Fast Best Beam Prediction and Overhead Reduction for 6G Networks: A Deep Learning Approach

Jalal Jalali†, Juan Roa†, Yifei Song‡, Renjian Zhao†, and Baoling Sheen†
†Wireless Research and Standards, Futurewei Technologies, Bridgewater, NJ 08807, USA
‡ Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060 USA
Email: {jfaghih, jroa, rzhao, bsheen}@futurewei.com and yifeisong@vt.edu

***Abstract*—Beam management (BM) plays a crucial role in maintaining reliable communication links in highly dynamic scenarios. To enhance BM performance, the 3rd Generation Partnership Project (3GPP) is actively exploring the use of artificial intelligence (AI) and machine learning (ML) for beam prediction in the evolution toward sixth-generation (6G) communications. The main goals of these 3GPP-based standard studies are to minimize the overhead from reference signals (RSs) and to reduce the number of beam sweepings at the user equipment (UE), which arise due to frequent beam measurements caused by UE movement and rotation. This paper delves into an AI/ML algorithm design that supports spatial domain beam prediction tailored for BM in 6G. This includes forecasting the optimal beam (pairs) and anticipating beam changes. Simulations are based on a data-driven strategy that uses RS receive power (RSRP) measurements as input for fast beam pair prediction with an advanced convolutions neural network (CNN) architecture. Results indicate that the proposed AI/ML model outperforms conventional BM techniques, reducing beam sweeping overhead and thereby validating the AI/ML's potential in BM. Additionally, our proposed algorithm achieves up to 40.58% higher beam prediction accuracy and improves the mean RSRP difference of the predicted best beam pair by up to 2.89 dB.**

***Index Terms*—Sixth-generation (6G), Machine Learning (ML), millimeter-Wave (mmWave), Spatial Domain, Beam Management (BM).**

I. Introduction

THE exploration and subsequent implementation of millimeter wave (mmWave) and terahertz (THz) frequencies in communication frameworks present a promising frontier, particularly because of the high data throughput they can achieve [1]. This is attributed to the vast bandwidth these frequencies have access to [2]. However, the potential of such frequencies is also accompanied by significant challenges. One primary concern is the pronounced free-space path loss inherent to these frequency bands. To counteract this drawback, communication systems are increasingly turning to beamforming methodologies, using large antenna arrays at both transmitting and receiving points, which aid in enhancing the signal's strength and directionality [3], [4]. Central to this strategy is the concept of beam management (BM), which revolves around the meticulous process of determining the optimal beam pair to establish and sustain a robust communication link, a process discussed in depth by [5], [6]. Yet, it is imperative to note that the efficiency of BM is put to the test, especially in highly dynamic wireless environments. In urban areas with high mobility, the channel's conditions can fluctuate rapidly [7]. Such volatile conditions amplify the complexities of ensuring a seamless BM [8].

Building on the complexities inherent in mmWave BM, especially in dynamic environments, there emerges a clear and pressing need for more adaptive and intuitive mechanisms to ensure robust communication. Turning to Artificial Intelligence (AI) and Machine Learning (ML), AI/ML-integrated BM stands out as a promising avenue [9]. As we transition toward sixth-generation (6G) communications, AI/ML-assisted BM is proposed to offer both accuracy and adaptability in beam prediction [10]. These AI/ML technological paradigms provide transformative potential in reshaping how BM is approached, constantly evolving as standard impacts in the 3rd Generation Partnership Project (3GPP), a forefront entity in shaping communication standards [11]. Stepping away from traditional methods like exhaustive beam switching and pilot-based beamforming [12], which may falter in adaptability, AI/ML strategies present real-time learning capabilities. They not only assure the consistent choice of beam pairs but also dynamically adapt to the fluid challenges inherent in mmWave communication channels [13].

3GPP Release 18 has set up a study item to explore the potential of AI/ML-based solutions for identified use cases [14] Within this progressive release, BM has emerged as a prime example of integrating AI/ML into the New Radio (NR) Air-interface [15]. Within the scope of AI/ML-based BM, spatial-domain beam prediction has been identified as one of the representative sub-use cases. The essence of this approach is profound yet elegant: by utilizing an AI/ML model, the User Equipment (UE) needs to measure only a limited set of transmit-receive (TX-RX) beam pairs. Thereafter, either the UE or the next generation Base Station (gNodeB) can predict the optimal beam pair, a feat achieved without the overhead of requiring exhaustive measurements. AI/ML-based BM algorithms aim to minimize beam measurement overhead while simultaneously amplifying the accuracy of beam predictions, all under the same operational constraints. Leveraging vast amounts of data from communication networks, AI/ML models can learn the beam behavior patterns to the ever-changing environmental dynamics in higher frequencies.

