A Machine Learning approach to Forecast 5G Data in a Commercial and Operational 5G Platform*

Ana Almeida*†, Pedro Rito*, Susana Brás‡, Filipe Cabral Pinto§, Susana Sargento*†
*Instituto de Telecomunicações, 3810-193 Aveiro, Portugal
†Departamento de Eletrónica, Telecomunicações e Informática, Universidade de Aveiro, 3810-193 Aveiro, Portugal
‡IEETA, DETI, LASI, Universidade de Aveiro, 3810-193 Aveiro, Portugal
§Altice Labs, Aveiro, Portugal

***Abstract*—The demand for more secure, available, reliable, and fast networks emerges in a more interconnected society. In this context, 5G networks aim to revolutionize how we communicate and interact. However, studies using 5G data are sparse since there are only a few number of publicly available 5G datasets (especially about commercial 5G network metrics with real users). In this work, we analyze the data of a commercial 5G deployment with real users, and propose forecasting techniques to help understand the trends and to manage 5G networks. We propose the creation of a metric to measure the traffic load. We forecast the metric using several machine learning models, and we choose LightGBM as the best approach. We observe that this approach obtains results with a good accuracy, and better than other machine learning approaches, but its performance decreases if the patterns contain unexpected events. Taking advantage of the lower accuracy in the performance, this is used to detect changes in the patterns and manage the network in real-time, supporting network resource elasticity by generating alarms and automating the scaling during these unpredictable fluctuations.**

***Index Terms*—5G Networks, NWDAF, Dimensionality Reduction, PCA, Forecasting, LightGBM**

I. Introduction

Fifth Generation (5G) networks aim to revolutionize cellular networks by supporting billions of devices connected to the Internet without compromising the user's Quality-of-Experience (QoE). They support an elevated user demand, machine-to-machine communication, a massive amount of Internet of Things (IoT) devices, ultra-high-definition video and virtual reality applications. Thus, 5G networks have to deal with high scalability of devices, high data rate (10-50 Gbps), low end-to-end latency (less than 5ms), while increasing energy efficiency and reducing the cost [1]–[3].

Acknowledging the potential of 5G, by September 2023, 173 countries and territories have invested in this technology, from initial trials to network deployment and launches. From those, 114 countries and territories had launched 3GPP-compliant 5G services, and 113 had deployed 5G mobile services [4].

3GPP proposed a function named Network Data Analytics Function (NWDAF) to incorporate data analytics and machine learning in 5G networks, which allows the implementation of different use cases, such as monitoring and forecasting the network load performance, studying and detecting network anomalies, and analyzing and predicting traffic congestion [5]. On the other hand, the concept of network slicing enables an end-to-end network behavior for different verticals in the 5G networks [6]. We can use, for instance, Software-Defined Networks (SDNs) to create slices as virtual networks, not requiring infrastructure changes [7]. The big advantage of creating slices is to serve multiple use cases and end-users that expect different service requirements, in an isolated manner with low latency, high bandwidth, high throughput, high mobility and high security [7], [8]. The big advantage of using network slicing in 5G and NWDAF is the ability to optimize communication. Thus, network slicing can respond to the need of the different use cases and several end-users, providing infrastructure optimization and flexibility [7], but its dynamic management requires an accurate understanding of the traffic trends and their requirements at runtime.

There is a significant lack of public real-world 5G datasets available from networks or experimental testbeds, as the authors of [9] have recognized, which can impact the number of studies on the subject. It is critical to investigate how 5G networks are being used, understand 5G trends, and analyze what can be needed as more users start to use 5G. The work in [9] publicly released a 5G traffic dataset that was created by measuring various packet traffic. This dataset contains 328 hours of data (almost two weeks). To simplify the use of the dataset, the authors trained machine learning models to generate two types of traffic. Compared to the original dataset, they concluded that the generated traffic is similar to the original one. The work in [10] published a synthetic 5G network dataset based on a high-traffic event such as a major sports game.

This work uses real commercial 5G data, provided by a 5G network operator, to analyze and obtain insights about the trends in 5G usage and the observed patterns. To the best of

Ana Almeida acknowledges the Doctoral Grant from Fundação para a Ciência e Tecnologia (2021.06222.BD). Susana Brás is funded by national funds, European Regional Development Fund, FSE, through COMPETE2020 and FCT, in the scope of the framework contract foreseen in the numbers 4, 5 and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19. This work is also funded by the University of Aveiro through the funding 2.14.300.21 - 'VERBAS LIVRES - SUSANA SARGENTO', and by the European Union / Next Generation EU, through Programa de Recuperação e Resiliência (PRR) Project Nr. 29: Route 25. Furthermore, Ana Almeida would like to acknowledge the support of Dr. Koojana Kuladinithi and Researcher Daniel Stolpmann from the Institute of Communication Networks of the School of Electrical Engineering, Computer Science and Mathematics of the Hamburg University of Technology (TUHH).

ISBN: 978-3-903176-61-4

© 2024 International Federation for Information Processing (IFIP).
ISBN: 978-3-903176-61-4

our knowledge, this is the first work using real 5G commercial network data, and therefore, the first one providing this type of analysis and trend prediction with real data. We analyze trends, the relationship between features and temporal patterns. We propose a traffic load metric, and use Light Gradient-Boosting Machine (LightGBM) to forecast the network traffic with different horizons (1 day and 8 days) with high accuracy. Furthermore, we observe that in most cases, the best results are achieved using Gradient Boosting Decision Tree (GBDT) as the boosting method. We also observe that forecasting can be very useful in detecting data drift and can help manage network resources. By managing network resources as needed, we can achieve network resource elasticity. Furthermore, the forecasting model could be incorporated into the NWDAF to manage and optimize the resources needed for different network slices.

