Model Adaptation for Lifecycle Management in AI-native 5G/6G Networks

Jewoo Go, Taeje Park, and Wonjin Sung
Department of Electronic Engineering, Sogang University, Seoul, Korea
Email: wsung@sogang.ac.kr

***Abstract*—As part of ongoing standardization efforts in 5G-Advanced for artificial intelligence (AI) and machine learning (ML)-based beam management (BM), model adaptation framework is proposed in this paper as a key lifecycle management (LCM) mechanism to ensure reliable model performance in dynamic wireless environments. The proposed approach enables the network to proactively respond to performance degradation by selecting and activating a more suitable model based on support information and user-side reporting. The superiority of the proposed method over non-adaptive or dataset-only approaches in maintaining beam prediction accuracy is demonstrated by simulation results, highlighting its effectiveness as a robust and scalable solution for future AI-native 6G systems.**

***Index Terms*—lifecycle management, 3GPP, beam management, machine learning, artificial intelligence.**

I. Introduction

THE transition from 5G to 6G marks a paradigm shift toward intelligent, data-centric networks, where connectivity is fundamentally redefined through the native integration of artificial intelligence (AI) and machine learning (ML) technologies [1]. With the emergence of 5G-Advanced, AI-native air interface beam management (BM) has been highlighted as a key use case in 3GPP Release 18 [2]. To address the inefficiencies of traditional BM caused by measurement and reporting overheads, AI/ML-based beam prediction has been proposed to reduce signaling and latency [3].

Only a subset of beams is measured in AI/ML-based beam prediction, and the results are leveraged as input to an AI/ML model for predicting the optimal downlink beam. In this framework, Set A denotes the complete set of candidate downlink beams, whereas Set B refers to a subset used for model input, which may either be a subset of Set A or consist of wider beams. As AI/ML-driven communication systems grow in complexity within 3GPP, model lifecycle management (LCM) has emerged as a critical issue [4].

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) under RS-2024-00397480 (System Development of Upper-mid Band Smart Repeater) and RS-2025-02283217 (Development of AI-based Frequency Interference Analysis and Electromagnetic Wave Prediction Technology).

In particular, 3GPP RAN working group (WG) 2 has explored functionality-based LCM approaches [5]. However, current specifications only define signaling procedures for the exchange of AI/ML functionalities, without addressing mechanisms to handle performance degradation. In such scenarios, model management operations such as retraining on the user equipment (UE), replacing the current model with an alternative, or training a new model architecture become necessary. Accordingly, clearly defined signaling procedures and standardized specifications are required to support these operations.

In this work, we define model adaptation as a model management operation triggered under performance degradation. We propose a signaling procedure that enables performance-aware model management by leveraging network (NW)-side knowledge of the operating environment. Unlike conventional approaches based solely on static signaling, the proposed method ensures model reliability through environment-aware performance evaluation, thereby enabling robust model adaptation. The ongoing discussions on LCM are also addressed in this work, with the need for advanced management strategies beyond simple model exchange signaling being highlighted.

II. System Model

A. Signal Model

In this work, we investigate a multi-user multiple-input multiple-output (MU-MIMO) system where a base station (BS), equipped with a uniform planar array (UPA) of M antennas, simultaneously serves U single-antenna users. The received signal vector at the users, denoted by $\mathbf{y} = [y_1, y_2, \ldots, y_U]^T$, is given by

$$\mathbf{y} = \mathbf{HFs} + \mathbf{n}. \tag{1}$$

Here, $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_U]^H \in \mathbb{C}^{U \times M}$ represents the channel matrix, where each row $\mathbf{h}_u^H$ corresponds to the channel vector between the BS and the u-th user. The matrix $\mathbf{F} \in \mathbb{C}^{M \times U}$ denotes the beamforming matrix applied at the BS. The transmitted signal vector is represented by $\mathbf{s} = [s_1, s_2, \ldots, s_U]^T$, with the elements satisfying the normalization condition $\mathbb{E}[|s_u|^2] = 1$.

And $\mathbf{n}$ denotes the additive white Gaussi
with $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}_U, \sigma_n^2 \mathbf{I}_U)$.

