QoS and Capacity Prediction for 5G Network Slicing

1st Nathalie Romo Moreno*, 2nd Felix Dsouza*, 3th Andreas Kassler†, 4rd Florian Pullem*, 5th Bangnan Xu*, 6th Markus Amend*, 7th Changsoon Choi*
*Deutsche Telekom AG
Email: {nathalie.romo-moreno, felix.dsouza, florian.pullem, B.Xu, markus.amend, changsoon.choi}@telekom.de
†Faculty of Applied Computer Science, Deggendorf Institute of Technology (THD), Deggendorf, Germany
Email: andreas.kassler@th-deg.de

***Abstract*—Efficient resource allocation is fundamental to enabling 5G network slicing with guaranteed Quality of Service (QoS). Achieving this requires accurate forecasting of cell utilization, which remains challenging due to dynamic factors such as user mobility, radio channel variability, and fluctuating traffic loads. Existing approaches predominantly rely on simulated data and static radio models, limiting their applicability in live deployments. In this work, we propose a hybrid forecasting framework that combines machine learning-based time series prediction with radio channel modeling grounded in live network measurements. The method estimates service-specific radio resource demand at fine-grained spatial and temporal granularity and supports slice admission control through proactive feasibility assessment. This is realized by jointly forecasting required radio resources and the expected utilization of physical resource blocks (PRBs) in each cell. In our evaluation, we use data from a European Tier-1 operator and demonstrate a Mean Absolute Error (MAE) of ±0.74 for uplink interference prediction and a MAE of ±1.72 for cell load estimation when using an LSTM predictor.**

***Index Terms*—5G Network Slicing, Quality of Service (QoS), Spectral Efficiency, Radio Access Network (RAN), Machine Learning (ML), Time Series Forecasting, Predictive Analytics**

I. Introduction

The introduction of Network slicing in 5G has enabled operators to dynamically allocate resources across shared infrastructure and provision customized virtual networks tailored to specific services and user requirements. [1]. Efficient slice design and deployment require (i) accurate estimation of the resources needed to meet service demands and (ii) awareness of available network capacity. As network conditions evolve over time, forecasting key performance indicators (KPIs)—such as required and utilized Physical Resource Block (PRB)s—is essential to maintain service assurance and comply with Service Level Agreements (SLAs). This is particularly challenging in the RAN domain, where PRB demand is tightly coupled to spectral efficiency and radio channel quality, both of which are strongly influenced by environmental conditions that vary across time and location.

Conventional 5G RAN resource estimation for Network Slicing design relies on static planning and theoretical models, which fail to account for real-world dynamics. As a result, operators often adopt conservative, overprovisioning strategies, leading to inefficient resource use and limited scalability. Addressing these limitations requires advanced estimation techniques that incorporate live network conditions with spatial and temporal granularity. Such methods must enable forecasting of future network states within the relevant service time windows to support accurate assessment of both resource demand and available RAN capacity.

Estimating spectral efficiency is central to determining the radio resources required for a given service. Existing research explores diverse machine learning (ML) techniques, leveraging network metrics that might influence radio channel quality. However, most studies rely on simulated data [2], and the limited work using live measurements often neglects the temporal variability of radio conditions [3], [4]. As a result, current approaches struggle to fully capture the dynamics of operational networks. While other studies focus on temporal prediction of network KPIs to support 5G slice resource allocation [5], [6], these efforts primarily forecast traffic volume, which cannot be directly translated into required PRBs.

In response to the challenges identified in prior work, this paper addresses the problem of modeling and forecasting RAN spectral efficiency to estimate the number of PRBs required to meet specific QoS targets at a given time and location. We integrate these predictions with forecasts of available radio resources to support efficient and scalable 5G network slice deployment. Our approach leverages live network measurements, applying machine learning time-series forecasting in conjunction with channel quality estimation based on Shannon's law [7]. Our key contributions are:

- We develop a forecasting framework that couples interference ML predictions with radio modeling to derive spectral efficiency at a given location and time, enabling precise PRB demand estimation for specific QoS requirements.
- We propose a novel unified forecasting and admission-control pipeline tailored for 5G Network Slicing. The pipeline combines predicted resource demand and cell load forecasting to achieve proactive and efficient radio resource allocation.
- We conduct an empirical evaluation of different ML forecasting models using Tier-1 operator data from a live commercial 5G network, demonstrating a Mean Absolute Error (MAE) of ±0.74 for uplink interference prediction and a MAE of ±1.72 for cell load estimation using LSTM.

