A Low-Complexity Machine Learning Design for mmWave Beam Prediction

Muhammad Qurratulain Khan, Abdo Gaber, Mohammad Parvini, *Member, IEEE*, Philipp Schulz, and Gerhard Fettweis, *Fellow, IEEE*

***Abstract*—Machine learning (ML) for fifth generation (5G)-Advanced air interface is currently being studied by the 3rd Generation Partnership Project (3GPP), where millimeter-wave (mmWave) beam prediction is an important use case. Thereby the targets are to reduce reference signal (RS) overhead, latency, and power consumption, which are currently imperative for frequent beam measurements. To this end, a low-complexity ML design is presented, that exploits the spatial correlation between beam qualities to expedite the spatial-domain beam prediction. Evaluation results showcase that the proposal achieves a beam prediction accuracy of $96\,\%$ with $75\,\%$ reduction in RS overhead and lower computational complexity as compared to the state of the art. Further, to demonstrate the practicality of the proposed design, we analyze its generalization behavior across different communication scenarios.**

***Index Terms*—Beam prediction, low-complexity, machine learning (ML), millimeter-wave (mmWave).**

I. Introduction

THE AVAILABILITY of abundant bandwidth at millimeter-wave (mmWave) bands makes it a requisite for higher throughput. However, to achieve an adequate link margin, beamforming via large antenna arrays is essential [1]. Consequently, the evaluation of beam qualities through frequent beam measurements and beam qualities reporting is imperative to help the base station (BS) and the user equipment (UE) decide an optimal beam pair for link establishment. Within the 3rd Generation Partnership Project (3GPP) this is referred to as the beam management (BM) procedure [2].

In order to enable the UE to measure the beam qualities, beamformed reference signals (RSs), i.e., synchronization sequence blocks (SSBs) are periodically transmitted by the BS. This allows the UE to measure the qualities of all the BS transmit beams in terms of their reference signal received powers (RSRPs) by sweeping all of its receive beams. This exhaustive beam scan (EBS)-based BM procedure suffers from two major limitations. First, periodic transmission of large number of SSBs results in huge RS overhead, latency and RS transmit power consumption. Second, frequent measurements, quantization, and reporting of beam qualities consume a large amount of UE power [3], [4]. To reduce this RS overhead, a hierarchical beam scan (HBS) with fewer parent (wide) and more child (narrow) beams was considered [5]. Nevertheless, it suffers from higher latency and inaccuracy of beam selection.

Recently, machine learning (ML) methods have been exploited for efficient beam prediction and selection [3]. A simple approach to reduce the RS overhead is to utilize contextual information of the UE location [6] or other sensory information [7] to train an ML model for beam prediction. However, acquisition of such contextual information requires additional sensors and additional feedback overhead. To avoid this issue, this letter in [8] fuses the concept of HBS with a supervised convolutional neural network (CNN) and exploits the spatial correlation among beam qualities to predict the optimal child beam. A similar approach in [9] utilizes the received signal vector of parent beams as an input to a CNN. Alternatively, other works propose to reduce the RS overhead and latency by transmitting a uniformly sampled subset of child beams in [10] and a predefined subset of child beams in [11]. The beam measurements obtained via these subsets of child beams are then utilized by a CNN for optimal beam prediction. However, their performance is highly sensitive to the selection of sampled child beams. Here, only the most relevant studies for this letter were summarized. For a more detailed overview, the interested reader is referred to the comprehensive survey in [3].

Starting from 2022, this letter of ML-based BM with a focus on spatial and temporal domain beam prediction is an important project at 3GPP [12]. Here, the focus is to explore the benefits of augmenting the fifth generation (5G)-Advanced New Radio (NR) air interface with ML for enhanced performance and/or reduced RS overhead and complexity. Following 3GPP guidelines, companies report their proposed evaluation methodology and results on ML-based beam prediction [13], [14]. Some of the recent investigations on spatial domain beam prediction include [15], [16], where based on the received power of a subset of the transmit beams, a CNN is trained to predict the RSRPs of the non-transmitted beams.