To this end, several works have also been developed to study the AI/ML-based BM performance in mmWave future networks [16]–[20]. For instance, a deep learning-assisted BM prediction method was proposed in [16] to assess a subset of downlink beam pairs to reduce overhead compared with an exhaustive search while determining the optimal pair by predicting the reference signal receive power (RSRP) across all

979-8-3503-8741-4/24/$31.00 ©2024 IEEE

beams. Nguyen *et al.* introduced a novel learning framework for user-specific beam selection and transmit power optimization to minimize costs in unknown channels, tackling missing data issues using the long-short term memory for temporal input processing [17]. Drawing on spatial channel characteristics from the sub-6 GHz band to mitigate mmWave beam training overhead, Alrabeiah *et al.* crafted a deep learning model and empirically assessed its capability in beam/blockage prediction [18]. Similarly, Sim *et al.* introduced a deep learning-based beam selection compatible with the 5G NR standard, utilizing sub-6 GHz channel information and employing a deep neural network (DNN) to estimate the power delay profile of a sub-6 GHz channel as its input [19]. Echigo *et al.* presented a deep learning-driven, low-overhead analog beam selection strategy using super-resolution technology, where DNNs estimate beam quality from partial measurements, addressing the challenges of swift beam alignment for rapid wireless link establishment in codebook-based beamforming scenarios [20]. However, none of the aforementioned research works considered a *RSRP-enabled AI/ML-based mmWave spatial domain beam pair prediction using convolutional neural networks (CNNs)*.

In this paper, our focus is on AI/ML modeling and its performance evaluation for spatial-domain beam prediction in 6G networks. The key contributions are as follows:

- We formulate the BM process as a classification problem by training a model to learn to identify the best BM pair on each time step. This approach incorporates BM measurements from a small subset of all available beam pairs based on the mmWave 3GPP-compliant downlink signaling and measurement framework.
- We explore AI/ML-based spatial domain beam pair prediction technique(s) using an advanced CNN architecture for broader applicability and computational efficiency for BM in future networks. The proposed architecture is compatible with various beam measurement configurations while considering how to reduce the training and storage requirements typically associated with conventional CNNs.
- Our numerical analyses validate the effectiveness of our proposed CNN model, showing that beam selection accuracy and L1-RSRP difference performance outperform traditional approaches, such as sparse beam sweeping, in determining the best-serving beam pair.

The remainder of this paper is organized as follows. Section II discusses the system model and performance metric. The AI/ML-based BM algorithm is described in Section III. Section IV presents the simulation setup and provides a comprehensive explanation of the proposed deep learning-based BM strategy and results. Finally, Section V concludes the paper and summarizes the key findings.

II. System Model and Performance Metric

We study an mmWave multi-cell network with a gNodeB and several UEs in each cell. There are K users randomly positioned in each tri-sector cell, where the set of users is

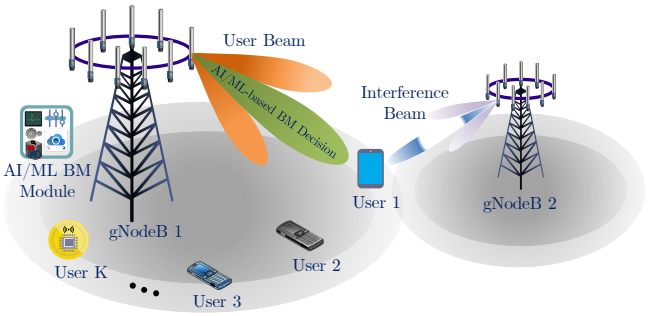

Fig. 1. Spatial domain BM in a multi-cell multi-user mmWave network.

denoted by $\mathcal{K} = \{1, \ldots, K\}$. As illustrated in Fig. 1, both the gNodeB and the UEs possess multiple beams, where the gNodeB of the neighboring cell interferes with the intended user beam. For enhanced coverage and data transfer rates, the optimal beam pair should be chosen for each UE and gNodeB during data transmission. In the conventional downlink BM method, the correct beam pairing from the N_{TX} beams at the gNodeB end and N_{RX} beams at the UE end is determined based on reference signal measurements. The gNodeB sends out each RS in a distinct beam direction, allowing the beam's quality to be inferred from the received power of the corresponding RS on the UE's end. The UE can report the quality of the assessed beams back to the gNodeB. From the beams that are reported, the gNodeB will then choose the most suitable one. The reported data comprises the RS identification, and its associated RS received power (RSRP) or the signal-to-interference-plus-noise-ratio (SINR) of the indicated RS.

The mmWave channel response for the k-th UE is derived using a 3GPP-compliant 3D geometry-based stochastic channel model generator called the QUAsi Deterministic RadIo channel GenerAtor (QuaDRiGa) [21]. Accordingly, the mmWave channel can be represented as follows:

$$\boldsymbol{H}_k = \sum_{l \in \mathcal{L}} \xi_{l,k} \boldsymbol{a}_{r,k,l}(\theta, \phi) \boldsymbol{a}_{t,k,l}^H(\theta, \phi) e^{-j2\pi\tau_{l,k} f}, \forall k \in \mathcal{K}, \quad (1)$$