978-3-903176-75-1

II. Related Work

Several studies have explored spectral efficiency prediction using ML techniques, each with distinct objectives and data sources. [2] employs ML models with simulated data to predict spectral efficiency in Multiple-Input Multiple-Output (MIMO) systems, achieving a Mean Absolute Percentage Error (MAPE) below 10% with gradient boosting and neural networks across different precoding schemes. [3] targets spectral efficiency prediction in real-world 5G deployments using drive test data, focusing primarily on RSRP as input. Their approach integrates domain knowledge into model selection and evaluation. [4] proposes a cell-level forecasting framework for capacity planning, evaluating models such as linear regression, feedforward neural networks, and XGBoost. In the context of temporal forecasting for RAN resource demand, particularly for network slicing, recent work has focused on traffic-based predictions. [5] introduces an Intelligent Resource Scheduling Strategy (iRSS) that combines LSTM for long-term traffic forecasting with A3C-based reinforcement learning for short-term decisions. [6] proposes an X-LSTM model to predict the REVA metric for estimating resource needs of highly active bearers.

While these approaches offer valuable insights, they primarily emphasize traffic volume or coarse-grained metrics. In contrast, our work focuses on fine-grained, location- and time-specific forecasting of spectral efficiency and PRB utilization using live network data to support accurate and dynamic slice feasibility assessment. There is extensive literature exploring the integration of time series forecasting into self-aware systems [8] and decision-making processes [9], showing that accurate predictions of future states in complex systems enable more efficient and proactive control-loop decisions. Building on these insights, this work demonstrates a practical implementation of a forecasting and admission-control pipeline for Network Slicing.

III. Methodology

This research focuses on a QoS and capacity prediction system designed for a Live Video Production (LVP) service delivered via a 5G Network Slice. In this use case, the service requires a sustained uplink (UL) throughput of 8 Mbps to support professional-grade live video streaming without degradation. While the LVP scenario is used to demonstrate the system's capabilities, the approach is generalizable to other services with different QoS requirements. The LVP slice is provisioned through Radio Resource Partitioning (RRP), which reserves the necessary number of PRBs in the serving cell. Customers request bookings by specifying a target location and time. The system is then designed to estimate the number of PRBs required for the service, and to predict the serving cell load to verify whether the resources can be allocated without impacting existing users.

Figure 1 presents the flow diagram executed upon a customer booking request. In Step 1, the provided location is used to identify the serving cell and to retrieve its configuration data, the past seven days of Uplink (UL) interference and PRB utilization, and Downlink (DL) Reference Signal Received Power (RSRP) at the specified location.

In Step 2, UL interference is forecasted over the requested time window. Combined with cell configuration parameters and estimated UL RSRP measurements, this forecast is used to compute spectral efficiency, and subsequently, the required number of PRBs. In Step 3, PRB utilization is forecasted for the requested time interval. The system then compares the predicted PRB demand with the forecasted utilization and existing reservations to decide whether the booking can be accepted.

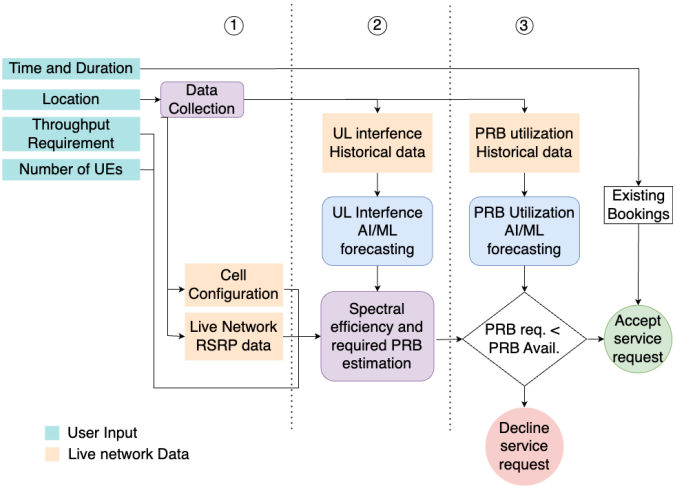

Fig. 1. QoS and capacity prediction system

The following subsections provide a detailed description of the data collected and used in this research; the process for estimating spectral efficiency and required PRBs; and the implementation of ML models for forecasting UL interference and PRB utilization.

