 0762–0767, doi: 10.1109/UEMCON47517.2019.8993066
15. Anwar, Muhammad Zohaib, Zeeshan Kaleem, and Abbas Jamalipour. "Machine learning inspired sound-based amateur drone detection for public safety applications." IEEE Transactions on Vehicular Technology 68.3 (2019): 2526–2534, doi: 10.1109/TVT.2019.2893615
16. Khan, Muhammad Asif, et al., “Real-time throughput prediction for cognitive Wi-Fi networks,” Journal of Network and Computer Applications 150 (2020): 102499, doi: 10.1016/j.jnca.2019.102499
17. Kousias, Konstantinos, et al., “Estimating downlink throughput from end-user measurements in mobile broadband networks,” 2019 IEEE 20th International Symposium on,” A World of Wireless, Mobile and Multimedia Networks"(WoWMoM). IEEE, 2019, doi: 10.1109/WoWMoM.2019.8792968
18. Liu, Yan, and Jack YB Lee, “An empirical study of throughput prediction in mobile data networks,” 2015 IEEE Global Communications Conference (GLOBECOM). IEEE, 2015, doi: 10.1109/GLOCOM.2015.7417858
19. Chaudhry, Aizaz U., and HM Roshdy Hafez, “On Finding Hidden Relationship among Variables in WiFi using Machine Learning,” 2020 International Conference on Computing, Networking and Communications (ICNC). IEEE, 2020, doi: 10.1109/ICNC47757.2020.9049741
20. RONG, Shen et BAO-WEN, Zhang. The research of regression model in machine learning field. In: MATEC Web of Conferences. EDP Sciences, 2018. p. 01033. https://doi.org/10.1051/matecconf/201817601033
21. HAO, Jiangang et HO, Tin Kam. Machine learning made easy: A review of scikit-learn package in Python programming language. Journal of Educational and Behavioral Statistics, 2019, vol. 44, no 3, p. 348–361. https://doi.org/10.3102/1076998619832248
22. TIMOFEEV, Roman. Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, 2004, vol. 54.

23. A. Hajjem, F. Bellavance and D. Larocque, “Mixed-effects random forest for clustered data”, Journal of Statistical Computation and Simulation (2014), 84:6, 1313–1328, DOI: 10.1080/00949655.2012.741599
24. Bélanger, M., El-Jabi, N., Caissie, D., Ashkar, F. & Ribi, J. M. (2005). Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. Revue des sciences de l'eau / Journal of Water Science, 18(3), 403–421. https://doi.org/10.7202/705565ar
25. CHITSAZ, Hamed, SHAKER, Hamid, ZAREIPOUR, Hamidreza, *et al.* Short-term electricity load forecasting of buildings in microgrids. Energy and Buildings, 2015, vol. 99, p. 50–60. https://doi.org/10.1016/j.enbuild.2015.04.011
26. MEI, Lifan, GOU, Jinrui, CAI, Yujin, *et al.* Realtime mobile bandwidth and handoff predictions in 4G/5G networks. Computer Networks, 2022, p. 108736. https://doi.org/10.1016/j.comnet.2021.108736
27. Moon, S., Kim, H., You, YH. *et coll.* Réseau de neurones profonds pour la prédiction des faisceaux et des blocages dans les environnements de points d'accès intérieurs basés sur 3GPP. Pers Commun sans fil (2022). https://doi.org/10.1007/s11277-022-09513-4
28. Zhen Tang, Xiaobin Fu, Pingping Xiao, "Mobile Performance Intelligent Evaluation of IoT Networks Based on DNN", *International Journal of Antennas and Propagation*, vol. 2022,ArticleID 4038830, 7 pages, 2022. https://doi.org/10.1155/2022/4038830
29. NGUYEN, Chi et CHEEMA, Adnan Ahmad. A deep neural network-based multi-frequency path loss prediction model from 0.8 GHz to 70 GHz. Sensors, 2021, vol. 21, no 15, p. 5100. https://doi.org/10.3390/s21155100
30. PURUSHOTHAMAN, K. E. et NAGARAJAN, V. Evolutionary multi-objective optimization algorithm for resource allocation using deep neural network in 5G multi-user massive MIMO. International Journal of Electronics, 2021, vol. 108, no 7, p. 1214–1233. https://doi.org/10.1080/00207217.2020.1843715
31. L. A. Garrido, P. -V. Mekikis, A. Dalgkitsis and C. Verikoukis, "Context-Aware Traffic Prediction: Loss Function Formulation for Predicting Traffic in 5G Networks," ICC 2021 - IEEE International Conference on Communications, 2021, pp. 1–6, doi: 10.1109/ICC42927.2021.9500735
32. ABDELLAH, Ali R., ALSHAHRANI, Abdullah, MUTHANNA, Ammar, *et al.* Performance Estimation in V2X Networks Using Deep Learning-Based M-Estimator Loss Functions in the Presence of Outliers. Symmetry, 2021, vol. 13, no 11, p. 2207. https://doi.org/10.3390/sym13112207
33. A. Al-Tahmeesschi, K. Umebayashi, H. Iwata, J. Lehtomäki and M. López-Benítez, "Feature-Based Deep Neural Networks for Short-Term Prediction of WiFi Channel Occupancy Rate," in IEEE Access, vol. 9, pp. 85645–85660, 2021, doi: 10.1109/ACCESS.2021.3088423
34. J. Thrane, D. Zibar and H. L. Christiansen, "Model-Aided Deep Learning Method for Path Loss Prediction in Mobile Communication Systems at 2.6 GHz," in IEEE Access, vol. 8, pp. 7925–7936, 2020, doi: 10.1109/ACCESS.2020.2964103
35. TIMOFEEV, Roman. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin, 2004, p. 1–40.