The channel is modeled ba
Saleh–Valenzuela model with L resolv
and is expressed as

$$\mathbf{h}_u = \frac{1}{\sqrt{L}} \sum_{\ell=1}^{L} \alpha_{u,\ell}\, \mathbf{a}(\theta_{u,\ell}, \phi$$

where $\alpha_{u,\ell} \sim \mathcal{CN}(0, \sigma_\ell^2)$ denotes the
gain, and $\theta_{u,\ell}, \phi_{u,\ell}$ represent the zenit
angles of departure, modeled as Lapla
with angular spread σ_θ [7]. The array response vector
$\mathbf{a}(\theta, \phi)$ of a $M_h \times M_v$ UPA is given by

$$\mathbf{a}(\theta,\phi) = \big[\, 1,\, e^{j\kappa d(\sin\phi\sin\theta+\cos\theta)}, \\ \ldots,\, e^{j\kappa d((M_h-1)\sin\phi\sin\theta+(M_v-1)\cos\theta)} \,\big]^T \tag{3}$$

where $\kappa = 2\pi/\lambda$ is the wavenumber, and d is the inter-element spacing [8].

In 5G NR, the beamforming codebook $\mathcal{C}$ is constructed using the Kronecker product of discrete Fourier transform (DFT)-based vectors along horizontal and vertical dimensions [9]. The horizontal and vertical components are defined as

$$\mathbf{x}_h = \frac{1}{\sqrt{M_h}} \left[1, e^{j\frac{2\pi h}{O_h M_h}}, \ldots, e^{j\frac{2\pi(M_h-1)h}{O_h M_h}}\right]^T, \tag{4}$$

$$\mathbf{y}_v = \frac{1}{\sqrt{M_v}} \left[1, e^{j\frac{2\pi v}{O_v M_v}}, \ldots, e^{j\frac{2\pi(M_v-1)v}{O_v M_v}}\right]^T, \tag{5}$$

where $h = 0, \ldots, Q_h - 1$, $v = 0, \ldots, Q_v - 1$, and $Q_h = O_h M_h$, $Q_v = O_v M_v$. The full codebook is given by

$$\mathcal{C} = \{\mathbf{x}_h \otimes \mathbf{y}_v \,|\, h \in [0, Q_h - 1],\, v \in [0, Q_v - 1]\}. \tag{6}$$

B. 3GPP AI/ML for Beam Prediction

Set A represents the complete set of beams available for downlink transmission, while Set B denotes a subset of beams selected from Set A specifically for measurement purposes [4]. The prediction process utilizes the measurement results obtained from Set B to determine the optimal downlink beam within Set A, enabling efficient spatial-domain beam selection. Set A corresponds to the full codebook $\mathcal{C}$, and Set B comprises a subset of codevectors extracted from $\mathcal{C}$. The relationship between Set A and Set B is depicted in Fig. 1, along with the variability in Set B index patterns. The beams included in Set B are shaded in blue in the figure. Assume that the BS employs a uniform planar array with $M = M_v \times M_h = 4 \times 8 = 32$ antenna elements. For the construction of the codebook $\mathcal{C}$, the oversampling factors are set to $(O_h, O_v) = (1, 2)$, resulting in $N_\mathrm{A} = 64$ codevectors.

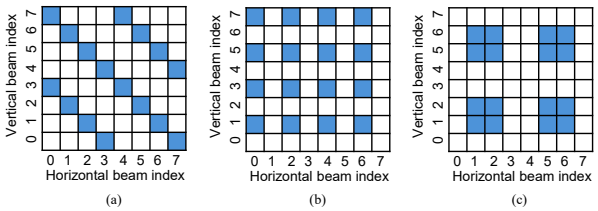

Fig. 1. Set A and Set B configurations for different index pattern.

The model performance is evaluated using two metrics: e_{RSRP}, which is based on the reference signal received power (RSRP), and *Top-K/1 accuracy*. The first metric, e_{RSRP}, quantifies the average difference in L1-RSRP between the true optimal beam index and the index predicted by the model. In addition, Top-K/1 accuracy is used as a key performance indicator (KPI) for prediction accuracy. It is defined as the ratio of instances where the true best beam index is included within the Top-K predicted beam indices. This metric is well-suited for evaluating two-stage beam selection process, where the Top-K beams predicted from Set B are further swept to identify the best beam.