The main contributions of this work are the following:

- Data exploration pipeline of 5G network metrics;
- 5G network data anonymization through feature reduction using Principal Component Analysis (PCA);
- Research of a metric to measure the traffic load on 5G networks;
- 5G network traffic load forecasting to assist network management;
- Analysis of several machine learning algorithms, including LightGBM, to predict the trends of 5G network data;
- Forecast network traffic load trends to help manage network resources in the future timeframes.

The remainder of this paper is organized as follows. Section II presents the related work. Section III presents the methodology adopted. Section IV contains the results and discussion. Lastly, section V concludes the paper and provides ideas for future work.

II. Related Work

This section presents related works on 5G network data analysis, feature reduction, and algorithms that can be applied to 5G contexts.

Analyzing 5G network datasets can be overwhelming, since we can easily have hundreds of network metrics and thousands or millions of nodes, having a high dimensionality dataset. Feature reduction techniques can be used to solve the problem of high dimensionality. Several techniques can be employed, but we should consider that our high-dimensional dataset is a multivariate time series; therefore, the chosen method should preserve the properties of the time series data. In this context, several methods have been used, such as PCA, Singular Value Decomposition (SVD), kernel Principal Component Analysis (kPCA), t-Stochastic Neighbor Embedding (t-SNE), and AutoEncoders, among many others. Although they might be proposed for general tabular data, they have been proven to work well with multivariate time series [11], [12].

PCA [13] is a multivariate statistical model that aims to extract the most significant information from a dataset. PCA reduces the size of the dataset by representing it using fewer components than the original set of variables. The new set of features are orthogonal variables called principal components that result from linear combinations of the original set of variables. This can be achieved by the eigen-decomposition and SVD. The first component of PCA explains the largest variance of the data; the second component of PCA explains the second largest possible variance, and so on. The work in [14] proposed t-SNE, an unsupervised non-linear dimensionality reduction technique based on matching distances between distributions to reduce high dimensional datasets to lower dimensions. This method is also widely used for tabular and time series data [11], [12]. The work in [15] proposed a framework based on Sparse Tensor Factorization (STF) to perform dimensionality reduction for traffic analysis on 5G data.

Regarding forecasting strategies, we can divide them into historical, statistical, and machine learning methods. The machine learning models are the most popular and efficient of these three categories. Over the years, several methods have been proposed, some based on traditional machine learning and others based on deep learning [16]. In our previous work [17], we compared different deep learning models to forecast traffic flows, such as *Feed Forward Neural Network*s (FNNs), *Long Short-Term Memory*s (LSTMs), *Convolutional Neural Network*s (CNNs) and a hybrid LSTM-CNN approach. We observed that CNNs achieved the best performance and a lower training time. A different approach, LightGBM, is a tree-based model that proved, recently, very effective during the fifth edition of the forecasting accuracy competition [18]. LightGBM is a GBDT that uses strategies such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to deal with big, and highly dimensional datasets [19]. The work in [20] proposed using forecasting algorithms (from *AutoRegressive Integrated Moving Average* (ARIMA) to neural networks) to improve multi-slice 5G network management, using 4G data. The authors developed a real-time distributed forecasting framework and evaluated its performance in a vehicular and mobile scenario. They also proposed using a dynamic threshold for slice management and avoid network traffic congestion.

To predict the quality of wireless IoT networks' RSSI and Packet Delivery Ratio (PDR), Miguel Sindjoung et al. [21] use machine learning, namely Random Forest. The authors did not use the values of these metrics; they converted them into categories (good, intermedium, and bad) to classify the quality of the links. They also tested other classification methods, such as Logistic Regression, Support Vector Machines (SVMs) and Linear SVMs. The work in [22] proposed a real-time 5G wireless communication forecasting framework based on an LSTM network. The authors simulated a 5G network in a container environment. The work in [23] showed pratical use cases for NWDAF in 5G networks. They evaluated two scenarios: network load performance forecasting, using Linear Regression (LR), *Recurrent Neural Network* (RNN), and LSTM models, and classification of network anomalies using LR, and eXtreme Gradient Boosting (XGBoost). In this case, the authors generated a synthetic dataset for 5G cellular

ISBN: 978-3-903176-61-4

networks. Several studies have been conducted to forecast network traffic metrics, although, to the best of our knowledge, none of them used real commercial 5G data.

There are several problems that can degrade the performance of the models deployed across the network over time. This phenomenon is known as model drift. There are two types of model drift: data drift and concept drift, and each one of these types can be divided into subtypes. Data drift can be associated with changes in the data distribution, such as covariate shift and prior probabilistic shift. Concept drift is associated with changes in the relationship between features and targets. The presence of model drift can lead to wrong insights and harming decision-making. The main cause of model drift is the lack of data representative of the entire population [24]. Some solutions have been developed to deal with model degradation, or model drift. In the context of 5G networks, the work in [25] proposed a module for model drift detection and adaptation using LSTMs. The authors simulated a drift in the user behavior and were able to capture the model drift. In the context of IoT Data Streams, the work in [26] proposed the use of an adaptive LightGBM model with Optimized Adaptive and Sliding Windowing (OASW) for anomaly detection with concept drift adaptation. This work tested two public IoT datasets and outperformed state-of-the-art methods. The model provided high accuracy while consuming low memory and processing time.