where $\mathcal{L} = \{1, \ldots, L\}$ is the set of channel's multi-path components with L as its set cardinally. Besides, $\xi_{l,k}$, $\tau_{l,k}$, and f represent the complex gain for sub-path l of UE k, delay values for sub-path l of UE k, and the subcarrier frequency, respectively. Furthermore, $\boldsymbol{a}_{t,k}$ and $\boldsymbol{b}_{r,k}$ denote the transmitter and receiver array responses based on sub-path l's of the k-th UE's elevation and azimuth angles of arrival and departure (θ, ϕ). The array responses for both gNodeB and UE, sized as $N_{\text{TX}} = N_{t,x} N_{t,y} N_{t,z}$ ($N_{t,x}$, $N_{t,y}$, and $N_{t,z}$ are x, y, and z of N_{TX}) and $N_{\text{RX}} = N_{r,x} N_{r,y} N_{r,z}$ ($N_{r,x}$, $N_{r,y}$, and $N_{r,z}$ are x, y, and z of N_{RX}), respectively, are as follows:

$$\boldsymbol{a}_{i,k,l}(\theta, \phi) = \boldsymbol{b}_{i,k,l}(\theta, \phi) \odot \boldsymbol{g}_{i,k,l}(\theta, \phi), \quad \forall i \in \{t, r\}, \forall k \in \mathcal{K}, \forall l \in \mathcal{L}, \quad (2)$$

where $\odot$ is the Hadamard product. Moreover, $\boldsymbol{g}_{t,k,l} \in \mathbb{C}^{N_{\text{TX}} \times 1}$ and $\boldsymbol{g}_{r,k,l} \in \mathbb{C}^{N_{\text{RX}} \times 1}$ are the linear gain for TX and RX antennas, respectively. Finally, $\boldsymbol{b}_{i,k,l}$ is defined as:

$$\boldsymbol{b}_{i,k,l}(\theta, \phi) = \frac{\boldsymbol{b}_{i,k,l}^z(\theta) \otimes \boldsymbol{b}_{i,k,l}^y(\theta, \phi) \otimes \boldsymbol{b}_{i,k,l}^x(\theta, \phi)}{\sqrt{N_{i,x} N_{i,y} N_{i,z}}}, \quad \forall i \in \{t, r\}, \forall k \in \mathcal{K}, \forall l \in \mathcal{L}, \quad (3)$$

where $\otimes$ denotes the Kronecker product, and the components $\boldsymbol{b}^x_{i,k,l} \in \mathbb{C}^{N_x \times 1}$, $\boldsymbol{b}^y_{i,k,l}\mathbb{C}^{N_y \times 1}$, and $\boldsymbol{b}^z_{i,k,l}\mathbb{C}^{N_z \times 1}$ are given as:

$$\boldsymbol{b}^x_{i,k,l}(\theta,\phi) = [1, e^{j\pi \sin\theta_{i,k,l}\cos\phi_{i,k,l}}, \ldots, e^{j\pi(N_x-1)\sin\theta_{i,k,l}\cos\phi_{i,k,l}}]^T \tag{4}$$

$$\boldsymbol{b}^y_{i,k,l}(\theta,\phi) = [1, e^{j\pi \sin\theta_{i,k,l}\sin\phi_{i,k,l}}, \ldots, e^{j\pi(N_y-1)\sin\theta_{i,k,l}\sin\phi_{i,k,l}}]^T \tag{5}$$

$$\boldsymbol{b}^z_{i,k,l} = [1, e^{j\pi\cos\theta_{i,k,l}}, \ldots, e^{j\pi(N_z-1)\cos\theta_{i,k,l}}]^T, \quad \forall i \in \{t,r\}, \forall k \in \mathcal{K}, \forall l \in \mathcal{L}. \tag{6}$$

The 3GPP BM framework consists of layer 1 (L1) and layer 2 (L2) procedures utilizing beam actions like sweeping, measurement, determination, and reporting to align gNodeB's and UE's beams [22]. These procedures, unofficially termed P1, P2, and P3 in technical discussions [22], encompass gNodeB transmit beam selection using broad beams (P1), refining that selection with narrower beams (P2) and post-gNodeB selection, allowing the UE to find the optimal receive beam (P3) for the gNodeB transmit beam identified in P2. However, the standard does not necessitate these procedures' implementation. The assumed beamwidth relationship between P1 and P2 beams is not standard-specified but is industry-accepted. Their application should be scenario-tailored to minimize latency and signaling overhead. In both P1 and P3 instances, the signal received by k-th UE can be expressed as:

$$y_k = \boldsymbol{\rho}_k^H \boldsymbol{H}_k \boldsymbol{\omega} s_k + \boldsymbol{\rho}_k^H \boldsymbol{n}, \forall k \in \mathcal{K}, \tag{7}$$

where $\boldsymbol{H}_k \in \mathbb{C}^{N_r \times N_t}$ represents the mmWave channel matrix as expressed in (1). The gNodeB's beamforming vector, $\boldsymbol{\omega} \in \mathbb{C}^{N_{\text{TX}} \times 1}$, encompasses the analog phase shifts for a beam, maintaining a constant modulus of $\frac{1}{\sqrt{N_{\text{TX}}}}$, which spatially processes the transmitted signal s_k. At the k-th UE, this signal is captured through a beam dictated by the analog phase shifts in the beamforming vector $\boldsymbol{\rho}_k \in \mathbb{C}^{N_{\text{RX}} \times 1}$, sustaining a constant modulus of $\frac{1}{\sqrt{N_{\text{RX}}}}$. Lastly, $\boldsymbol{n} \in \mathbb{C}^{N_{\text{RX}} \times 1} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_{\text{RX}}})$ is the receiver's noise, interpreted as a complex additive white Gaussian noise vector with a variance σ^2.

To ensure an optimal beam pair that adjusts to changing channel conditions, regular beam measurements and reporting are essential. Identifying the best beam pair involves thoroughly measuring $|\text{Set A}| = N_{\text{TX}} \times N_{\text{RX}}$ beam pair combinations. However, due to the typically large number of transmit beams at the gNodeB in mmWave systems, it is impractical to measure the quality of every beam pair exhaustively because of the extensive measurement overhead. Instead, a subset of these beams, referred to as Set B, where Set B $\subseteq$ Set A, is chosen for beam measurement, that is $m = |\text{Set B}|$ beams from all possible $N_{\text{TX}} \times N_{\text{RX}}$ transmit-receive beams. The right beam pair is then selected from the $p \times q$, $\forall p \leq N_{\text{TX}}, q \leq N_{\text{RX}}$ beam pair combinations. Still, there is a risk that the best beam pair might be overlooked since it might not be included in the measured set. In our study, we employ an AI/ML-driven approach to address the challenge of selecting the best beam pair, given the known quality of some beam pairs. The problem can be expressed as computing:

$$\arg\max \; \boldsymbol{p}_{\text{Set B}} \left([r_{1,1}, \ldots, r_{N_{\text{TX}},N_{\text{RX}}}]^T \right), \tag{8}$$

where $r_{u,v}, 1 \leq u \leq N_{\text{TX}}, 1 \leq v \leq N_{\text{RX}}$ stand for the RSRP associated with the beam pair comprised of the u-th Tx and the v-th Rx beam. The $\arg\max$ returns the beam pair index with the highest RSRP within Set A. The AI/ML model, defined by the parameters $\boldsymbol{p}_{\text{Set B}}$, aims to predict the RSRP for all Set A beam pairs, using the observed RSRP of the Set B pairs as a reference. Instead of predicting the RSRPs for all Set A beam pairs, which requires more effort in collecting the training data, we treat this as a classification problem, which predicts the position of the optimal beam pair in Set A. The AI/ML model aims to predict the probability of being the optimal beam for each Set A beam pair. During the model training, the optimizer tries to minimizes the cross-entropy loss, $\mathcal{H}(\boldsymbol{f}, \boldsymbol{g})$, which is the distance between what the model believes the output labels should be, $\boldsymbol{f}$, and what the actual labels are, $\boldsymbol{g}$, according to:

$$\mathcal{H}(\boldsymbol{f}, \boldsymbol{g}) = -\frac{1}{V} \sum_{u=1}^{U} \sum_{v=1}^{V} f_{u,v} \log(z_\theta(g_v, u)), \tag{9}$$

where V is the number of training examples, U is the number of classes, g_v is the input for training example v, $f_{u,v}$ represents the target label for training example v for class u, and finally z_θ is the model with neural network weights θ.

III. AI/ML-Driven BM Design and Proposed Solution

In this section, we delve into the AI/ML-centric BM process, followed by an introduction to the proposed AI/ML-driven scheme. Our approach hinges on the CNN to execute spatial domain beam predictions. We contemplate various beam measurement pattern configurations, that is, distinct Set B values, to anticipate the quality across all beam pairs. Initially, models rooted in CNN are trained for each distinct configuration and are trained as a "global" model combining data from all UEs in the dataset. The proposed CNN architecture is pictured in Fig. 2 and consists of the following:

- Input Layer contains the |Set A| normalized RSRP values converted into a 3D grid with shape 16x16x1 with one value per BM pair. For Set B elements, the measured RSRP value is used for training, whereas for all other elements, the lowest observed RSRP value is used.
- A sequence of four convolutional layers performs feature extraction with increasing channel depth up to 4x4x512. These layers use the same activation function: rectified linear unit function.
- A sequence of fully connected Layers of decreasing size from 8192 to 256 to compute the classification. A dropout layer precedes each layer to improve the model's generalization capability.
- Output Layer: Uses the softmax activation function to predict the top beam pair by converting the last computed 256 vectors into probabilities associated with each entry in Set A.

To train this classifier, we use the Categorical Cross-Entropy loss function, which is well-suited for multi-class classification problems [23]. The Adam optimizer is used for gradient-based optimization of the network's weights, combining the advantages of both AdaGrad and RMSProp methods [24].

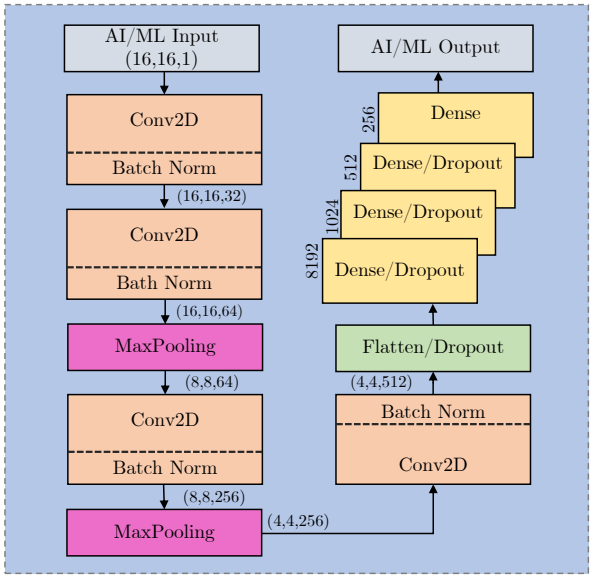

Fig. 2. Our proposed CNN architecture for AI/ML-based mmWave BM.