III. Model Lifecycle Management

A. Model Lifecycle Management in 3GPP

LCM of AI/ML-based beam management is currently under active discussion within 3GPP as a critical mechanism to ensure robust model operation throughout various stages, including data collection, model training, inference, monitoring, and transfer [4]. In particular, 3GPP RAN WG1 focuses on specifying signaling procedures for both UE-side and NW-side AI/ML model management, including aspects such as data collection and applicability. Additionally, mechanisms for model performance monitoring and inference result reporting are under consideration [10].

Within 3GPP WG2, three categories of functionalities are defined for managing UE-side AI/ML models: *supported*, *applicable*, and *activated* functionalities [5]. Supported functionalities indicate model operations that the UE can handle and are reported via the *UECapabilityInformation* message. Applicable functionalities are those ready for inference, while activated functionalities are currently in use. The signaling sequence begins with the NW sending a *UECapabilityEnquiry* to request the supported functionalities. The UE responds with a *UECapabilityInformation* message, after which the NW provides inference conditions through the *RRCReconfiguration* message. Based on this configuration and internal model states, the UE identifies applicable functionalities and reports them, and the NW activates the selected ones.

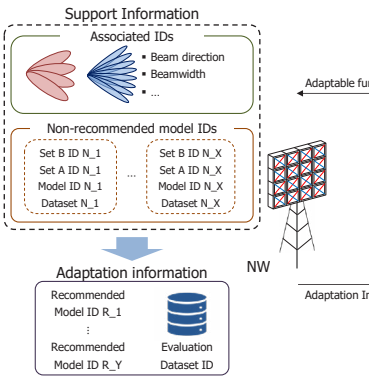

Fig. 2. Transmission of NW adaptation i
of adaptable functionalities for model ada

B. Proposed Model Lifecycle Ma

The LCM framework currently 3GPP does not provide explicit signaling procedures or information exchange to manage AI/ML models under performance degradation. To address this gap, we propose a model adaptation mechanism as an LCM technique designed to ensure performance stability in such scenarios. Model adaptation is defined as a set of management operations triggered when the KPI of beam management, based on the deployed AI/ML model, falls below a predefined threshold. The possible operations include retraining the existing model, switching to an alternative model, or deploying a new model architecture that is better suited to the current conditions. To ensure robust and environment-aware adaptation, the decision process is guided by support information maintained and provided by the NW.

Figure 2 illustrates the NW-side structure of the support information utilized during model adaptation. The UE transmits available model-related identifiers (IDs) to the NW. This includes beam set IDs which contains information about the Set A/B configurations used during model training and inference. Model ID which represents model-specific attributes such as architecture and type and dataset ID which indicates the specific dataset or dataset category used in model training procedure.

Based on the model information reported by the UE, the NW generates and delivers adaptation information by utilizing corresponding support information. The support information includes associated IDs that describe beam-related parameters such as beamwidth and direction, as well as IDs for models that are not suitable for the current environment. The adaptation information transmitted to the UE contains IDs of recommended models that are appropriate for the prevailing conditions, along with an evaluation dataset ID that specifies the dataset used to assess the performance of these recommended models.

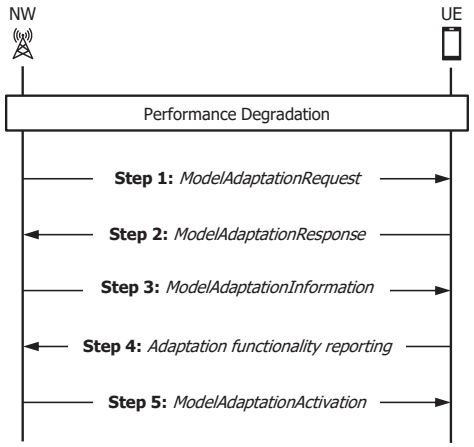

Fig. 3. Proposed signalling procedure.