The choice of the model to be used in the forecast task may vary according to the dataset characteristics. There are advantages and disadvantages to using each statistical, traditional machine learning, or deep learning method. While statistical methods are easier to explain, they also may work better with less data. Classical machine-learning techniques can be computationally less expensive and more explainable than deep-learning methods. On the other hand, deep-learning methods may capture more complex patterns, but require more data and are less explainable. Due to the limited amount of data available for training deep learning models, traditional machine learning models are more suitable in this case. Therefore, we selected traditional machine learning to predict network traffic metrics.

III. Methodology and Metrics

This section presents information about the network, the data, the traffic load metric and the methodology performed for the forecasting process.

A. Handling 5G commercial data

A commercial 5G network provider allowed us to explore a dataset containing commercial 5G data. The dataset contains 27 weeks of hourly data (a little bit more than six months) from several 5G base stations. For each base station, the operator provided 134 attributes/features. The attributes are fixed values about the base stations, and the features are network related performance metrics which vary over time. To focus our analysis, we select 13 base stations (named from A to M) and 15 5G network metrics. Due to privacy issues of the operator, the real map of the stations is not presented.

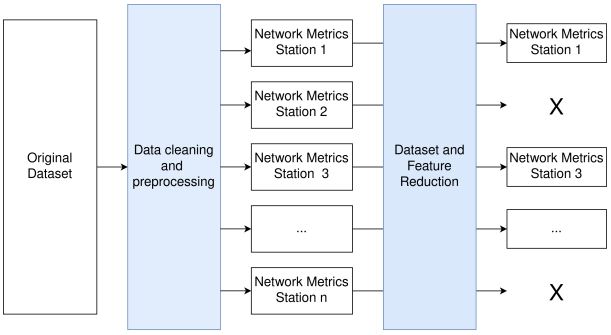

Fig. 1. Processing Pipeline

The processing pipeline is illustrated in Figure 1. We start by cleaning and preprocessing the dataset. We separate the dataset by base station, and obtain the data over time per base station with the information from that base station. Then, we select 13 base stations and 15 network metrics. Since some of the network metrics divide the information by uplink and downlink, and we do not need that type of differentiation, we decided to merge those fields. For instance, we add the uplink user data volume with the downlink user data volume. Then, we standardize the dataset to have a mean of 0 and a standard deviation of 1. After that, we test different feature reduction techniques (PCA and t-SNE) to obtain fewer features per base station.

B. Network Traffic Load Metric

In this section we propose a network traffic load metric to anonymize the information about the base stations and simplify the data analysis. We test two approaches: t-SNE and PCA. However, we observe that PCA can capture more patterns and variability using fewer components. Furthermore, the results for the different base stations are more similar and constant when using PCA than when using t-SNE.

Figure 2 contains the explained variance of PCA when using a different number of PCA components for base station D. By using two or more components in PCA, we can effectively account for over 90% of the data's variability. Increasing the number of components beyond this point offers diminishing returns in terms of information gain. Thus, we opted to pick the first two components.

Figure 3 contains the correlation matrix between the selected features and the PCA components. 'HS' stands for Handover Success, 'TPUT' for Throughput, 'Tx' stands for Transmission, and 'RB' stands for Resource Block. By applying PCA to the dataset, we observe that the first component of the PCA is significantly correlated with most of the selected features. The two features to which the first component is not significantly correlated are the average user throughput and the user data volume. Those features are captured in the second component of PCA, and we can observe that they are very correlated with the second component. We chose to

ISBN: 978-3-903176-61-4

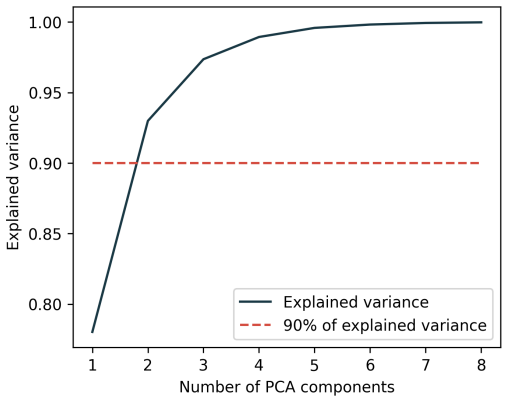

Fig. 2. PCA: Explained variance (base station D)

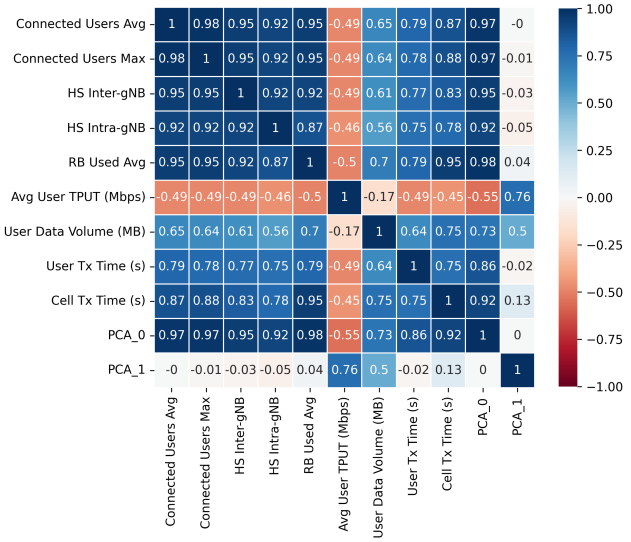

Fig. 3. PCA: correlation matrix between features (base station D).

concentrate our analysis on the primary component of PCA, as it effectively encapsulates the predominant patterns. However, for visualization purposes, we retained the second component. The components of PCA are a linear combination of the features. The features that are more related to the first PCA component can describe the data and give some notions of traffic load.