36. LIAW, Andy, WIENER, Matthew, et al. Classification and regression by randomForest. R news, 2002, vol. 2, no 3, p. 18–22
37. K. I. Ahmed, H. Tabassum and E. Hossain, "Deep Learning for Radio Resource Allocation in Multi-Cell Networks," in IEEE Network, vol. 33, no. 6, pp. 188–195, Nov.-Dec. 2019, doi: 10.1109/MNET.2019.1900029

Figures

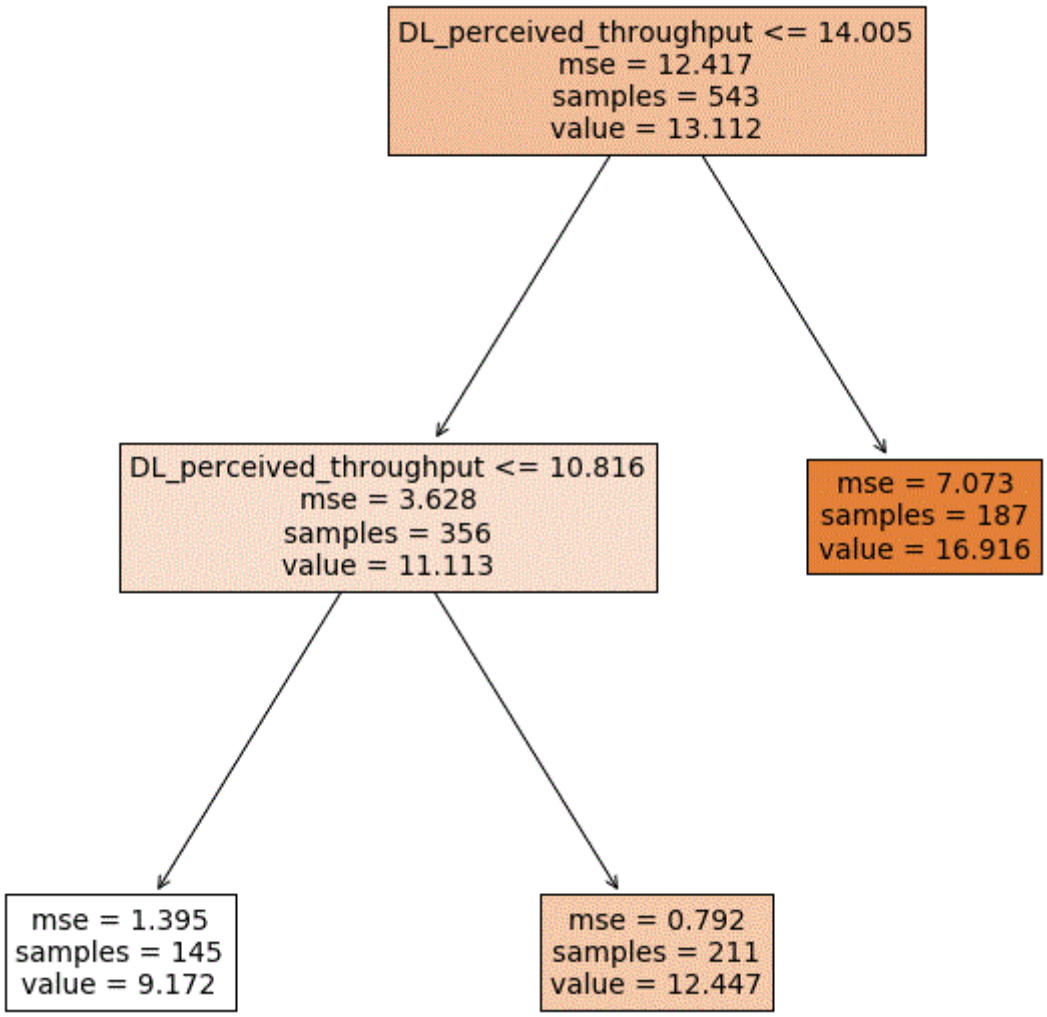