C. Model Adaptation Signaling Procedure

Figure 3 illustrates the signaling procedure for model adaptation in beam management. When the NW detects performance degradation of the beam management model operating at the UE, the following steps are executed:

- **Step 1:** Upon identifying the current model as inadequate, the NW appends its ID to the non-recommended model list and sends a *ModelAdaptationRequest* message to instruct the UE to deactivate the current model, revert to a legacy beam management scheme, and initiate the model adaptation process.
- **Step 2:** The UE acknowledges the request by replying with a *ModelAdaptationResponse* message, which also includes information about the adaptable functionalities.
- **Step 3:** Based on the support information and the functionalities reported by the UE, the NW selects recommended models and transmits a *ModelAdaptationInformation* message containing the recommended model IDs. If no appropriate candidate exists, only the support information is sent, and the UE is instructed to train a new model that avoids overlap with the non-recommended models.
- **Step 4:** After receiving the adaptation information, the UE evaluates each recommended model using the provided dataset and selects the one that meets the predefined KPI thresholds, such as beam prediction accuracy or throughput. The selected model ID is reported to the NW through a model activation request.
- **Step 5:** The NW finalizes the adaptation process by sending a *ModelAdaptationActivation* message to instruct the UE to activate the selected model.

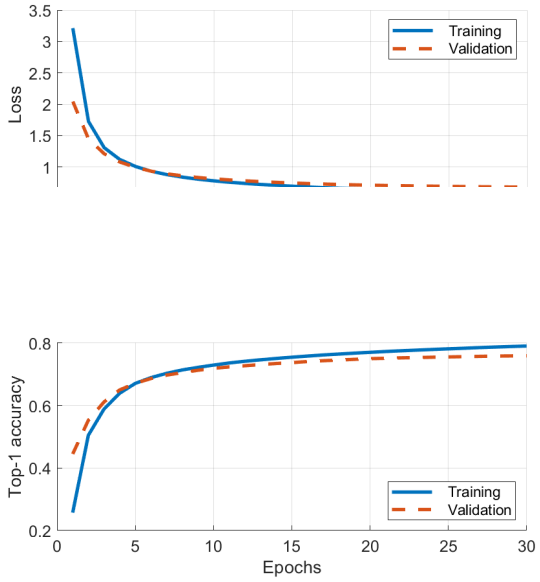

Fig. 5. Average top-1 accuracy curves for training and validation across epochs.

IV. Performance Evaluation

A. Simulation Setup

To assess the proposed framework, we consider a mobility scenario involving $U = 10$ single-antenna UEs moving within a hexagonal cell with an inter-site distance of 200m. In line with recent 3GPP discussions on data collection for UE-side AI/ML models via vendor-operated servers [11], the training dataset is constructed by partitioning the coverage sector into six angular regions centered around the BS. For each region, data samples were collected from 10,000 randomly distributed points. During inference, data collection points are determined based on each UE's mobility profile, assuming a default velocity of 3km/h and a sampling interval of 20ms. The BS is positioned at $(0, 0)$m at a height of $h_{\mathrm{BS}} = 25$m, while UEs are placed at a height of $h_{\mathrm{UE}} = 1.5$m. The BS operates at a carrier frequency of $f_c = 30$GHz within the FR2 band, with adjacent antenna elements spaced by half a wavelength. The array is configured with a downward tilt angle of 20° in elevation and 30° in azimuth. The channel model is based on $L = 6$ multipath components, each with equal average power and an angular spread of $\sigma_L = 5^\circ$, following the link-level simulation assumptions described in [4].

For model adaptation, three identifiers are considered: Set B ID, model ID, and dataset ID. As illustrated in Fig. 1, three distinct Set B patterns are used. Each model is implemented using a deep neural network (DNN), with 3, 4, or 5 hidden layers corresponding to each model ID. The dataset IDs correspond to the six angular training regions defined earlier. These six regions evenly divide the coverage area around the BS based on azimuthal angle. A total of 54 models were created by combining 3 Set B IDs, 3 model IDs, and 6 dataset IDs. We assume that the UE is provisioned with all 54 models stored locally. Model retraining is excluded from the evaluation, as this study focuses on the feasibility and effectiveness of selecting an appropriate model from the pre-stored candidates based on performance-adaptive signaling under degradation scenarios.

The details of the model training process are summarized as follows. The training dataset was partitioned into training and validation sets with a ratio of 9:1. All models were trained with a fixed learning rate of 0.001, LeakyReLU activation applied to each hidden layer, the Adam optimizer, and cross-entropy loss. The batch size was set to 50, and training was performed for 30 epochs. Figures 4 and 5 present the average loss and Top-1 accuracy of the 54 models on the validation and training datasets, respectively. The results show that both metrics begin to saturate after approximately 20 epochs, and the validation performance indicates that no overfitting occurred during training.