Though most of the discussed ML solutions significantly reduce the RS overhead, no significant attention has been paid to the ML model computational complexity, training time and its generalization capabilities. To bridge this research gap, this letter exploits the spatial correlation among parent and child beams to propose a low-complexity ML-based beam prediction approach that achieves a performance close to the optimal EBS but with much lower computational complexity

Manuscript received 5 February 2024; accepted 15 March 2024. Date of publication 26 March 2024; date of current version 11 June 2024. This work was supported in part by the European Union's SEMANTIC ITN Project under the Marie Skodowska-Curie Grant under Agreement 861165, and in part by the Federal Ministry of Education and Research (BMBF) as part of the Project 6G-ANNA under Grant 16KISK103. The associate editor coordinating the review of this article and approving it for publication was J. Tang. *(Corresponding author: Muhammad Qurratulain Khan.)*

Muhammad Qurratulain Khan is with the NI Dresden R&D, National Instruments Corporation, 01099 Dresden, Germany, and also with the Vodafone Chair for Mobile Communications Systems, Technische Universität, 01069 Dresden, Germany (e-mail: muhammad.khan@ifn.et.tu-dresden.de).

Abdo Gaber is with National Instruments Corporation, 01099 Dresden, Germany (e-mail: abdo.gaber@ni.com).

Mohammad Parvini, Philipp Schulz, and Gerhard Fettweis are with the Vodafone Chair for Mobile Communications Systems, Technische Universität, 01069 Dresden, Germany (e-mail: mohammad.parvini@tu-dresden.de; philipp.schulz2@tu-dresden.de; gerhard.fettweis@tu-dresden.de).

Digital Object Identifier 10.1109/LWC.2024.3381447

and power consumption as compared to the state of the art, resulting in fast and accurate beam prediction. Further, due to the transmission of a smaller number of parent beams, the RS overhead is significantly reduced leading to benefits in terms of latency and power consumption at the BS and UE. To ensure the practicality of our proposal, the model training and testing dataset is collected using 3GPP defined BM procedure [2], performance is evaluated over 3GPP specified key performance indicators (KPIs), and model generalization is investigated over 3GPP specified scenarios [17].

II. System Model

This section details channel and beam steering models, followed by an overview of the 3GPP BM procedure.

A. Channel and Beam Steering Model

We consider a downlink mmWave multiple-input multiple-output (MIMO) communication system with N_{T} and N_{R} antenna elements at the BS and UE, respectively. Using the clustered channel model, the channel $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}}$ is the sum of the line-of-sight (LOS) path and C non-line-of-sight (NLOS) clusters with L paths per cluster [18].

$$\mathbf{H} = \sqrt{\frac{K\Lambda}{K+1}}\alpha_{\mathrm{LOS}}\mathbf{a}_{\mathrm{R}}(\phi^{\mathrm{R}}_{\mathrm{LOS}}, \theta^{\mathrm{R}}_{\mathrm{LOS}})\mathbf{a}^{H}_{\mathrm{T}}(\phi^{\mathrm{T}}_{\mathrm{LOS}}, \theta^{\mathrm{T}}_{\mathrm{LOS}}) + \sqrt{\frac{\Lambda}{L(K+1)}}\sum_{c=1}^{C}\sum_{l=1}^{L}\alpha_{c,l}\mathbf{a}_{\mathrm{R}}(\phi^{\mathrm{R}}_{c,l}, \theta^{\mathrm{R}}_{c,l})\mathbf{a}^{H}_{\mathrm{T}}(\phi^{\mathrm{T}}_{c,l}, \theta^{\mathrm{T}}_{c,l}). \quad (1)$$

Here, the l-th path of the c-th cluster has azimuth (elevation) angle-of-arrival (AoA) $\phi^{\mathrm{R}}_{c,l}(\theta^{\mathrm{R}}_{c,l})$ and azimuth (elevation) angle-of-departure (AoD) $\phi^{\mathrm{T}}_{c,l}(\theta^{\mathrm{T}}_{c,l})$, while $\alpha_{c,l}$ is the complex path gain. The same variables are analogously defined for the LOS path and are indicated by the LOS index. Furthermore, $\mathbf{a}_{\mathrm{R}}(\cdot) \in \mathbb{C}^{N_{\mathrm{R}} \times 1}$ and $\mathbf{a}_{\mathrm{T}}(\cdot) \in \mathbb{C}^{N_{\mathrm{T}} \times 1}$ denote the UE and the BS array response, respectively, $(\cdot)^H$ denotes conjugate transpose, K is the Ricean factor, and Λ indicates the pathloss.