A. Data collection and pre-processing

As depicted in Figure 1, two ML models predict average UL PRB utilization and UL interference (interference plus noise) at the cell level. The data corresponding to these target variables was retrieved as PM counters from each gNB in the network. For training, we collected one month of data at the highest available granularity (15-minute intervals), added daily and weekly sinusoidal encodings, and normalized the features using z-score. The spectral efficiency and required PRB estimation process takes as input DL RSRP measurements and cell configuration information. DL RSRP measurements were retrieved from the Measurement Reports (MR) generated by the different UEs connected to the live network and averaged daily (approximately 8000 to 10000 measurements for the selected locations). Configuration information was retrieved from the network inventory data base and includes the frequency band, subcarrier spacing, and channel bandwidth. All data described above were collected from a Tier-1 operator's live network.

B. Spectral efficiency and PRB estimation

The spectral efficiency of the uplink channel is calculated based on Shannon's law [7] outlined in Equation 1

TABLE I
MODEL TRAINING CONFIGURATION PER TARGET VARIABLE.

Variable	PRB Utilization			Interference		
Model	LSTM	TCN	Hybrid	LSTM	TCN	Hybrid
N° of Epochs	50	50	80	50		
Batch size	8	8	16	8		
Optimizer	Adam			Adam		
Starting learning rate	0.001			0.001	0.001	0.005
Loss function	MSE			MSE		

$$\eta = \frac{C}{B} = log_2(1 + UL\_SINR) \quad (1)$$

where η is Spectral efficiency, C is the Maximum achievable channel capacity, B is the Channel bandwidth, and UL_SINR is the Uplink Signal to Interference Noise Ratio (SINR) calculated using Equation 2.

$$SINR[dBm] = UL\_RSRP[dBm] - UL\_I(t)[dBm] \quad (2)$$

The UL RSRP values are estimated from the DL RSRP measurements collected from the network MR reports, and the interference value at a given time (UL_I(t)), corresponds to the prediction generated by the interference forecasting model described in section III-C.

The number of PRBs required to transmit data at a specific throughput is then calculated using equation 3. The throughput is given by the service requirement (8Mbps), and the bandwidth per PRB is given by the cell configuration.

$$PRBs_{required} = \frac{Throughput}{\eta \times PRB_{Bandwidth}} \quad (3)$$

C. UL Interference and PRB utilization forecasting

Both forecasting models —PRB utilization and UL interference— were designed to predict three days (288 time steps) of future values using the previous 7 days (672 samples) of data as input, aligning with typical booking lead times. TWe implemented and evaluated three ML forecasting architectures: a standard LSTM, a Temporal Convolutional Network (TCN), and a hybrid model combining LSTM, TCN, and attention mechanisms. LSTM and TCN were chosen for their effectiveness in capturing long-term dependencies and temporal patterns in sequential data [10], [11]. The hybrid model integrates both architectures and adds an attention mechanism to dynamically focus on the most relevant time steps [12]. To improve generalization and reduce overfitting, , we used a random shifting-window approach [13], dynamic learning rate scheduling, and dynamic dropout. Table I summarizes the selected hyperparameters for the architectures implemented for each target variable.

IV. EVALUATION AND RESULTS

In our evaluation we answer the following questions:

- What is the accuracy of the different models when forecasting interference and PRB utilization (c.f. IV-A)?
- How accurate can we predict spectral efficiency and required PRBs to meet service demands (c.f. IV-B)?
- How can our framework be used to estimate service-level scalability, i.e. the number of simultaneous users without experiencing service quality degradation (c.f. IV-B)?

To address these questions, we conducted field measurements at four distinct locations in the city center of Bonn, Germany, which we compared with the predictions generated by our system. Measurements were taken at different times and at locations served by the same cell but situated at varying distances from it (see Figure 2), allowing us to capture both temporal and spatial variability.