The LCM procedure used in the experiment is described as follows. The UE-side model continuously performs inference over 5,000 time steps. To monitor performance degradation, the system periodically evaluates the model at intervals of T_{LCM}. A degradation event is triggered when the beam prediction accuracy metric, e_{RSRP}, drops below a predefined threshold γ_{RSRP}, relative to the reference value measured at the time of initial model activation. If the model is replaced, the reference value is updated based on the first activation of the newly selected model. Upon detecting degradation, model adaptation is initiated. The evaluation dataset used in the adaptation process consists of 5,000 randomly sampled locations within the current sector where the UE is situated. Each recommended model is evaluated using this dataset, and the model exhibiting the best performance is selected for subsequent operation.

B. Simulation Result

Figure 6 illustrates the variation in e_{RSRP} over 5,000 inference time steps for each LCM approach. In this experiment, T_{LCM} is set to 1 second, and γ_{RSRP} is set to 5dB. The *only dataset ID* method refers to a scheme in which model adaptation is performed solely based on the dataset ID corresponding to the UE's current region. In contrast, the *non-LCM* method uses the initially selected model throughout all inference steps without any LCM. As shown in the results, the proposed model adaptation approach maintains consistent performance over time, with minimal fluctuations in error values. On the other hand, the only dataset ID

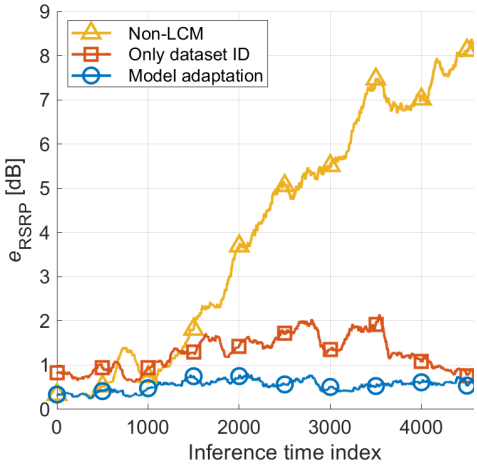

Fig. 6. Variations in e_{RSRP} over the inference time

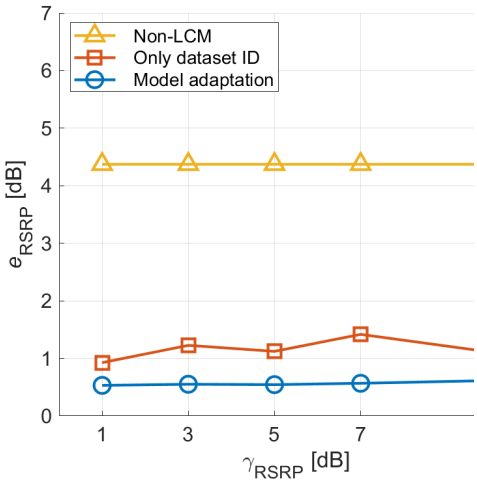

Fig. 7. Impact of γ_{RSRP} on e_{RSRP} for different LCM strategies.

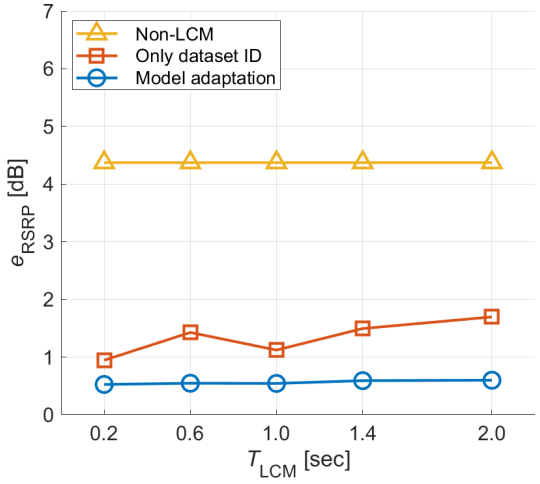

Fig. 8. Impact of T_{LCM} on e_{RSRP} for different LCM strategies.