We proceeded with additive decomposition of the first PCA component. The additive decomposition divides the observed values into three components. The first one contains the seasonal patterns, the second includes the trend, and the third has the remaining patterns that the previous components could not capture. As shown in Figure 4, we can observe the weekly and monthly patterns by performing additive decomposition of the first component of PCA (example for the base station E). Furthermore, we can also observe an increasing trend in the traffic. This trend can be observed in most of the base stations. This indicates an increase in the network traffic that can be associated with the fact that more users are using the 5G network. These patterns may vary from base station to

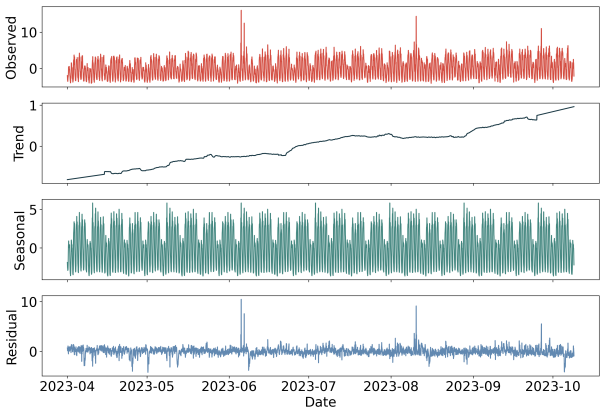

Fig. 4. Additive decomposition of the first component of PCA (base station E).

base station since they may exhibit different patterns. The base stations E, H, I, J, K, and L all show a consistent upward trend. On the other hand, stations A, C, and D have experienced a sharp increase in the last month or two months. Stations F and G, while showing an overall increasing trend, have had some fluctuations. Finally, stations B and M exhibit a more parabolic behavior.

Furthermore, we can conclude that the first component of the PCA could capture the seasonality present in the dataset. The residual component captures the noise existing in the dataset, the outliers, and other patterns that the first two components could not capture. The observed plot has four peaks that are captured by the residual component. Based on our analysis, we propose the creation of a traffic load metric based on the first component of the PCA.

C. Model Selection and Evaluation

We select LightGBM to perform forecasting. LightGBM[1] is an efficient and distributed framework for gradient boosting that employs tree-based learning algorithms, such as traditional GBDT, Dropouts meet Multiple Additive Regression Trees (DART), and GOSS. This framework uses low memory, can handle large datasets, and allows parallel and distributed computing. It also allows the use of the Graphics Processing Unit (GPU). Furthermore, it provides APIs in different languages, such as Python, C, and R. We select a traditional machine learning method, since we only have six months of data; deep neural networks require more data for the training process. Furthermore, we also add, as baseline models, LR, Lasso, Ridge, and ElasticNet[2]. All these baseline models are deterministic. Lasso, Ridge, and ElasticNet are regularization techniques applied to LR to address issues like multicollinearity and overfitting.

In forecasting scenarios, several evaluation metrics can be used to measure the models' performance by comparing the actual values with the predicted ones. We select two evaluation

[1]https://lightgbm.readthedocs.io/en/stable/
[2]https://scikit-learn.org/stable/

ISBN: 978-3-903176-61-4

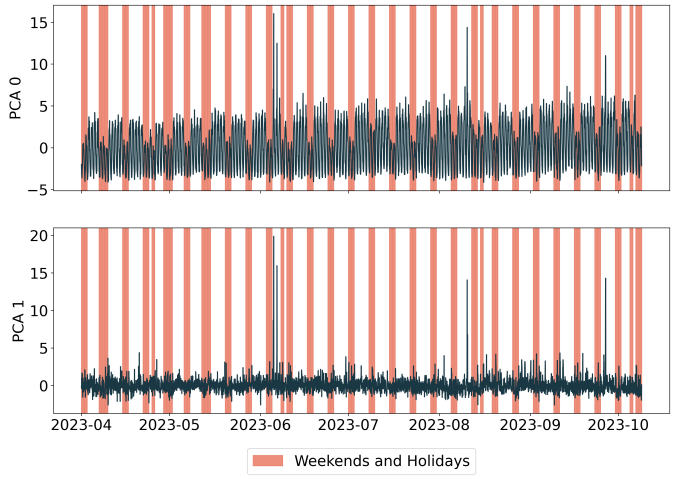

Fig. 5. Network Traffic in Weekends and Holidays (base station E)

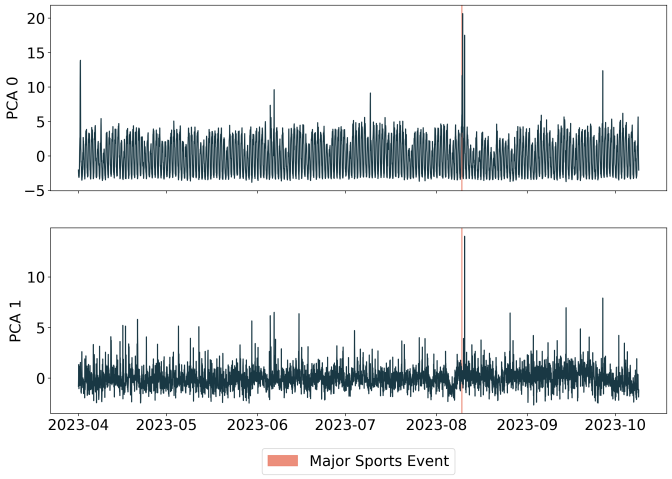

Fig. 6. Network Traffic and Major Sports Event (base station F)

metrics, Root Mean Squared Error (RMSE) and R^2-Score, and we decide to use R^2-Score to choose the best model [16], [17]. Since we are working with time series, we use cross-validation on a rolling basis without shuffling the data.