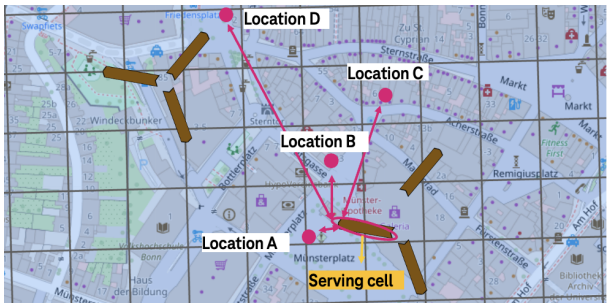

Fig. 2. Measurement locations in Bonn city center. Location A: public square. Location B: coffee shop. Location C: bakery. Location D: restaurant/bus stop

A. Forecasting Model Evaluation

We evaluated the three model architectures described in section III-C against the following baselines: (i) last observation, (ii) moving average over 10 slots, and (iii) simple exponential smoothing (SES) with α = 0.5. Performance was measured using Mean Squared Error (MSE) and Mean Absolute Error (MAE) and the corresponding results are outlined in Table II.

We observe that the Long Short-Term Memory (LSTM) architecture provides the best results in terms of MSE and MAE for both target variables. Furthermore, this model requires significantly less training effort and computational resources compared to the other two architectures, which makes it the more efficient and accurate choice for this use case. In addition, Figure 3 illustrates the performance of the built models and baselines across different forecasting horizons for the PRB utilization, using MAE as performance metric. These results show that the LSTM architecture has the best accuracy across the whole forecast horizon, maintaining lower errors even for the longest windows. The outperforming behavior of the LSTM model can be explained by the nature of the dependencies in the data and the feature engineering process. Usually, TCN models excel at capturing hierarchical and multi-scale dependencies; however, as the input dataset was enriched with weekly and daily sinusoidal encodings (see section III-A), the LSTM gating mechanism is sufficient to provide good performance when capturing these seasonalities. In this case, the hierarchical receptive fields of a TCN or the attention mechanism in the hybrid architecture do not bring significant extra benefit, while they still add computational complexity.

TABLE II
MODEL ACCURACY COMPARISON AVERAGED OVER 288 STEPS AHEAD

Model	PRB Cell Load		Interference	
	MAE	MSE	MAE	MSE
LSTM	**1.72**	**9.85**	**0.74**	**0.87**
TCN	2.24	13.33	0.75	**0.87**
Hybrid	2.15	12.74	0.79	0.97
Last observation	9.3	139.26	109.99	12107.15
Moving average	8.75	129.29	110.46	12211.26
SES	9.11	135.76	110.188	12149.36

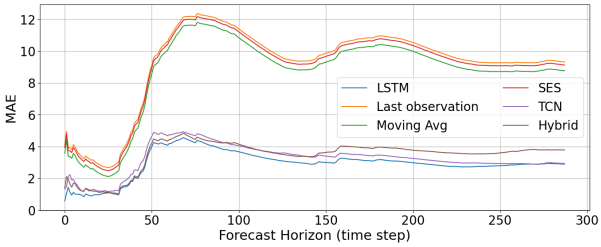

Fig. 3. PRB forecasting model comparison based on MAE over different Time Horizon

Figure 4 and Figure 5 show the corresponding predicted vs. measured values for UL Interference and PRB utilization of the serving cell in the selected locations. The predictions were generated using the best performing architecture i.e LSTM. The data patterns and seasonalities can be clearly observed for both target variables throughout the six days of data depicted in the plots, and the predicted values for the last three days show that these seasonalities are well captured by both models. In Figure 5, PRB utilization exhibits distinct outliers that can not be captured by the model based solely on historical data, leading to higher MAE and MSE values for this target variable. As depicted in Figure 2, the selected testing locations are placed in Bonn city center. Accordingly, the seasonality observed in the data reflects typical urban usage patterns, thus, the models should generalize to similar urban environments across cities and operators. Rural areas or event venues (e.g., stadiums, concert locations) may require additional features and/or training on different data.

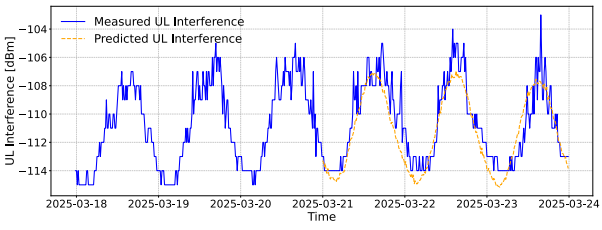

Fig. 4. LSTM Model: Predicted vs Measured UL Interference

B. *Spectral efficiency and PRB demand estimation evaluation*

To address the question of how accurately we can predict spectral efficiency and the number of PRBs required to meet service demands, we conducted experiments in which one User Equipment (UE) actively transmitted uplink (UL) data