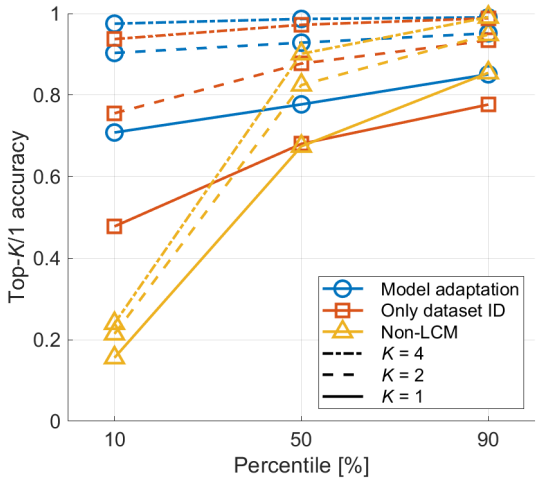

Fig. 9. Comparison of Top-K/1 accuracy for varying K.

method exhibits higher variability in e_{RSRP},y in , and the non-LCM method shows a gradual increase in error as time progress

Figures 7 and 8 illustrate the variation in e_{RSRP} with respect to different values of the degradation threshold γ_{RSRP} and the LCM evaluation interval T_{LCM}. The results from both experiments exhibit similar trends. The non-LCM method maintains a consistently high error value above 4 dB regardless of the parameter changes, as it does not incorporate any LCM. In contrast, the only dataset ID method shows sensitivity to the parameter settings, with increasing error values observed when the threshold or evaluation interval is increased. Meanwhile, the proposed model adaptation approach demonstrates robust performance across all parameter settings, indicating its resilience to variations in LCM configurations.

Figure 9 presents the Top-K/1 accuracy performance for the 10th, 50th, and 90th percentile users in the inference dataset. The proposed model adaptation consistently outperforms the other approaches across all values of K. Notably, for $K = 1$, it achieves over 70% accuracy for the 10th percentile users, surpassing the only dataset ID method by more than 20% and the non-LCM method by over 50%. For the 90th percentile users, however, all three methods exhibit similar performance.

V. Conclusion

In this study, we propose a model adaptation mechanism as a LCM technique to sustain the performance of AI/ML-based wireless communication models, a topic of increasing importance in the evolution from 5G to 6G. The proposed mechanism enables the network to deliver adaptation information upon detecting performance degradation, thereby facilitating the selection of a more robust model tailored to the user environment. Simulation results demonstrate that the proposed

approach consistently outperforms both non-LCM and alternative LCM strategies. Although the dataset used in this study is synthetic and the evaluation scenario simplified, the results provide a clear proof-of-concept for the feasibility of model adaptation. Future work will focus on validating the framework with standardized or real-world datasets, quantifying signaling overhead and latency, and addressing scalability concerns such as model storage at the UE. Further extensions include integrating model retraining into the LCM framework and expanding applicability beyond beam management to other AI/ML-driven use cases in wireless systems.

REFERENCES

[1] W. Chen *et al.*,"5G-Advanced toward 6G: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1592-1619, June 2023.

[2] 3GPP TSG RAN WG1#94e, RP-213599, *New SI: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface*, Dec. 2021.

[3] Q. Xue, J. Guo, B. Zhou, Y. Xu, Z. Li, and S. Ma, "AI/ML for beam management in 5G-Advanced: A standardization perspective," *IEEE Vehicular Technology Magazine*, vol. 19, no. 4, pp. 64-72, Dec. 2024.

[4] 3GPP TR 38.843, V18.0.0, *Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface (Release 18)*, Dec. 2023.

[5] 3GPP TSG RAN WG2#127 R2-2407848, *LS on Applicable Functionality Reporting for Beam Management UE-sided model*, Maastricht, Netherlands, Aug. 2024.

[6] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, Feb. 19871.

[7] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499-1513, Mar. 2014.

[8] C. A. Balanis, *Antenna Theory: Analysis and Design*, 3/e. Wiley, 2005.

[9] 3GPP TS 38.214, V18.3.0, *NR; Physical Layer Procedures for Data (Release 18)*, June 2024.

[10] 3GPP TSG RAN WG1#120, R1-2501595, *FL Summary #6 for AI/ML in Beam Management*, Athens, Greece, Feb. 2025.

[11] 3GPP TSG RAN WG1#117, R1-2405505, *FL Summary #5 for Other Aspects of AI/ML model and Data*, Fukuoka, Japan, May 2024.