This optimizer is known for its computational efficiency and capability to handle large datasets effectively. For the evaluation metrics, we focus on Accuracy — more precisely, the Top-K[1] Accuracy. The Accuracy metric measures the proportion of correctly classified instances to the total instances in the dataset, providing a straightforward assessment of the model's performance. Meanwhile, Top-K Accuracy considers the model's prediction to be correct if the true label is among the top K predicted classes. This latter metric is particularly useful when we are interested in identifying not just the most likely beam pair but also the next-best alternatives. By employing these settings, we aim to achieve a robust and accurate beam-pair prediction model that can adapt to various network conditions.

IV. Simulation Results

We introduce the performance evaluation derived from the discussions in Sections II and III for AI/ML-based beam prediction in the spatial domain. We consider 19 cells, with each cell comprising 3 sectors. The distribution of UEs is such that 80% are indoors. We employ the urban-macro channel model, as defined in 3GPP [26], for generating the dataset. The gNodeB is equipped with 32 TX antennas, while the UEs have 8 RX antennas. The transmit power is set to 20 dBm for a bandwidth of 40 MHz. The arrangement for the TX antennas is represented by the sequence $[4\ 8\ 2\ 1\ 1]$ with a horizontal and vertical spacing of 0.5λ. On the UE side, the antenna configuration follows $[1\ 4\ 2\ 1\ 1]$ pattern with the same uniform horizontal and vertical spacing. The working frequency is anchored at 30 GHz within the mmWave spectrum, incorporating a sub-carrier spacing of 120 kHz. We also assume that L1-RSRP for all TX-RX beam pairs are perfectly available to generate the training data. For every beam measurement setup, we produce a total of 500000 data samples. We allocate 90% of these samples for training the model, reserving the remaining 10% for testing. We now aim to assess the effectiveness of our proposed CNN model by:

- Examination of performance based on varying numbers of beam pairs inputted into the AI/ML model.
- Analysis of performance differences using various beam patterns: fixed, random, and pre-configured.
- Evaluating Top-$K/1$ beam prediction accuracy, assessing the L1-RSRP difference of the Top-$K/1$ predicted beam pair(s), and comparing this to the L1-RSRP difference from the sparse beam sweeping (traditional) method.

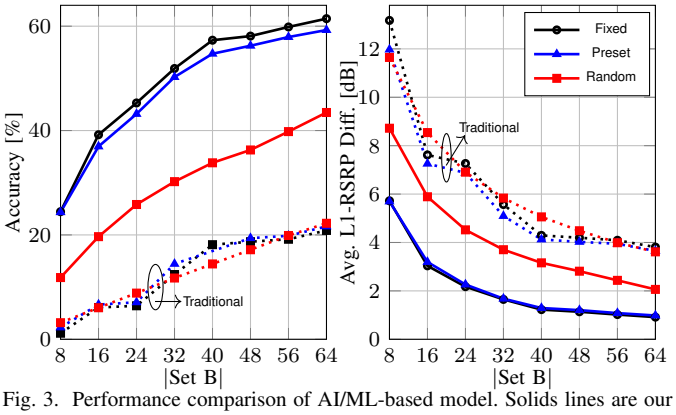

Fig. 3. Performance comparison of AI/ML-based model. Solids lines are our proposed CNN architecture with 9.724 million trainable parameters, and the dashed lines are based on the traditional sparse beam sweeping approach.

We begin by contrasting the performance of the AI/ML model over diverse Set B sizes and patterns. Table I contains the performance evaluation results when using fixed beam pattern sampling with various numbers of Set B beam pairs as input to the AI/ML model. Subsequently, Table II is for pre-configured beam patterns, while Table III shows results when using a random beam patterns sampling approach. It is evident that as Set B lengths grow, the accuracy improves. Conversely, the average L1-RSRP difference for top-$K/1$ diminishes with increasing Set B lengths. This trend remains consistent across all three sampling approaches. Fig. 3 visualizes this effect for Top-1 accuracy and L1-RSRP difference of the Top-1 predicted beam pair in our proposed CNN architecture based on Tables I-III. As seen in Fig. 3, the fixed beam pattern approach surpasses both the pre-configured and random methods, delivering higher accuracy and a reduced L1-RSRP difference using the proposed AI/ML model. Yet, the margin of improvement between the fixed and pre-configured patterns is small. Besides, although the performance of the traditional beam sweeping approach, i.e., sparse beam sweeping (in which the beam pair in Set B with the highest L1-RSRP is chosen as the best beam pair) correlates with the Set B length, there seems to be no discernible connection with various Set B sampling patterns. Furthermore, Fig. 3 shows our proposed data-driven strategy achieves up to 40.58% higher beam prediction Top-1 accuracy and decreases the average L1-RSRP difference of the Top-1 predicted beam pair by 2.89 dB compared to the traditional sparse beam sweeping method.

To understand the distribution of L1-RSRP differences for all test samples, we plot the corresponding cumulative

[1]We note Top-1 is defined as the percentage of "the Top-1 genie-aided beam is Top-1 predicted beam," and Top-$K/1$ is the percentage of "the Top-1 genie-aided beam is one of the Top-K predicted beams," where $K > 1$ [25].

TABLE I
FIXED BEAM PATTERN SAMPLING (TOTAL BEAM PAIRS = 256)

Set B Length	Accuracy (%)					Avg. L1-RSRP difference of Top-$K/1$ predicted beam				
	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1
8	24.45	38.22	54.52	64.31	71.25	5.72	3.86	2.26	1.54	1.13
16	39.20	56.56	72.61	80.65	85.62	3.04	1.81	0.90	0.56	0.38
24	45.29	63.37	78.83	85.87	89.79	2.17	1.22	0.56	0.33	0.23
32	51.89	70.0	83.63	89.24	92.41	1.65	0.87	0.38	0.23	0.15
40	57.31	75.37	87.82	92.28	94.68	1.23	0.61	0.24	0.14	0.09
48	58.10	76.16	88.10	92.58	94.93	1.14	0.56	0.22	0.13	0.08
56	59.85	77.73	89.22	93.39	95.49	1.02	0.48	0.19	0.11	0.07
64	61.44	79.29	90.32	94.28	96.22	0.92	0.42	0.16	0.09	0.06

TABLE II
PRE-CONFIGURED PATTERN SAMPLING (TOTAL BEAM PAIRS = 256)

Set B Length	Accuracy (%)					Avg. L1-RSRP difference of Top-$K/1$ predicted beam				
	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1
8	24.35	38.41	54.53	64.37	71.12	5.70	3.83	2.25	1.55	1.14
16	36.95	54.3	70.62	78.69	83.7	3.19	1.93	0.99	0.62	0.44
24	43.18	61.92	77.41	84.46	88.65	2.26	1.27	0.60	0.37	0.25
32	50.25	68.63	82.56	88.47	91.76	1.68	0.88	0.39	0.23	0.16
40	54.72	73.53	86.18	91.20	93.95	1.29	0.65	0.28	0.16	0.10
48	56.25	74.88	87.21	92.08	94.62	1.20	0.58	0.24	0.13	0.09
56	57.92	76.36	88.53	93.09	95.26	1.08	0.51	0.20	0.11	0.07
64	59.26	77.93	89.47	93.82	95.93	0.98	0.45	0.18	0.09	0.06

TABLE III
RANDOM BEAM PATTERN SAMPLING (TOTAL BEAM PAIRS = 256)

Set B Length	Accuracy (%)					Avg. L1-RSRP difference of Top-$K/1$ predicted beam				
	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1	Top-1	Top-2/1	Top-4/1	Top-6/1	Top-8/1
8	11.83	21.54	34.96	44.85	52.30	8.72	6.58	4.45	3.35	2.69
16	19.67	33.86	50.29	60.21	67.34	5.89	4.10	2.49	1.76	1.32
24	25.83	41.68	58.87	68.43	74.77	4.52	3.03	1.73	1.18	0.87
32	30.21	48.09	65.25	73.83	79.37	3.70	2.35	1.27	0.85	0.61
40	33.82	52.22	69.04	77.49	82.54	3.16	1.96	1.04	0.67	0.48
48	36.28	55.59	71.92	79.78	84.44	2.81	1.67	0.86	0.56	0.40
56	39.80	58.91	74.89	82.35	86.63	2.44	1.43	0.70	0.44	0.31
64	43.46	62.91	78.00	84.87	88.89	2.06	1.16	0.55	0.34	0.23

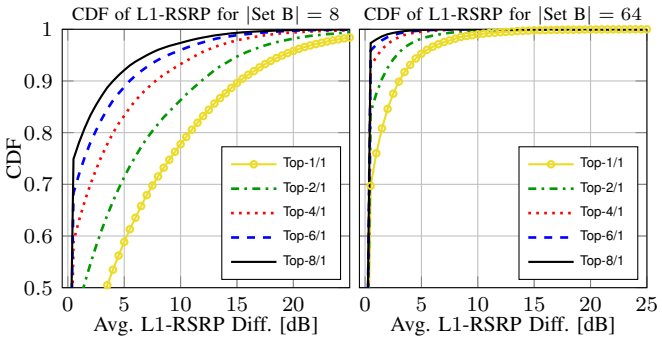

Fig. 4. CDF of L1-RSRP difference for fixed input beam pattern.

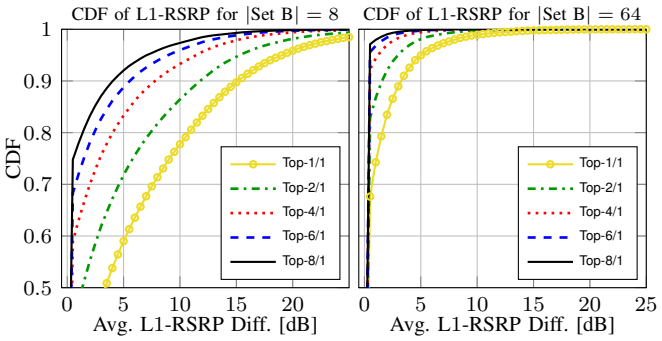

Fig. 5. CDF of L1-RSRP difference for pre-configured input beam pattern.

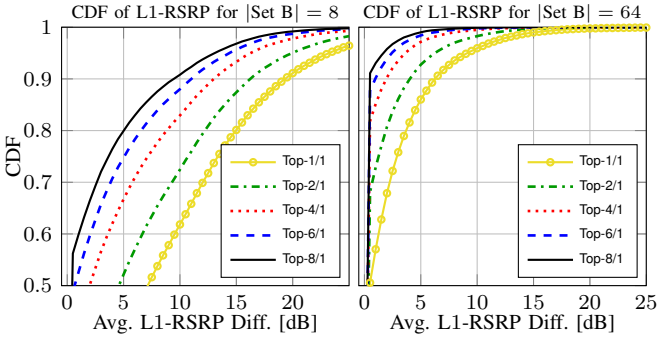

Fig. 6. CDF of L1-RSRP difference for random input beam pattern.

distribution function (CDF) as a function of average L1-RSRP difference as depicted in Fig. 4 (for fixed input beam pattern), Fig. 5 (for pre-configured- input beam pattern), and Fig. 6 (for random input beam pattern). From the CDF figures, we can see that given an equal quantity of training data, the utilization of fixed beam patterns as input yields superior performance in comparison to either random or pre-configured beam patterns. This insight corroborates our earlier findings from Fig. 3. Such a trend might arise because leveraging an increased number of different beam patterns as input to an AI/ML model necessitates more training data. This ensures the model can proficiently discern the mapping function between all the input beam patterns and the corresponding optimal beam pairs (outputs) in Set A.

After evaluating the Top-$K/1$ prediction accuracy and average L1-RSRP difference of Top-1 (and Top-K) predicted beam pair(s) across various Set B lengths, a natural next step is to determine proper Set B length and K value, which may help the gNB to decide the number of L1-RSRP measurements it may request the UE to measure in the next round. Hence, we analyzed the L1-RSRP difference between the ideal L1-RSRP of the Top-1 genie-aided beam and that of the Top-K genie-aided beams in the dataset. This combined metric provides a

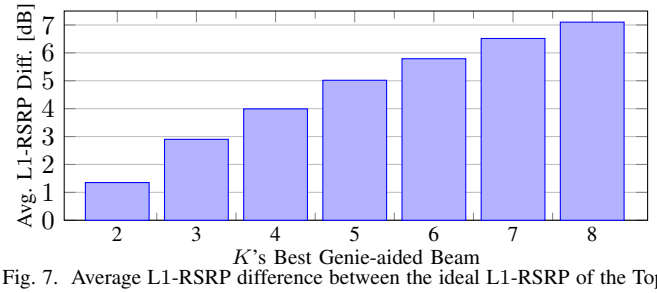

Fig. 7. Average L1-RSRP difference between the ideal L1-RSRP of the Top-1 genie-aided beam and the ideal L1-RSRP of the Top-K genie-aided beams in the dataset

clearer picture of performance. As an illustration, in Fig. 7, the true L1-RSRP difference between the Top-1 beam and the Top-3 beam in our dataset is 2.9 dB, and the difference between the Top-1 beam and the Top-4 beam is 3.9 dB. Hence, the value for the average L1-RSRP difference of the Top-1 predicted beam pair prediction performance below 3.9 dB may be considered decent. Jointly considering this information and the results depicted in (Table I) when using a fixed Set B sampling pattern, we may select $|\text{Set B}| = 32$ and choose 8 as the value for K as it indicates $> 92\%$ probability the true optimal beam pair is within the Top-8 predicted beam pairs and the average L1-RSRP difference of the Top-1 predicted beam pair is 1.65 dB. If gNB chooses to perform another round of beam sweeping for the Top-8 predicted beam pairs, then it would further reduce the average L1-RSRP difference of the Top-1 predicted beam pair to 0.15 dB.

V. Conclusion and Future Work

In this paper, we explored AI/ML-based spatial beam prediction, focusing on both performance and the intricacies of AI/ML model architecture. Our analysis revealed that fixed input beam patterns consistently outperformed both random and pre-configured patterns when subjected to equivalent training samples and beam measurements. Interestingly, AI/ML-driven spatial beam prediction markedly surpassed the sparse beam sweeping approach in terms of accuracy and average L1-RSRP difference of Top-1 predicted beam pair. Yet, one should tread carefully when using the average L1-RSRP difference of the Top-1 (or Top-K) predicted beam as an indicator of performance. Without juxtaposing it with the average L1-RSRP deviation between the ideal metrics of Top-1 and Top-K genie-aided beams from the testing dataset, the metric might offer a skewed perspective. In our future work, we envisage a consolidated model suitable for diverse Set B values, which bolsters the model's adaptability. Furthermore, a transfer learning-based approach could be studied to enhance model generalization capability and minimize the training overhead and storage requirements associated with developing multiple models.

References

[1] M. Qurratulain Khan, A. Gaber, P. Schulz, and G. Fettweis, "Machine learning for millimeter wave and terahertz beam management: A survey and open challenges," *IEEE Access*, vol. 11, pp. 11880–11902, Feb. 2023.

[2] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, pp. 28–41, Jul. 2019.

[3] J. Jalali and A. Khalili, "Optimal resource allocation for MC-NOMA in SWIPT-enabled networks," *IEEE Commun. Lett.*, vol. 24, pp. 2250–2254, June 2020.

[4] J. Jalali, *Resource allocation for SWIPT in multi-service wireless networks*. M.S. thesis, Dept. Telecommun. Inf. Process., TELIN/IMEC, Ghent Univ., Ghent, Belgium, Jun. 2020.

[5] *Evaluation on AI/ML for Beam Management*. R1-2209978,3GPP, Qualcomm, RAN1-110bis-e, Oct. 2022.

[6] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surv. Tutor.*, vol. 21, pp. 173–196, Sept. 2019.

[7] T. S. Cousik, V. K. Shah, T. Erpek, Y. E. Sagduyu, and J. H. Reed, "Deep learning for fast and reliable initial access in AI-driven 6G mm Wave networks," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–12, 2022.

[8] E. Onggosanusi, M. S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant, R. Chen, R. Tamrakar, and Q. Gao, "Modular and high-resolution channel state information and beam management for 5G new radio," *IEEE Commun. Mag.*, vol. 56, pp. 48–55, Mar. 2018.

[9] C. Sun, L. Zhao, T. Cui, H. Li, Y. Bai, S. Wu, and Q. Tong, "AI model selection and monitoring for beam management in 5G-Advanced," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 38–50, 2024.

[10] T. S. Cousik, V. K. Shah, J. H. Reed, T. Erpek, and Y. E. Sagduyu, "Fast initial access with deep learning for beam prediction in 5G mmWave networks," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, pp. 664–669, 2021.

[11] *On the support of AI in PHY for 5G Advanced*. RWS-210185,3GPP, Samsung, 3GPP TSG RAN Rel-18 Wkshps., Jun. 2021.

[12] W.-T. Shih, C.-K. Wen, S.-H. Tsai, and S. Jin, "Fast antenna and beam switching method for mmWave handsets with hand blockage," *IEEE Trans. Wirel. Communi.*, vol. 20, pp. 8134–8148, Dec. 2021.

[13] Q. Li, P. Sisk, A. Kannan, T. Yoo, T. Luo, G. Shah, B. Manjunath, C. Samarathungage, M. T. Boroujeni, H. Pezeshki, and H. Joshi, "Machine learning based time domain millimeter-wave beam prediction for 5G-advanced and beyond: Design, analysis, and over-the-air experiments," *IEEE J. Sel. Areas Commun.*, vol. 41, pp. 1787–1809, Jun. 2023.

[14] *Discussion and Evaluation of AI/ML for Beam Management*. R1-2306433, 3GPP, Futurewei, 3GPP TSG RAN1-114 WG1, Aug. 2023.

[15] *Discussion on Other Aspects of AI/ML for Beam Management*. R1-2306434, 3GPP, Futurewei, 3GPP TSG RAN1-114 WG1, Aug. 2023.

[16] X. Li, B. Gao, Y. Wang, Q. Luo, S. Shao, X. Yang, W. Yan, H. Wu, and B. Han, "Compressed beam selection for single/multi-cell beam management," in *Proc. IEEE 95th Veh. Techno. Conf. (VTC2022-Spring)*, pp. 1–5, 2022.

[17] T. T. Nguyen and K.-K. Nguyen, "A deep learning framework for beam selection and power control in massive MIMO-millimeter-wave communications," *IEEE Trans. Mob. Comput.*, vol. 22, pp. 4374–4387, Aug. 2023.

[18] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using Sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, pp. 5504–5518, Jun. 2020.

[19] M. S. Sim, Y.-G. Lim, S. H. Park, L. Dai, and C.-B. Chae, "Deep learning-based mmWave beam selection for 5G NR/6G with Sub-6 GHz channel information: Algorithms and prototype validation," *IEEE Access*, vol. 8, pp. 51634–51646, Mar. 2020.

[20] H. Echigo, Y. Cao, M. Bouazizi, and T. Ohtsuki, "A deep learning-based low overhead beam selection in mmWave communications," *IEEE Trans. Veh. Technol.*, vol. 70, pp. 682–691, Jan. 2021.

[21] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, pp. 3242–3256, Jun. 2014.

[22] *Study on New Radio Access Technology Physical Layer Aspects (Release 14)*. 3GPP TR 38.802 V14.2.0, Tech. Rep., Sept. 2017.

[23] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, Dec. 2020.

[24] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of Adam and RMSProp," in *Proc. IEEE/CVF Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11119–11127, 2019.

[25] *Feature Lead Summary #4 Evaluation of AI/ML for Beam Management*. R1-2301959, 3GPP, Moderator (Samsung), 3GPP TSG RAN1-112 WG1 Rel-18, Mar. 2023.

[26] *Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 16)*. 3GPP TR 38.901 V16.1.0, Tech. Rep., Nov. 2020.