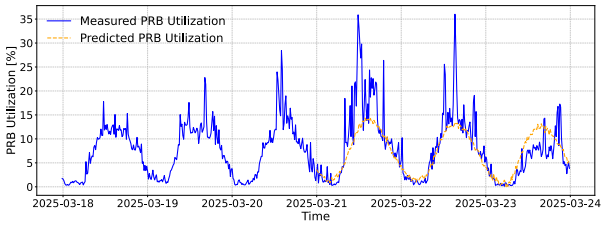

Fig. 5. LSTM Model: Predicted vs. Measured PRB Utilization

at 8 Mbps at each selected location. A debugging tool on the UE was used to record the number of PRBs allocated for UL transmission. Each experiment was conducted over a 30-minute period at each location, with measurements taken every 5 minutes. On each test the UE was served by a 5G New Radio (NR) base station operating in the n78 band (3.6Ghz) with a 90Mhz channel bandwidth and 30KHz Subcarrier Spacing (SCS). This configuration yields 245 available PRBs and a per-PRB bandwidth of 360 kHz [14] [15]. With the collected data we performed the evaluation using the following procedure:

- Step 1: we forecast the PRB resource demand for the LVP service for each given location.
- Step 2: we forecast the base PRB cell load generated by regular traffic using the built LSTM model.
- Step 3: we compare the total predicted PRB load (Predicted resource demand from Step 1 + Predicted PRB load from regular traffic from Step 2) with the measured PRB utilization (during an LVP service) using the Mean Error (ME) as statistical measure (equation 4).

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(P_i - M_i) \tag{4}$$

where P_i is the predicted number of PRBs for the i-th time instance, M_i is the measured number of PRBs at the given time, and n is the total number of test samples.

The predicted PRB resource demand for the target throughput was calculated with Equation 2, using the forecasted interference levels of the serving cell generated by the built LSTM model, the location based UL RSRP estimations, and the previously mentioned cell configurations.

Figure 6 shows the results of the evaluation procedure with the plots of the base PRB load forecast, the total PRB utilization prediction (base forecast + required PRBs) and the measured PRB load during the LVP service. The ME values for the predictions at each location are the following: Location A = -0.73, Location B = +3.96, Location C = +1.70, location D = +1.80. It is worth to mention that PRB demand prediction accuracy depends strongly on the accuracy of DL RSRP measurements. In our test locations, abundant UE MR reports made RSRP estimates robust; however, during the data collection and analysis for this research, several locations with very scarce or no recent MR reports were found. For predictions on such scenarios, additional ML algorithms or radio channel modeling estimations might be necessary.

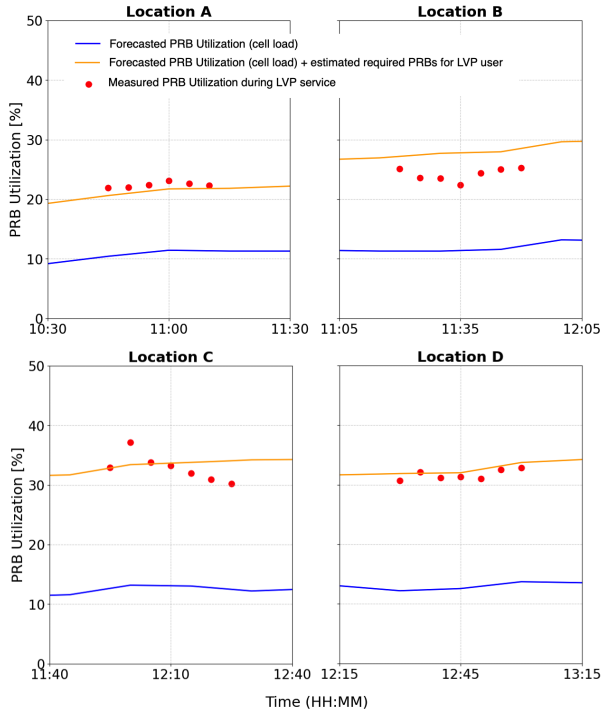

Fig. 6. Predicted vs Measured PRB utilization during LVP service.