IV. Results and Discussion

This section presents the analysis and discussion of the results. It starts with a study about the relation of the network traffic load with events. Thereafter, the forecasting results are presented, showcasing a comparison of different models with the selected LightGBM. It concludes by demonstrating how the forecast can aid in detecting events and predicting the network traffic load for the subsequent month.

A. Network Traffic Load and Events

We start by comparing both PCA components, including the proposed traffic load metric, with weekends and holidays data. We observe that the network traffic load (first PCA component) decreases on the weekends and holidays for some base stations, while for others, it increases, depending on the location. For the particular case of base station E presented in Figure 5, the network traffic decreases on the weekends and holidays. This particular base station is located in an industrial area; therefore, most people visit the area during the weekdays due to their working routine. An example of a base station that increases its network traffic load value during the weekends and holidays is the base station near a shopping mall, where most people shop during those days.

We also study the impact of major sports events on the network traffic by comparing the values from a base station close to a sports field to important sports events. Usually, this sports field does not attract many sports fans. So, the baseline use of the network is, on average, not significant. However, there was an important sports event between two major teams in August. As visualized in Figure 6, there was a peak in the network traffic during the sports event. Furthermore, that peak was propagated to the time after the event. This might have happened because of the celebrations happening after the sports event. This is also noticeable in the second PCA component.

We study a base station near a school and compare the network traffic load metric with the school calendar. As can be observed in Figure 7, the network traffic metric is highly affected by school events. During the term, the regular event is attending classes on a daily basis between 9a.m. and 7p.m.. Therefore, we did not use any color to indicate it in the Figure. During the Easter holidays, school break, and teaching break, we can observe a lower network traffic compared to the normal days between April and mid-June. Furthermore, we can observe slightly lower network traffic during the regular and repeat exam seasons. During the summer break, the network traffic decreases until the end of August, increasing with the special exams session. It starts to increase during the special exam session in September, and we can see the biggest values during the beginning of the new school year in mid-September and October of 2023. As we expected, the network traffic metric is very affected by the events occurring nearby.

As we can observe in these three scenarios (weekends and holidays, sports events, and school events), some events contribute to changes in the patterns. The impact of these events in forecasting depends on how frequent these events are. For instance, the impact of weekends in forecasting is not a problem since they occur frequently and at regular intervals; however, the effects of holidays will be more significant since they happen less frequently and their pattern is more difficult to capture, especially when having only a small amount of data. The major sports event was sporadic, meaning that it will probably not repeat itself, and it will be more difficult to forecast for those conditions. Regarding the school events, with less than one year of data, it will be difficult for the forecasting algorithm to learn their patterns.

B. Forecasting Network Traffic Load

Before starting the forecasting, we decided to select the best temporal lags. The seasonal decomposition and autocorrelation

ISBN: 978-3-903176-61-4

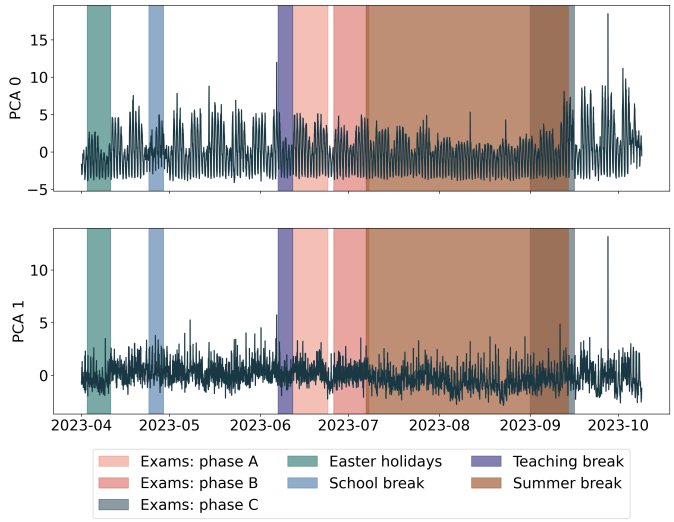

Fig. 7. Network Traffic and School Events (base station A)

TABLE I
LIGHTGBM PARAMETERS

Parameter	Value
Boosting type	GBDT, DART, GOSS
Subsample ratio of the training instance	0.2, 0.5, 0.8
Subsample ratio of columns when building each tree	0.2, 0.5, 0.8
Number of leaves	15, 31, 127
Maximum depth	-1, 3, 5, 7, 9
Learning rate	0.01, 0.05, 0.1
Number of estimators	50, 100, 200

plots gave us some insights regarding the best values for the temporal lags. Figure 8 depicts the correlation matrix between the first component of PCA and the temporal lags. As we can observe, the chosen temporal lags, such as 24h (1 day before), 48h (2 days before), and so on, strongly correlate with the first PCA component.

We also consider different parameters for LightGBM, present in Table I, resulting in 3645 combinations of parameters, and we repeat each experiment five times. Furthermore, we fix the seeds for the different experiments, so that we can reproduce the results. We experiment with three boosting types: GBDT, DART and GOSS. The maximum depth of -1 means that there is no limit.

Table II and Table III contain the best results for the different base stations obtained using the selected modes and forecasting horizons of 24 and 192 steps ahead (1 day and 8 days), respectively. Regarding Table II, the best overall method is LightGBM with an R^2-Score of 0,9293±0,0208. LightGBM achieves the best results for 6 base stations, Ridge Regression for 4, and LR for 3 base stations. When predicting values for the next day using LightGBM, all base stations have an R^2-Score greater than 0.89, and the best performance is achieved for base station D. We can also observe that LightGBM is the more stable method, having the lowest standard deviation, especially when compared to Ridge and Linear regression. For the forecasting horizon of 8 days, the best overall method is

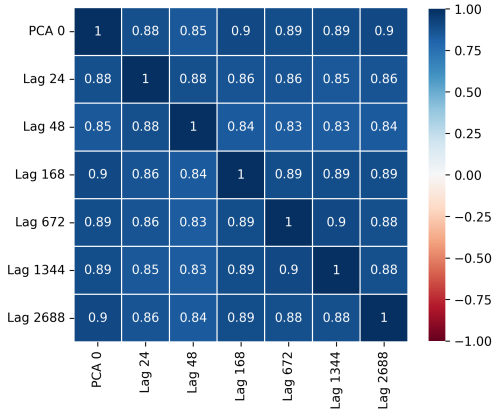

Fig. 8. PCA: correlation matrix between temporal lags (base station D)

still LightGBM with an R^2-Score of 0,8602±0,0768. For this case, there are only three base stations where LightGBM is not the best model, and LR performs better. Using LightGBM, most base stations achieve an R^2-Score greater than 0.85. The worst results are obtained for base station A when forecasting 192 steps ahead; this happened with all methods. Base station A is the base station close to the school and is highly affected by the events at the school. Once more, most of the events that affect the network traffic in this area are not seen, and it is more difficult for the model to forecast when there are patterns that the model did not observe. The best performance is achieved for base station I. We can observe small losses when comparing the performance of forecasting one day versus eight days ahead, as expected. Since we are increasing the forecasting horizon, there will be an increase in the error. The only exception is base station A, which presents a significantly larger difference. Given that we have only six months of data, meaning that there are patterns the model never saw, we consider these results a good first baseline. Furthermore, we also observe that Lasso regression achieves the worst performance for both cases. In some situations, LR might outperform LightGBM. This might be explained due to the linearity present in the data, meaning that introducing complexity might lead to overfitting. We could also use just one model to forecast all stations; however, we would also increase model dependency. If we needed to add more stations or remove stations from the model, we would have to train the model from the start. Furthermore, the train would be more difficult if the amount of data in the new stations is lower than the one of the old stations.

Regarding the best-boosting type, GBDT is the best for most of the forecasts, giving the top results for 8 base stations when forecasting one day, and 11 base stations when forecasting eight days. GOSS and DART are similar, with GOSS being the second best for 3 base stations in the first case and 1 base station in the second one, and DART is 2 and 1, respectively.

The patterns and statistical properties observed in the base station A are changing for the model, which means that we are in the presence of data drift. Thus, we achieve a worse

ISBN: 978-3-903176-61-4

TABLE II
RESULTS: FORECAST 24 STEPS AHEAD (1 DAY)

Station	LR		Ridge		Lasso		ElasticNet		LightGBM	
	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE
A	0,9538	0,3497	**0,9538**	0,3497	**0,9216**	0,4559	0,9471	0,3745	0,9443±0,0011	0,3842±0,0040
B	0,9516	0,5390	0,9516	0,5390	**0,9392**	0,6042	**0,9546**	0,5218	0,9524±0,0034	0,5340±0,0196
C	0,9200	0,7081	**0,9200**	0,7081	**0,8968**	0,8043	0,9114	0,7452	0,9176±0,0019	0,7186±0,0084
D	0,9449	0,5540	0,9449	0,5540	**0,8938**	0,7690	0,9123	0,6988	**0,9566±0,0032**	0,4912±0,0182
E	**0,9444**	0,5030	0,9444	0,5031	**0,8904**	0,7064	0,9155	0,6202	0,9410±0,0051	0,5176±0,0228
F	0,9438	0,5224	0,9438	0,5224	**0,8978**	0,7043	0,9166	0,6362	**0,9467±0,0013**	0,5087±0,0065
G	**0,5658**	2,0193	0,5658	2,0193	0,6232	1,8811	0,6003	1,9375	**0,8982±0,0073**	0,7406±0,0265
H	**0,9467**	0,5489	**0,9467**	0,5489	**0,8933**	0,7766	0,9261	0,6465	0,9455±0,0041	0,5546±0,0211
I	**0,9479**	0,8848	0,9479	0,8849	**0,8892**	1,2912	0,9099	1,1642	0,9380±0,0019	0,9655±0,0149
J	0,9116	0,7900	**0,9116**	0,7900	**0,8803**	0,9194	0,8956	0,8587	0,9107±0,0061	0,7937±0,0272
K	0,7602	0,9747	0,7602	0,9747	**0,6898**	1,1086	0,7269	1,0403	**0,9022±0,0042**	0,9721±0,0215
L	0,9181	0,7455	0,9180	0,7455	**0,8895**	0,8657	0,9028	0,8118	**0,9253±0,0040**	0,7113±0,0196
M	**0,7872**	0,7139	0,7872	0,7139	0,8283	0,6412	0,8266	0,6444	**0,9021±0,0052**	0,4840±0,0130
Avg±Std	0,8843±0,1145	0.7579±0,4163	0,8843±0,114	0,7579±0,4163	**0,8564±0,0931**	0,8868±0,3682	0,8727±0,1009	0,8231±0,3936	**0,9293±0,0208**	0,6443±0,1878

TABLE III
RESULTS:FORECAST 192 STEPS AHEAD (8 DAYS)

Station	LR		Ridge		Lasso		ElasticNet		LightGBM	
	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE	R²-Score	RMSE
A	**0,5728**	2,0096	0,5728	2,0096	0,6125	1,9138	0,6044	1,9339	**0,6261±0,0039**	1,8799±0,0100
B	**0,9032**	0,9919	0,9032	0,9920	**0,8652**	1,1708	0,8862	1,0758	0,9009±0,0006	1,0036±0,0031
C	0,8857	1,1631	0,8857	1,1631	**0,8551**	1,3098	0,8623	1,2768	**0,8901±0,0042**	1,1402±0,0219
D	**0,8806**	0,9932	0,8806	0,9933	**0,8420**	1,1425	0,8597	1,0766	0,8655±0,0017	1,0538±0,0066
E	**0,8590**	1,0438	0,8590	1,0438	0,8622	1,0318	0,8685	1,0079	**0,8706±0,0051**	0,9993±0,0197
F	0,8618	0,9864	0,8618	0,9863	**0,8363**	1,0734	0,8648	0,9754	**0,9038±0,0007**	0,8224±0,0033
G	0,8097	1,0964	0,8097	1,0963	**0,8011**	1,1208	0,8158	1,0785	**0,8396±0,0025**	1,0073±0,0109
H	0,8660	0,9472	0,8660	0,9472	**0,8591**	0,9713	0,8850	0,8773	**0,8900±0,0001**	0,8580±0,0006
I	0,9320	0,8184	0,9320	0,8183	**0,9087**	0,9482	0,9310	0,8248	**0,9393±0,0009**	0,7729±0,0058
J	0,8871	1,0830	0,8871	1,0831	**0,8398**	1,2904	0,8546	1,2292	**0,8874±0,0014**	1,0814±0,0070
K	0,8114	1,2600	0,8114	1,2600	**0,7975**	1,3055	0,8114	1,2600	**0,8136±0,0037**	1,2522±0,0126
L	**0,8926**	1,0048	0,8926	1,0048	**0,8560**	1,1637	0,8736	1,0900	0,8903±0,0004	1,0153±0,0021
M	0,8530	0,9505	0,8530	0,9505	**0,8278**	1,0289	0,8408	0,9894	**0,8652±0,0006**	0,9103±0,0022
Avg±Std	0,8472±0,0891	1,1037±0,2930	0,8473±0,0891	1,1037±0,2930	**0,8279±0,0707**	1,1900±0,2484	0,8429±0,0779	1,1304±0,2769	**0,8602±0,0768**	1,0613±0,2783

performance when trying to predict bigger horizons. Even though base station A's performance is lower than the remaining ones, this is still useful for the network management. The performance is lower because the network traffic usage on the base station A increased in the last weeks with the beginning of the new school year. In this case, the users are mostly school students and staff. They are consuming more resources, and the real needs surpass the forecasted needs. If this pattern continues, eventually, more resources will be needed to be available. So this means that we require some additional intervention. We can throw an alert or act to adapt the necessary resources. By applying forecasting methods, we can help to manage the resources available for the network. Therefore, forecasting is very useful for ensuring network elasticity by forecasting the resources needed and by identifying moments of unexpected fluctuations that require changes in the resources available. Furthermore, the network can help to predict and detect events, even though it might not know the context of the events.

In order to understand how the network traffic load would behave in the upcoming month, we calculate the estimated increase or decrease in the network traffic load for October. To achieve this, we use the best model for each station to predict the whole month of October. As the PCA gives negative values, we normalize the observed and forecasted values between 0 and 1. Next, we calculate the mean of the network traffic load for each month and use it to determine the percentage of change from one month to the following one.

TABLE IV
FORECASTING THE TREND FOR THE NEXT MONTH

Station	Min	Max	Forecast next month
A	-15,05%	47,87%	3,65% ± 0,06
B	-8,16%	6,43%	6,46% ± 0,02
C	-3,22%	22,75%	2,28% ± 0,42
D	-3,29%	4,80%	2,63% ± 0,09
E	2,05%	9,54%	-2,78% ± 0,16
F	-9,85%	8,10%	-0,23% ± 0,07
G	-4,16%	13,11%	-0,20% ± 0,61
H	6,14%	20,33%	1,94% ± 0,05
I	-2,38%	11,03%	3,52% ± 0,52
J	0,71%	14,33%	4,61% ± 0,13
K	0,36%	15,59%	2,29% ± 1,03
L	2,49%	11,26%	0,66% ± 0,11
M	-17,04%	10,78%	5,64% ± 0,09
		Avg±std	2,29% ±2,55

Table IV contains the minimum and maximum percentage of change of the observed months and the forecasted percentage of change for October. Depending on the station, we can observe some changes in the network traffic load, including both increases and decreases. However, on average, we can observe an increase in the network traffic load, corresponding to an increase in the usage of 5G. Using these models on the NWDAF, network resources can be managed more effectively, including resource allocation for each slice in a dynamic approach.

The forecasting of the network usage can indicate the necessity to create/modify network slices to guarantee specific

ISBN: 978-3-903176-61-4

requirements of bandwidth and latency for critical services such as emergency use cases. A particular example of network slicing appliance is the selection of User Plane Functions (UPFs). Due to the utilization (and demand forecasting) of a certain gNB, the network slices can be programmed to utilize a different UPF to redirect the traffic through different network resources, and to better utilize edge computing. The orchestration of the UPFs is done by the Session Management Function (SMF), and the registration of the network slices is stored in the Network Slice Selection Function (NSSF).

V. Conclusions and Future Work

Monitoring network metrics is a first step to improve the network usage, resource allocation, network traffic optimization, predictive maintenance, security and network slicing. Integrating machine learning models with NWDAF in 5G networks enables intelligent decision-making and automation, leading to a more efficient, reliable, secure and proactive network. 5G network datasets tend to have a high number of features, being highly dimensional, which can make the analysis and application of models difficult. This work proposes using PCA to simplify the study of 5G networks. The proposed approach is a first step to build a forecasting framework for 5G analysis and usage prediction. We achieved good forecasting results for two forecasting horizons: 1 day and 8 days. As expected, increasing the forecasting horizon leads to a decrease in the performance. Our work showcases the practical application of machine learning models using NWDAF, which enables the support of dynamic network slicing in the expansion of the 5G networks and deployments. In future research, we plan to test more methods to propose the best forecasting method for 5G expanding deployments. Since we expect more data, we aim to experiment with deep learning approaches. Furthermore, we also expect to contribute to detecting data drift in 5G networks.

References

[1] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[2] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5g networks for the internet of things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[3] N. Al-Falahy and O. Y. Alani, "Technologies for 5g networks: Challenges and opportunities," *IT Professional*, vol. 19, no. 1, pp. 12–20, 2017.

[4] "5G-Market Snapshot November 2023 - GSA," 11 2023. [Online]. Available: https://gsacom.com/paper/5g-market-snapshot-november-2023/

[5] ETSI, *technical Specification: 5G; 5G System; Network Data analytics Services; Stage 3*, Example City, CA, July 2023, available at https://www.etsi.org/deliver/etsi_ts/129500_129599/129520/17.11.00_60/ts_129520v171100p.pdf.

[6] R. Trivisonno, X. An, and Q. Wei, "Network slicing for 5g systems: A review from an architecture and standardization perspective," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2017, pp. 36–41.

[7] P. Subedi, A. Alsadoon, P. W. C. Prasad, S. Rehman, N. Giweli, M. Imran, and S. Arif, "Network slicing: a next generation 5g perspective," *EURASIP Journal on Wireless Communications and Networking 2021 2021:1*, vol. 2021, pp. 1–26, 4 2021. [Online]. Available: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-021-01983-7

[8] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network slicing for 5g: Challenges and opportunities," *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.

[9] Y.-H. Choi, D. Kim, M. Ko, K.-y. Cheon, S. Park, Y. Kim, and H. Yoon, "Ml-based 5g traffic generation for practical simulations using open datasets," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 130–136, 2023.

[10] K. R. K M and R. K. Lanke, "5g network metrics for high traffic event," 2023. [Online]. Available: https://dx.doi.org/10.21227/1ryt-wb82

[11] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625351930377X

[12] M. Ashraf, F. Anowar, J. H. Setu, A. I. Chowdhury, E. Ahmed, A. Islam, and A. Al-Mamun, "A survey on dimensionality reduction techniques for time-series data," *IEEE Access*, vol. 11, pp. 42 909–42 923, 2023.

[13] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.101

[14] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2002. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

[15] J. Wang, H. Han, H. Li, S. He, P. Kumar Sharma, and L. Chen, "Multiple strategies differential privacy on sparse tensor factorization for network traffic analysis in 5g," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1939–1948, 2022.

[16] A. Almeida, S. Brás, S. Sargento, and F. Pinto, "Time series big data: a survey on data stream frameworks, analysis and algorithms," *Journal of Big Data*, vol. 10, 05 2023.

[17] A. Almeida, S. Brás, I. Oliveira, and S. Sargento, "Vehicular traffic flow prediction using deployed traffic counters in a city," *Future Generation Computer Systems*, vol. 128, 10 2021.

[18] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021001874

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[20] D. Ferreira, A. Braga Reis, C. Senna, and S. Sargento, "A forecasting approach to improve control and management for 5g networks," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1817–1831, 2021.

[21] M. L. F. Sindjoung and P. Minet, "Wireless link quality prediction in iot networks," in *2019 8th International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN)*, 2019, pp. 1–6.

[22] R. Reddy, Y. Munoz, C. Lipps, and H. D. Schotten, "Supervised wireless communication: An analytic framework for real-time model inference in the 5g core network," in *2023 International Balkan Conference on Communications and Networking (BalkanCom)*, 2023, pp. 1–5.

[23] S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz, and T. Tugcu, "Intelligent network data analytics function in 5g cellular networks using machine learning," *Journal of Communications and Networks*, vol. 22, no. 3, pp. 269–280, 2020.

[24] D. M. Manias, A. Chouman, and A. Shami, "Model drift in dynamic networks," *IEEE Communications Magazine*, vol. 61, no. 10, pp. 78–84, 2023.

[25] Dimitrios Michael Manias and Ali Chouman and Abdallah Shami, "A model drift detection and adaptation framework for 5g core networks," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.06852

[26] L. Yang and A. Shami, "A lightweight concept drift detection and adaptation framework for iot data streams," *IEEE Internet of Things Magazine*, vol. 4, no. 2, pp. 96–101, 2021.

ISBN: 978-3-903176-61-4