Figure 1

Decision Tree

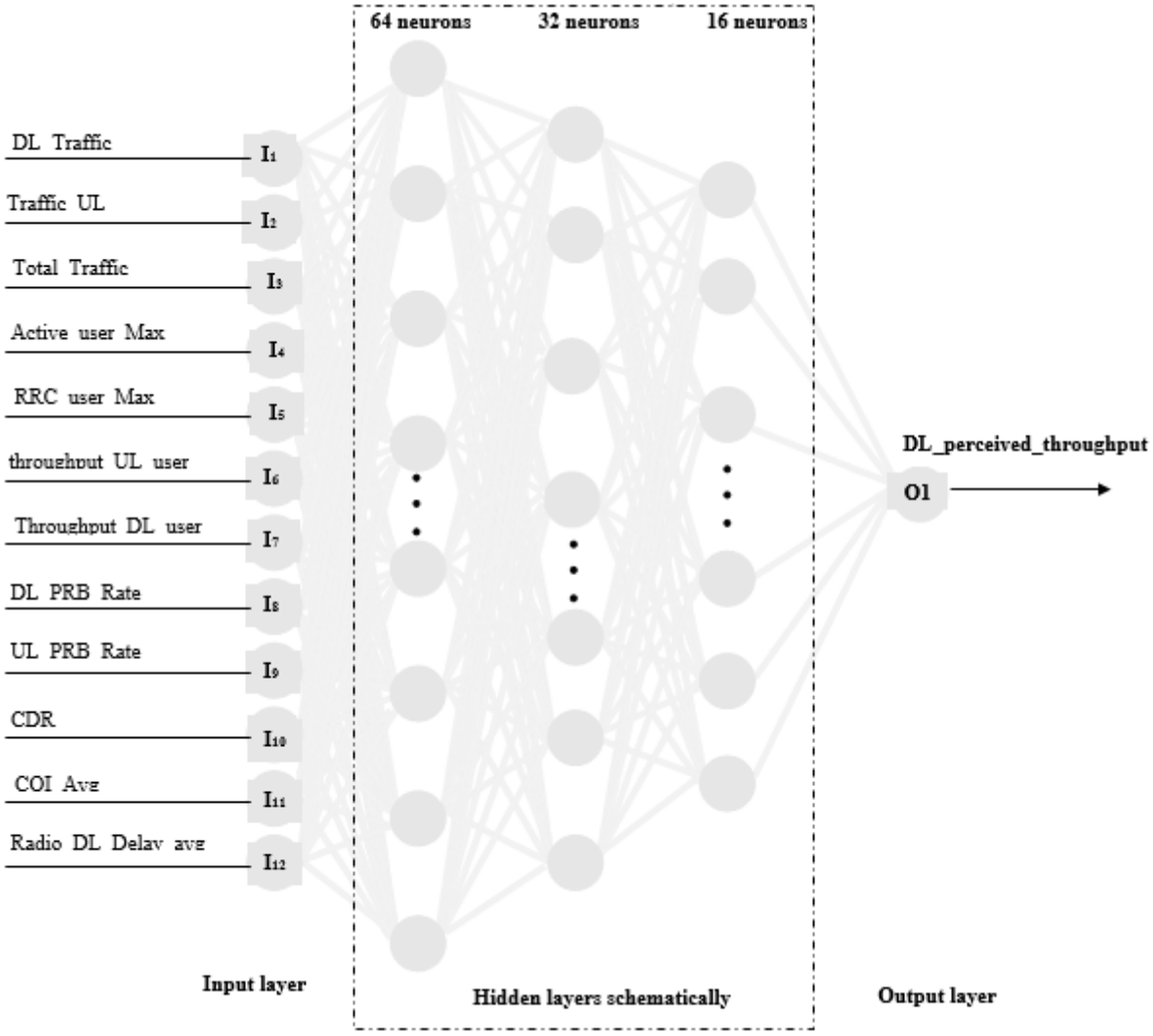

Figure 2

Typical DNN model

Figure 3

Distribution of training and testing datasets

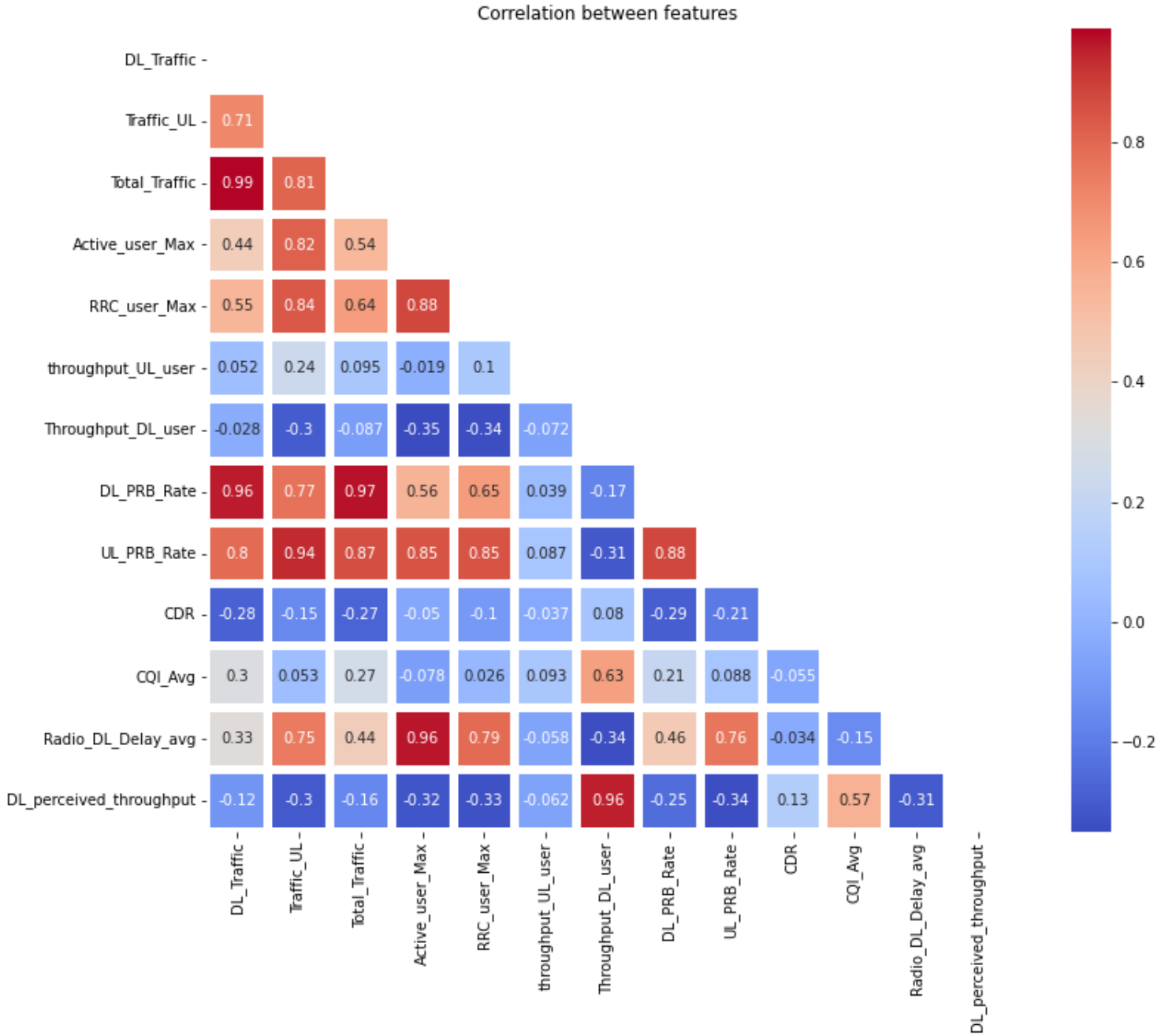

Figure 4

Correlation between data features

Figure 5

Actual VS. predicted values of throughput

Figure 6

Prediction performance for each model

Figure 7

Evolution of training and validation losses using the MAE metric

Figure 8

Prediction performance for DNN model

Figure 9

Actual VS. predicted values of throughput

Figure 10

Prediction accuracy of DNN, RF, LR, DT and MLP models

Figure 11

Performance measurement using MAE

Figure 12

Performance measurement using RMSE

Figure 13

Distributing the dataset after preprocessing.