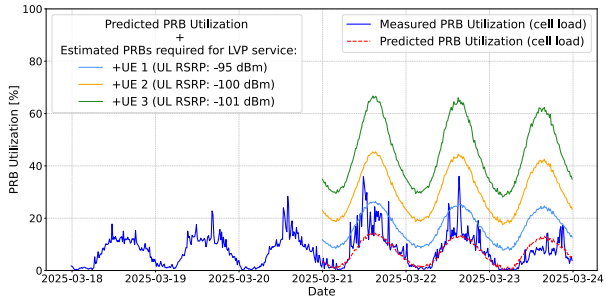

Fig. 7. Predicted impact of additional LVP users.

Finally, Figure 7 shows forecasted PRB utilization with additional predicted demand from three LVP users. As observed, the overall utilization is below saturation, demonstrating that one cell can support up to three concurrent LVP users, and that the designed forecasting and control pipeline enables reliable service scalability. The predicted PRB demand for the different LVP users is generated using Equations 3, 1, and 2 with the estimated UL RSRP values depicted in the figure, and forecasted interference for the serving cell.

V. Conclusion and Future Work

This paper presented a QoS- and capacity-aware prediction framework for 5G Network Slicing, integrating time series forecasting with radio channel modeling to estimate the number of PRBs required to meet specific service-level demands and to forecast cell level PRB utilization. Validation results demonstrate the system's effectiveness in accurately forecasting uplink interference and PRB utilization, enabling the admission of services — such as Live Video Production — without degrading existing traffic. By combining predicted demand with forecasted resource availability, the system supports proactive and efficient radio resource allocation in dynamic network environments.

Future work will extend evaluation scenarios to assess the impact of radio channel modeling in more diverse environments and enhance forecasting capabilities by adding contextual features (e.g., weather, scheduled public events). These enhancements aim to improve prediction accuracy during atypical traffic conditions, further increasing the robustness and applicability of the system across broader use cases.

Acknowledgment

This work has been partly funded by the Bavarian State Ministry of Education, Science, and Art through the High-Tech Agenda (HTA).

References

[1] GSM Association, "An Introduction to Network Slicing," 2017. [Online]. Available: https://www.gsma.com/solutions-and-impact/technologies/networks/wp-content/uploads/2020/01/1.0_An-Introduction-to-Network-Slicing.pdf

[2] E. Bobrov *et al.*, "Machine Learning Methods for Spectral Efficiency Prediction in Massive MIMO Systems," 12 2021. [Online]. Available: http://arxiv.org/abs/2112.14423

[3] Z. Xing *et al.*, "Spectrum Efficiency Prediction for Real-World 5G Networks Based on Drive Testing Data," in *IEEE WCNC*, vol. 2022-April. IEEE Inc., 2022.

[4] I. Tomic *et al.*, "Predictive Capacity Planning for Mobile Networks-ML Supported Prediction of Network Performance and User Experience Evolution," *Electronics (Switzerland)*, 2 2022.

[5] M. Yan *et al.*, "Intelligent resource scheduling for 5G radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691–7703, 8 2019.

[6] C. Gutterman *et al.*, "RAN resource usage prediction for a 5G slice broker," in *MobiHoc Proceedings*. ACM, 7 2019.

[7] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 7 1948.

[8] A. Bauer *et al.*, "Time Series Forecasting for Self-Aware Systems," pp. 1068–1093, 7 2020.

[9] F. Zito *et al.*, "Data-driven forecasting and its role in enhanced decision-making," *Engineering Applications of Artificial Intelligence*, vol. 154, 8 2025.

[10] B. Lindemann *et al.*, "A survey on long short-term memory networks for time series prediction," in *Procedia CIRP*, vol. 99. Elsevier B.V., 2021, pp. 650–655.

[11] S. Bai *et al.*, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," 3 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

[12] Y. Lin *et al.*, "Temporal Convolutional Attention Neural Networks for Time Series Forecasting," in *2021 IJCNN*, 2021.

[13] Q. Wen *et al.*, "Time Series Data Augmentation for Deep Learning: A Survey," 2 2020. [Online]. Available: http://arxiv.org/abs/2002.12478http://dx.doi.org/10.24963/ijcai.2021/631

[14] 3rd Generation Partnership Project (3GPP), "TS 138 104 - V16.4.0 - 5G; NR; Base Station (BS) radio transmission and reception (3GPP TS 38.104 version 16.4.0 Release 16)," 2020. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[15] ——, "TS 138 211 - V16.2.0 - 5G; NR; Physical channels and modulation (3GPP TS 38.211 version 16.2.0 Release 16)," 2020. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx