Forecasting Wireless Network Traffic and Channel Utilization Using Real Network/Physical layer Data

Su Pyae Sone*, Janne Lehtomäki*, Zaheer Khan* and Kenta Umebayashi†

*Centre for Wireless Communications (CWC), University of Oulu, Finland.
†Tokyo University of Agriculture and Technology, Japan.

***Abstract*—Prediction of wireless network parameters, such as traffic (TU) and channel utilization (CU) data, can help in proactive resource allocation to handle the increasing amount of devices in an enterprise network. In this work, we examined the medium-to-long-scale forecasting of TU and CU data collected from an enterprise network using classical methods, such as Holt-Winters, Seasonal ARIMA (SARIMA), and machine learning methods, such as long short-term memory (LSTM) and gated recurrent unit (GRU). We also improved the performance of conventional LSTM and GRU for time series forecasting by proposing features-like grid training data structure which uses older historical data as features. The wireless network time series pre-processing methods and the verification methods are presented as time series analysis steps. The model hyper-parameters selections process and the comparison of different forecasting models are also provided. This work has proven that physical layer data has more predictive power in time series forecasting aspect with all forecasting models.**

Keywords*—Forecasting, GRU, Holt-Winters, LSTM, Neural Networks, Real Network Data, SARIMA, Time Series, WLAN.

I. Introduction

Next generation wireless networks including fifth generation (5G) and beyond networks are expected to provide not only higher data rate but also ultra-reliable low latency communications (URLLC) for billions of devices since they will play the main role in the growth of Internet of Everything (IoE). Network data analytic function (NWDAF) which can perform data analysis, forecasting and executing decisions for network planning has been introduced by 3rd Generation Partnership Project (3GPP) [1]. Data analysis and prediction is important in data driven centralized frameworks. Hence, analyzed data with a good prediction about how much network parameters will be utilized at an access point (AP) or a group of APs can help in effectively allocating appropriate resources at those APs. Consequently, not only cellular networks but also enterprise networks will be extended to provide the connections for enormous amount of devices to incorporate IoE [2].

Hence, data driven automated frameworks for data analysis and good predictions of network parameters to make decisions on network performances are also required for enterprise networks. Wireless network parameters collected from an enterprise network such as traffic utilization indicating the data rate, number of users connected and channel utilization indicating the channel occupancy percentage, can be modeled as time series and many mechanisms have already been designed for time series analysis and prediction [3]. Various time series forecasting methods can be divided into two major groups: classical methods including exponential smoothing (ES) and auto-regressive integrated moving average (ARIMA) [4], and machine learning methods including support vector machine (SVR) and neural network based (LSTM) [5]. The gated recurrent unit (GRU) method is famous alternative to LSTM with faster operation and simpler structure [6]. Both classical and machine learning methods have their own advantages such as classical methods are better for the series with strong seasonality and trend [7]. Machine learning methods have the ability to handle non-linear patterns and rapid changes by extracting the features of the historical data [8].

According to the various results of researches from past decade, traffic time series patterns can be vary for different networks and there is no prediction method which can be fit all types of network time series patterns. There is abundant literature on detailed analyzing and forecasting network traffic data of cellular networks [9], [10]. The channels utilized in enterprise networks are shared channels operating with IEEE 802.11 wireless network standards which are different from the cellular networks. Therefore, separate forecasting studies are required for enterprise networks. So far only a few researches presented results for the enterprise networks using classical methods such as [11]. We did correlation-based detailed analysis and forecasting using both classical and machine learning methods for traffic time series data of a real enterprise network in [12]. In fact, a channel in an enterprise network is shared among multiple wireless devices. Hence, the traffic occupied on the physical layer channel can be more than the network layer traffic of a certain AP or a group of APs. Having low traffic usage at the APs with high channel utilization becomes the problem in proactive allocation by forecasting network layer traffic data. In [12], physical layer data was not considered.

Analyzing and forecasting physical layer data, such as channel usage time series, to help in allocating appropriate resources at those APs with shared channel can be a potential solution in the enterprise networks. Short-scale estimation of the channel duty cycle using neural network based forecasting methods was performed in [13] and capacity utilization prediction in point-to-point microwave communication links using ARIMA as well as neural network based forecasting method was done [14]. Moreover, it will be useful in network management to perform medium-to-long-scale network parameters analysis and forecasting for both network layer traffic and physical layer channel utilization time series of a real enterprise network. It is also important to investigate which layer data is more predictive in time series forecasting aspect to help in proactively allocating appropriate resources at a certain AP or a group of APs.

Therefore, we attempt to do time series analysis and forecasting with classical and machine learning methods for both physical layer data and network layer data of a real enterprise network. The main contributions of our work include:

- We evaluate forecasting performances of analyzed time series for both traffic utilization (TU) and channel utilization (CU) data using 4 methods: Holt-Winters, Seasonal ARIMA (SARIMA), LSTM and GRU.
- We propose features-like grid (FLG) training data structure to improve the performances of machine learning methods for both medium and long-scale forecasting.
- We present challenges of network parameter time series forecasting of TU data of the APs deployed at the student lounge of University of Oulu and CU data of a channel which is shared among different devices including above APs.
- We compare forecasting performances of network layer data and physical layer data to show that physical layer data has more predictive power in time series forecasting aspect to help in proactively allocating appropriate resources at the APs.

II. Time Series Data Analysis and Forecasting

A. *Description of collected data*

We collected both network and physical layer data from APs deployed around the Linnanmaa campus of the University of Oulu, Finland. The received and transmitted traffic data rate, number of users, locations and the names of each AP of a total 470 APs around the campus are collected as the network layer data. Each data point of total 5040 of the time series provides the measurement at every 10-minute interval within the period of January 5, 2019 to February 8, 2019. We defined the transmitted traffic time series of an AP i as $\mathbf{R}^{i}_{tx}$ and consider it (which dominates received traffic) as the network TU data. The collected TU dataset, which is named Wireless Network Traffic Time Series of an Enterprise Network, has open access.

Physical layer data CU indicates the percentage of the total amount of transmission from all kinds of sources including APs operating on the same specific channel within time period t. A data point of CU time series is defined as: $CU_t = \frac{1}{T}\sum_{j=1}^{T} D_j$, where D_j is the j^{th} binary decision of the signal presence or absence and T is the number of signal detection iteration. We collected CU of a channel with 20MHz bandwidth operating in 2.4GHz using three measuring devices placed in one of the locations of the University of Oulu where four APs are operating with high traffic transmissions. The devices were configured to collect mean CU (CU-mean) and maximum CU (CU-max) values of the channel at every 20 seconds interval between February 11 to February 24, 2019. Hence, each CU time series has over 50,000 data points. The details of collected CU time series and the measurement devices we used can be found in [15].

B. *Time Series Analysis System*

Analyzing and forecasting TU and CU of an enterprise network within medium-to-long time periods, such as one hour (1-hr) and one day (1-day) ahead, will help to make predictions about the behavior of the network and channel to effectively allocate the appropriate resources in the network. First we defined the forecast horizon (FH) which is the number of future time periods for which forecasts must be produced. The training dataset with the total number of samples T is divided equally into the pairs of input sample and target. The first pair of input sample and target can be defined as $\mathbf{X}_1 = \{x_1, x_2, \ldots, x_{FH}\}$ and $\mathbf{Y}_1 = \{y_1, y_2, \ldots, y_{FH}\}$, for instance, if x_i is the current traffic usage, y_i will be the traffic usage of next hour for 1-hr FH. The testing dataset is also divided equally where the first testing sample and forecasted values are defined as $\hat{\mathbf{X}}_1 = \{x_{T+1}, x_{T+2}, \ldots, x_{T+FH}\}$ and $\mathbf{P}_1 = \{p_{T+1}, p_{T+2}, \ldots, p_{T+FH}\}$ to evaluate the forecasting performance by comparing them.

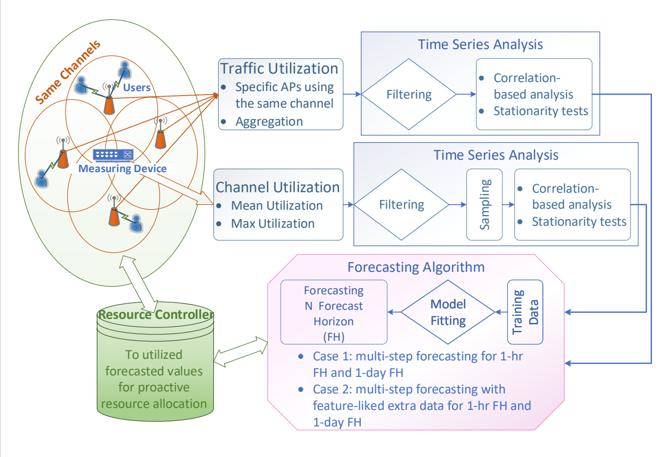

Same Channels
Users
Measuring Device
Resource Controller
To utilized forecasted values for proactive resource allocation
Traffic Utilization
Specific APs using the same channel
Aggregation
Time Series Analysis
Filtering
Correlation-based analysis
Stationarity tests
Channel Utilization
Mean Utilization
Max Utilization
Time Series Analysis
Filtering
Sampling
Correlation-based analysis
Stationarity tests
Forecasting Algorithm
Forecasting N Forecast Horizon (FH)
Model Fitting
Training Data
Case 1: multi-step forecasting for 1-hr FH and 1-day FH
Case 2: multi-step forecasting with feature-liked extra data for 1-hr FH and 1-day FH

Fig. 1. An illustration of traffic and channel utilization forecasting system

Before forecasting, we need to perform time series analysis of the data which in turn can improve the forecasting performance. As both TU and CU of typical enterprise network are very low during weekends, we focus our investigation to weekdays (working days) data of both TU and CU time series. The illustration of our system can be seen in Fig. 1. For TU data series, we only selected the APs near to a specific location with shared channel that we collected CU data from. We already observed that aggregating traffic series of numerous APs makes the resultant time series more stable [12]. Consequently, the traffic time series of K APs with shared channel at a specific location are aggregated into a total traffic time series, defined as $\mathbf{R}_{TU} = \mathbf{R}^{1}_{tx} + \mathbf{R}^{2}_{tx} + \ldots + \mathbf{R}^{K}_{tx}$, to utilize the benefits of resulting less variability.

In general, network parameter time series collected from any network exhibit non-stationarity and their statistical properties change over time [11]. However, a specific filter is applied to the traffic series of wireless home network to get the similar properties of stationary series in [16]. Therefore, we also try applying median filter to both our collected TU and CU data series as a process of time series smoothing. Median filter is widely used for its ability to keep the important pattern of the time series preserving the edges by distinguishing the outliers [17]. Since we would like to perform hourly forecasting, median filter with hourly window length is used to smooth out both aggregated TU and CU time series in this work. Then, we down sampled the collected CU-mean and CU-max time

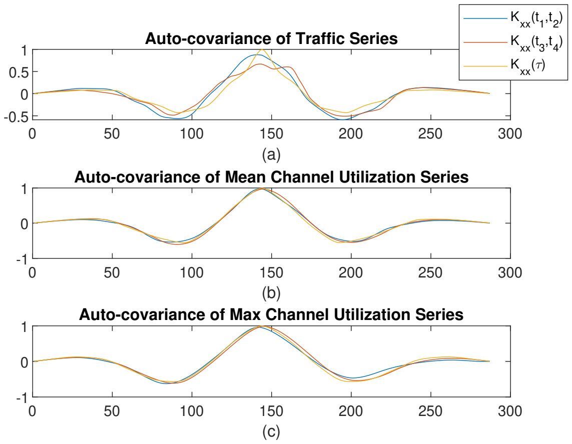


Fig. 2. Auto-covariance of non-overlapping sliding with 24-hour windows

series so that each data point provides the measurement of every 10-min period as in aggregated TU series, since CU data is collected for two weeks of every 20 second. Sampling is not required for TU data.

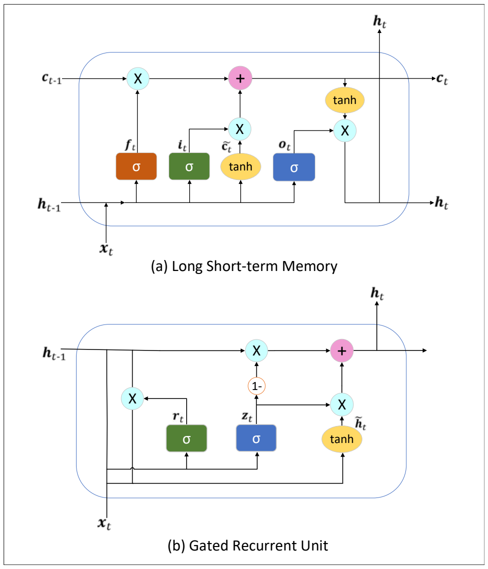
As the last steps for time series analysis, the correlation-based analysis of the collected time series and testing consistency of statistical properties for different types of data series are followed. Investigation of the auto-correlation results before and after applying the medium filter with various hourly window lengths tells that 1-hour and 2-hour window lengths provides the decent increments in the correlation of each aggregated TU and CU time series. After aggregation, filtering and correlation-based analysis, it is time to test stationarity of the analyzed time series since the behavior and properties of the time series strongly influence the forecasting performances. Unit root tests such as Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are widely used as the stationary tests. However, absence of unit root in a time series does not grantee for stationarity [18].

Hence, we used simple and guaranteed method called weak stationarity test by checking constant mean, variance over time and co-variance between two time periods to depend only on the time difference but not the actual time at which the co-variance is computed. For both aggregated TU and CU time series do not have constant mean and variance over time for any time windows. Nevertheless, CU-max time series appears to have flatter mean and variance changes over time. The covariance of different time periods such as t_1 to t_2, t_3 to t_4 and their time difference τ are also not the same so that none of the time series satisfy the weak stationarity properties. However, Fig. 2 shows that properties of CU time series are more similar to stationary time series than aggregated TU time series. After time series analysis, the prepared time series to be able to enhance the forecasting performances are ready to pass through the forecasting algorithms.

C. Forecasting Methods and Performances

1) Holt-Winters and SARIMA:

As network time series data usually exhibit seasonal patterns, one approach that is available for the analysis of such



(a) Long Short-term Memory

(b) Gated Recurrent Unit

Fig. 3. The block structures of LSTM and GRU networks

data is Holt-Winters which is also known as triple exponential smoothing method. It is used to forecast the time series algorithm by assigning exponentially decreasing weights and values on the historical data. Based on seasonality, Holt-Winters has two type of models, additive model and multiplicative model. However, we used only additive model as a research [19] stated that additive model is more suitable for wireless mobile traffic data.

ARIMA, is widely used for univariate time series data forecasting. However, it is not suitable for our work as it does not support time series with a seasonal component. Seasonal ARIMA (SARIMA) which is an extension of ARIMA supports the direct modeling of the seasonal component of a time series and is therefore perfect to use for forecasting network data usage. This method has AR term, MA term and integrated (I) term to fit the seasonal data as well as possible. SARIMA can be expressed as ARIMA$(p,d,q)(P,D,Q)s$, where p is the number of AR terms, d is the number of difference, q is the number of MA terms while the capital P,D,Q are for seasonal terms respectively, and s is seasonal period of time series. SARIMA model can be trained to fit the data by adjusting the above parameters [20].

2) LSTM and GRU:

LSTM is a variation of recurrent neural networks (RNNs). It is one of the most powerful tools in time series forecasting and hence makes a strong case for using it in our work. LSTM can remember the long time information as in RNN and can also even delete unnecessary information from its memory. One LSTM block can be considered as a cell with 3 main regulation structures to control the amount of information flow which are called Input gate $(\mathbf{i}_t)$, Forget gate $(\mathbf{f}_t)$ and Output gate $(\mathbf{o}_t)$ [21]. LSTM computes not only the states of the gates but also the Candidate cell state $(\tilde{\mathbf{c}}_t)$, the Current cell state $(\mathbf{c}_t)$ and the Final output $(\mathbf{h}_t)$. The activation function used in the gates of a cell in LSTM are sigmoid activation functions and the ones used in calculating cell states are tanh activation functions. The block structure can be seen in Fig. 3(a).

TABLE I. Considered hyper-parameters for LSTM and GRU

Hyper-parameter	Considered values
No. of layers (l)	1, 2, 3
No. of neurons (nn)	32, 64
Dropout (d_p)	0.3, 0.5, 0.8
Learning rate	0.001
Losses	MAE
Optimizer	Adam
Epochs	50

GRU also handles the information as in LSTM but without calculating the cell states. GRU structures consists of two gates called Reset gate ($\mathbf{r}_t$) and Update gate ($\mathbf{z}_t$) with sigmoid activation functions. The computation of Current memory ($\tilde{\mathbf{h}}_t$) and the Final output ($\mathbf{h}_t$) with tanh activation function are followed after the gating operations. The benefit is that GRU has less gating units resulting in faster operation than LSTM and the forecasting performance difference between LSTM and GRU can be insignificant [6]. The GRU is shown in Fig. 3(b).

D. Hyper-parameter selections of forecasting methods

Optimizing the hyper-parameters of both classical and machine learning methods is important to achieve the optimal forecasting performance. In Holt-Winters algorithm, the initial values such as l_0, b_0 and s_0 are computed based on the input time series as in [22]. The level, trend and seasonal smoothing factors are also optimized to fit the training samples by solving the minimum of constrained nonlinear multi-variable function. The optimal level, trend and seasonal smoothing factors used in our work are $\alpha = 0.004$, $\beta = 0$, $\gamma = 0.23$ for TU series and $\alpha = 0.001$, $\beta = 0$, $\gamma = 0.75$ for CU series, respectively. Moreover, there are various ARIMA model to choose for a certain time series. In fact, $\text{ARIMA}(0,1,1)(0,1,1)s$ is the most widely used model for seasonal time series with repetitive patterns [22]. According to the above correlation-based results, s and the seasonal MA terms Q are assigned as 144 to get the optimal performance for both TU and CU time series in this work.

In LSTM and GRU, which are the neural network-based machine learning methods, the hyper-parameters such as no. of layers, no. of neurons, dropout value and activation functions influence the performance of the forecasting model. The possible combinations of considered hyper-parameters presented in Table I are run through to find the optimal ones. Mostly, more complex features of the input can be extracted with deep and narrow neural networks than the shallow and wide ones [23]. However, as stated in [24] that applying deep or wide network for time series does not always improve the performance, the optimal hyper-parameters of LSTM for both TU and CU series are 2-layer each with 32 neurons and dropout value 0.3, where value 1 means no dropout is applied. Then, the optimized LSTM model is followed by a dense layer with linear activation function. Moreover, the batch size is set to 1 since [25] convinced that the method called on-line learning, which is with batch size 1, is good for pattern recognition problems and is faster than batch training. As an example, the performance comparison of forecasting CU-max time series with 1-hour medium filter length using LSTM models for different hyper-parameters can be seen in Table II.

TABLE II. Performance comparison of LSTM with different hyper-parameters for CU-max series in 1-hr FH (l is the number layers, nn is the number of neurons, and d_p is the dropout)

Hyper-parameters	d_p = 0.3	d_p = 0.5	d_p = 0.8
l = 1, nn = 32	0.8500	0.8482	0.8462
l = 2, nn = 32	**0.8633**	0.8626	0.8330
l = 3, nn = 32	0.8220	0.8365	0.8057
l = 1, nn = 64	0.8493	0.8473	0.8435
l = 2, nn = 64	0.8631	0.8623	0.8479
l = 3, nn = 64	0.8630	0.8522	0.8350

E. Network time series data forecasting

Time series forecasting is the process of extracting useful information from historical data to determine possible future data values. Different from classical methods, the main advantage of neural network based methods is that they can exploit the features or extra information to improve the forecasting performance [26]. In conventional LSTM and GRU for time series forecasting, only historical data sequence is given as input data if there is no extra feature available. On the other hand, it will require extra time and resources to compute extra features of all time series in the whole enterprise network.

Therefore, we proposed simple yet effective training data structure, which is named features-like grid (FLG), to improve the performance of neural network based methods. In proposed training data structure, 6 consecutive data points (as 1-hr period) are used as features instead of computing extra features. The first training sample with FLG structure is defined as $\mathbf{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{FH}\}$ where $\mathbf{x}_i = \{x_i, x_{i+1}, ..., x_{i+5}\}$ and the improvement of forecasting performance is also proved in this work. For all neural network based models, multi-step forecasting is used to learn the complex dependent structures of inputs and outputs of the model and to predict the multi-points ahead at once [27]. Since each data point provides the value at every 10-min, an input training sample is in the shape of 6x6 grid for 1-hr FH and 6x144 grid for 1-day FH. To make sure that samples and targets of the training data are not overlapped, we used 2 x FH separation period between samples and targets while preparing the training data.

The data collection days of TU series are longer than CU series. Therefore, we used the last 9 weeks of TU data to have the same amount of days as in CU data and represented as TU-R. We also investigated the impact of different training data size on forecasting performance of TU series by using the whole collected TU data and represented as TU-all. For CU series, we evaluated forecasting performance of both CU-mean and CU-max time series. As the performance evaluation metric, we used R Squared (R^2) normalized standard metric, which is scaled between 0 and 1 with an intercept, since the common metrics such as root mean squared error (RMSE) and normalized RMSE varied with different time series without any specific range [28]. In addition, only averaged accuracy of 10 times evaluation are presented in this work by considering the stochastic nature of neural network where results can be different for each prediction with the same model.

For 1-hr FH, the forecasting performances of considered time series with 1-hour median filter length for different models are presented in Table III. Among classical methods and conventional LSTM and GRU for time series models, SARIMA gave the decent accuracy for all time series. The

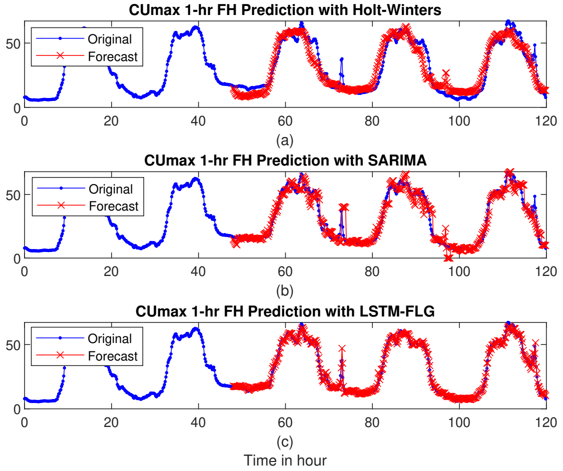


Fig. 4. Forecasting performances for CU-max time series with 1 hr FH

TABLE III. R^2 SCORE ACCURACY COMPARISON OF DIFFERENT METHODS FOR ALL TIME SERIES IN 1-HR FH

1-hr FH	HW	SARIMA	LSTM	GRU	LSTM-FLG	GRU-FLG
TU-all	0.8290	0.8852	0.7100	0.7114	0.9596	0.9644
TU-R	0.7328	0.8645	0.7145	0.7132	0.9669	0.9645
CU-mean	0.8323	0.8935	0.8276	0.8274	0.9896	0.9893
CU-max	0.8708	0.9213	0.8633	0.8542	0.9901	0.9879

accuracy are improved by using LSTM-FLG and GRU-FLG in all cases. Mostly, LSTM-FLG has the overall better accuracy than GRU-FLG for 1-hr FH. However, the performance difference between LSTM-FLG and GRU-FLG is not significant while GRU-FLG has the advantage of simpler and faster operation than in LSTM-FLG. According to given results in Table III, both CU series have better accuracy than TU series and CU-max series has the highest accuracy with LSTM-FLG for medium-scale prediction. The performance comparison of Holt-Winters, SARIMA and LSTM-FLG for CU-max series with 1-hr FH is presented in Fig. 4.

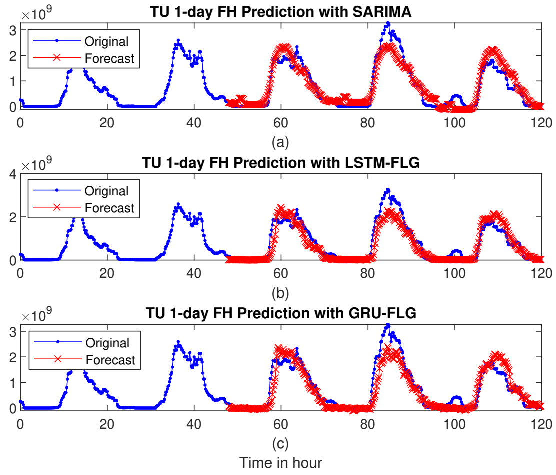


Fig. 5. Forecasting performances for aggregated TU-all time series with 1 day FH

TABLE IV. R^2 SCORE ACCURACY COMPARISON OF DIFFERENT METHODS FOR ALL TIME SERIES IN 1-DAY FH

1-day FH	HW	SARIMA	LSTM	GRU	LSTM-FLG	GRU-FLG
TU-all	0.8194	0.8574	0.7654	0.8513	0.8145	0.8169
TU-R	0.8424	0.8449	0.7557	0.7431	0.7991	0.8005
CU-mean	0.8653	0.8641	0.7766	0.7743	0.7908	0.8213
CU-max	0.9017	0.8996	0.8889	0.8984	0.9159	0.8992

For 1-day FH, 2-hour median filter length is used for all considered time series to enhance the long-scale performance. As shown in Table IV, classical methods gave better accuracy than conventional LSTM and GRU in all cases. Although LSTM-FLG and GRU-FLG improve the performances of conventional LSTM and GRU for time series in all cases, they can not outperform classical methods in long-scale prediction of TU-all, TU-R and CU-mean series. SARIMA gave the highest accuracy for TU series and Holt-Winters is best for CU-mean series. The performance differences between SARIMA, LSTM-FLG and GRU-FLG for TU-all series with 1-day FH are shown in Fig. 5. Mostly, GRU-FLG has the overall better accuracy than LSTM-FLG for 1-day FH. However, LSTM-FLG outperformed classical methods and gave the highest accuracy for CU-max series. Same as in 1-hr FH prediction, CU-max series has the overall highest accuracy with LSTM-FLG for long-scale prediction.

III. Challenges in Forecasting TU and CU series

Time series forecasting itself has its own challenges but the challenges are more specific for a particular wireless network parameter time series. The important factors and challenges to consider before forecasting wireless network parameter time series include training data size, training frequency, multiple seasonality, irregular changes and data processing. In Section II-E, we presented the impact of different training data size on forecasting performance by using two training data size for TU series. Results for 1-hr FH in Table III shows that accuracy of classical methods reduced along with the training data size while LSTM-FLG and GRU-FLG did improve the accuracy for TU-R when training data is reduced. However, overall accuracy for TU-R is reduced in both classical and machine learning models for 1-day FH as can be seen in Table IV. Hence, it is impossible to decide how much is enough training data size to train models in wireless network parameters forecasting.

Wireless network parameters are time-varying and non-stationary so that their statistical properties are also changing with time. The forecasting models need to be retrained with a certain frequency for a time period and how to set the optimal frequency is the interesting question. Moreover, most wireless network parameter of an enterprise network have strong multiple seasonal patterns. Multiple seasonality brings extra complexities for the forecasting models and proper handling methods are still an open issue. Some robust and strong forecasting models can predict the irregular changes in the network parameters time series up to some extent for short and medium scale forecasting. However, to predict the sudden changes in the network for long-scale forecasting is almost impossible as shown in Fig. 5. Pre-processing the data before forecasting is necessary for all models to enhance the accuracy by aggregating, filtering and so on. And, the utility of these pre-processed methods can be verified by doing correlation-based analysis and stationary test. However, all these steps require selection of right parameters and determining the amount of increased utility is also one of the challenging tasks.

IV. Conclusion

To manage the increasing amount of devices in an enterprise network, predictions of TU and CU can help in proactive resource allocation. Therefore, we examined the forecasting performances of TU data with different training data amounts, CU-mean data and CU-max data of an enterprise network. We investigated for 1-hr FH (medium-scale) and 1-day FH (long-scale) predictions using Holt-Winters, SARIMA, the conventional LSTM and GRU as well as the proposed LSTM-FLG and GRU-FLG. The time series processing such as aggregating, filtering and sampling are done and the utility of these methods are verified with correlation-based analysis and stationary tests. It is shown that the proposed machine learning methods are suitable for medium-scale predictions since they can learn the complex relations of input and output data better than classical methods. For long-scale prediction, the classical methods performed well for most of the cases due to their insensitivity of outliers from the data set. The physical layer CU data, especially CU-max data, has the highest accuracy in all cases and it answers the question of which layer data is more predictive in time series forecasting aspect to help in proactively allocating appropriate resources. The challenges that encountered during the investigation of different forecasting models for all time series are also presented.

Acknowledgment

This work is supported by the Infotech Oulu, Academy of Finland 6Genesis Flagship (grant no. 318927), and the European Commission in the framework of the H2020-EUJ-02-2018 project 5GEnhance (Grant agreement no. 815056), by "Strategic Information and Communications R&D Promotion Programme (SCOPE)" of Ministry of Internal Affairs and Communications (MIC) of Japan (Grant no. JPJ000595).

References

[1] 3GPP, “Study of enablers for network automation for 5G (release 16),” *TS 23.791, Version 16.0.0*, 2018.

[2] “The internet of things for enterprises: An ecosystem, architecture, and iot service business model,” *Internet of Things*, vol. 7, 2019.

[3] G. Bontempi, S. B. Taieb, and Y. Le Borgne, “Machine learning strategies for time series forecasting,” in *European business intelligence summer school*, pp. 62–77, Springer, 2012.

[4] M. Pawlikowski and A. Chorowska, “Weighted ensemble of statistical models,” *International Journal of Forecasting*, January 2019.

[5] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, “Big data-driven optimization for mobile networks toward 5G,” *IEEE Network*, vol. 30, pp. 44–51, January 2016.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *NIPS Workshop on Deep Learning*, December 2014.

[7] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” *Public Library of Science (PloS) one*, vol. 13, March 2018.

[8] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, “An empirical comparison of machine learning models for time series forecasting,” *Econometric Reviews*, vol. 29, pp. 594–621, August 2010.

[9] C. Zhang and P. Patras, “Long-term mobile traffic forecasting using deep spatio-temporal neural networks,” in *Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 231–240, ACM, June 2018.

[10] C. W. Huang, C. T. Chiang, and Q. Li, “A study of deep learning networks on mobile traffic forecasting,” in *28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, October 2017.

[11] B. Krithikaivasan, Y. Zeng, K. Deka, and D. Medhi, “ARCH-based traffic forecasting and dynamic bandwidth provisioning for periodically measured nonstationary traffic,” *IEEE/ACM Transactions on Networking (TON)*, vol. 15, no. 3, pp. 683–696, 2007.

[12] S. P. Sone, J. J. Lehtomäki, and Z. Khan, “Wireless traffic usage forecasting using real enterprise network data: Analysis and methods,” *IEEE Open Journal of the Communications Society*, vol. 1, pp. 777–797, 2020.

[13] A. Al-Tahmeesschi, K. Umebayashi, H. Iwata, M. López-Benítez, and J. Lehtomäki, “Applying deep neural networks for duty cycle estimation,” in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–7, IEEE, 2020.

[14] A. Mahmood *et al.*, “Capacity and frequency optimization of wireless backhaul network using traffic forecasting,” *IEEE Access*, vol. 8, pp. 23264–23276, 2020.

[15] Z. Khan and J. J. Lehtomäki, “FPGA-assisted real-time RF wireless data analytics system: Design, implementation, and statistical analyses,” *IEEE Access*, vol. 8, pp. 4383–4396, 2020.

[16] K. Mirylenka, V. Christophides, T. Palpanas, I. Pefkianakis, and M. May, “Characterizing home device usage from wireless traffic time series,” *Nineteenth International Conference on Extending Database Technology (EDBT)*, June 2016.

[17] M. G. Kendall and A. Stuart, *The advanced theory of statistics*, vol. 3 of *Kendall’s Advanced Theory of Statistics*. John Wiley & Sons, 1945.

[18] J. Davidson, “Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes,” *Journal of Econometrics*, vol. 106, pp. 243–269, February 2002.

[19] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, “Mobile traffic forecasting for maximizing 5G network slicing resource utilization,” in *Conference on Computer Communications (INFOCOM)*, pp. 1–9, IEEE, May 2017.

[20] J. Shi, G. He, and X. Liu, “Anomaly detection for key performance indicators through machine learning,” in *International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, Aug. 2018.

[21] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, December 1997.

[22] S. Wheelwright, S. Makridakis, and R. J. Hyndman, *Forecasting: methods and applications*, vol. 2 of *Business forecasting*. John Wiley & Sons, 1998.

[23] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” in *International Conference on Learning Representations (ICLR)*, April 2014.

[24] Z. Cui, R. Ke, Z. Pu, and Y. Wang, “Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction,” *arXiv: Computer Science*, 2018.

[25] “The general inefficiency of batch training for gradient descent learning,” *Neural Networks*, vol. 16, no. 10, pp. 1429 – 1451, 2003.

[26] G. P. Zhang, B. E. Patuwo, and M. Y. Hu, “A simulation study of artificial neural networks for nonlinear time-series forecasting,” *Computers & Operations Research*, vol. 28, no. 4, pp. 381–396, 2001.

[27] J. Brownlee, *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018.

[28] S. P. Washington, M. G. Karlaftis, and F. Mannering, *Statistical and econometric methods for transportation data analysis*. Chapman and Hall, December 2010.

[29] N. Shamsi, A. Mousavinia, and H. Amirpour, “A channel state prediction for multi-secondary users in a cognitive radio based on neural network,” in *2013 International Conference on Electronics, Computer and Computation (ICECCO)*, pp. 200–203, 2013.

[30] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” *arXiv preprint arXiv:1312.6026*, 2013.