We assume a uniform planar array (UPA) in the y-z plane at the BS and UE with N_{y} and N_{z} antenna elements $(N_{\mathrm{y}}N_{\mathrm{z}} = N)$ on the y- and z-axis, respectively. Here, for ease of notation we drop the subscript for the BS and UE. The array response vector for a UPA with antenna element spacing of $d = \frac{\lambda}{2}$, where λ indicates wavelength, can then be written as

$$\mathbf{a}(\phi, \theta) = \frac{1}{\sqrt{N}}[e^{j\frac{2\pi}{\lambda}d(y'\sin(\phi)\sin(\theta)+z'\cos(\theta))}]^{T}_{\substack{y'=0,\ldots,N_{\mathrm{y}}-1 \\ z'=0,\ldots,N_{\mathrm{z}}-1}}. \quad (2)$$

For beam steering, we consider phase shifter based analog beamforming with one radio frequency (RF) chain. At the BS the transmit signal is beamformed by a beamforming vector $\mathbf{f} = [f_1, f_2, \ldots, f_{N_{\mathrm{T}}}]^T \in \mathbb{C}^{N_{\mathrm{T}} \times 1}$ and at the UE the received signals are combined with a combining vector $\mathbf{w} = [w_1, w_2, \ldots, w_{N_{\mathrm{R}}}]^T \in \mathbb{C}^{N_{\mathrm{R}} \times 1}$. Here, f_i and w_j denote the complex weight on the i-th transmit and j-th receive antenna element, respectively. The transmit and receive beams are selected from the predefined codebooks $\mathcal{C}_{\mathrm{T}}$ and $\mathcal{C}_{\mathrm{R}}$, consisting of F and W candidate beams, respectively. The codebooks are designed on the following beam steering scheme.

$$\mathbf{f} \in \mathcal{C}_{\mathrm{T}} = \{\mathbf{a}_{\mathrm{T}}(\bar{\phi}^{\mathrm{T}}_1, \bar{\theta}^{\mathrm{T}}_1), \mathbf{a}_{\mathrm{T}}(\bar{\phi}^{\mathrm{T}}_2, \bar{\theta}^{\mathrm{T}}_2), \ldots, \mathbf{a}_{\mathrm{T}}(\bar{\phi}^{\mathrm{T}}_F, \bar{\theta}^{\mathrm{T}}_F)\} \quad (3)$$

$$\mathbf{w} \in \mathcal{C}_{\mathrm{R}} = \{\mathbf{a}_{\mathrm{R}}(\bar{\phi}^{\mathrm{R}}_1, \bar{\theta}^{\mathrm{R}}_1), \mathbf{a}_{\mathrm{R}}(\bar{\phi}^{\mathrm{R}}_2, \bar{\theta}^{\mathrm{R}}_2), \ldots, \mathbf{a}_{\mathrm{R}}(\bar{\phi}^{\mathrm{R}}_W, \bar{\theta}^{\mathrm{R}}_W)\} \quad (4)$$

Here, $\bar{\phi}^{\mathrm{T}}_m(\bar{\theta}^{\mathrm{T}}_m)$ for the m-th transmitting beam $\mathbf{f}_m$, $m \in \mathcal{F} = \{1, 2, \ldots, F\}$ and $\bar{\phi}^{\mathrm{R}}_n(\bar{\theta}^{\mathrm{R}}_n)$ for the n-the receiving beam $\mathbf{w}_n$, $n \in \mathcal{W} = \{1, 2, \ldots, W\}$ are the quantized azimuth (elevation) AoD and AoA, respectively. Given the channel matrix $\mathbf{H}$, the transmit signal x, the m-th transmitting beam $\mathbf{f}_m$ and the n-th receiving beam $\mathbf{w}_n$, the received signal $y_{m,n}$ is

$$y_{m,n} = \sqrt{P}\mathbf{w}^H_n\mathbf{H}\mathbf{f}_m x + \mathbf{w}^H_n\boldsymbol{\eta}, \quad (5)$$

where P is the transmit power and $\boldsymbol{\eta} \in \mathbb{C}^{N_{\mathrm{R}} \times 1}$ is the additive white Gaussian noise (AWGN).

B. Beam Management in 5G NR

The 3GPP specified EBS-based BM procedure aims to find the optimal beam pair $\{\mathbf{f}_{m^*}, \mathbf{w}_{n^*}\}$ by maximizing the RSRP given as: $\mathrm{RSRP}_{m,n} = |y_{m,n}|^2$. The optimization problem can be formulated as

$$\{m^*, n^*\} = \underset{m \in \mathcal{F}, n \in \mathcal{W}}{argmax} \quad \mathrm{RSRP}_{m,n}. \quad (6)$$

To solve this optimization problem, EBS requires $F \cdot W$ beam measurements resulting in huge RS overhead and latency. Alternatively, HBS utilizes a multi-resolution codebook and the problem of beam selection is divided into two levels as

$$\{\hat{m}^{\mathrm{p}}, \hat{n}^{\mathrm{p}}\} = \underset{m^{\mathrm{p}} \in \mathcal{F}^{\mathrm{p}}, n^{\mathrm{p}} \in \mathcal{W}^{\mathrm{p}}}{\mathrm{argmax}} \quad \mathrm{RSRP}^{\mathrm{p}}_{m^{\mathrm{p}}, n^{\mathrm{p}}}, \quad (7a)$$

$$\{\hat{m}, \hat{n}\} = \underset{m \in \mathcal{F}^{\mathrm{c}}(\hat{m}^{\mathrm{p}}), n \in \mathcal{W}^{\mathrm{c}}(\hat{n}^{\mathrm{p}})}{\mathrm{argmax}} \quad \mathrm{RSRP}^{\mathrm{c}}_{m,n}. \quad (7b)$$

Here, $\mathcal{F}^{\mathrm{p}} = \{1, 2, \ldots, F^{\mathrm{p}}\}$ and $\mathcal{W}^{\mathrm{p}} = \{1, 2, \ldots, W^{\mathrm{p}}\}$ contain the parent beams at the BS and UE, respectively. Similarly, $\mathcal{F}^{\mathrm{c}}(m^{\mathrm{p}}) \subset \mathcal{F}$ and $\mathcal{W}^{\mathrm{c}}(n^{\mathrm{p}}) \subset \mathcal{W}$ contains the child beams of the BS parent beam m^{p} and the UE parent beam n^{p}, respectively. Further, $F^{\mathrm{p}} = \frac{F}{s_{\mathrm{T}}}$, $W^{\mathrm{p}} = \frac{W}{s_{\mathrm{R}}}$, $|\mathcal{F}^{\mathrm{c}}(\hat{m}^{\mathrm{p}})| = s_{\mathrm{T}}$, and $|\mathcal{W}^{\mathrm{c}}(\hat{n}^{\mathrm{p}})| = s_{\mathrm{R}}$, where s_{T} and s_{R} define the number of child beams within each parent beam at the BS and UE, respectively. Notably, HBS requires $F^{\mathrm{p}} \cdot W^{\mathrm{p}} + s_{\mathrm{T}} \cdot s_{\mathrm{R}}$ beam measurements resulting in reduced RS overhead. However, due to multi-level search it suffers from higher latency.

III. Low-Complexity Machine Learning Design for mmWave Beam Prediction

In this section, we leverage the angular domain spatial correlation to propose a low-complexity ML model for fast and efficient beam prediction. Considering the fact that very large antenna arrays can only be employed at the BS due to size constraints, we limit our discussion to the identification of the optimal transmit beam.

A. Algorithm Framework

Motivated by the two-level beam search, we propose to cover the whole angular region with a smaller number of first-level parent beams. By doing so, we observe that there exists a strong angular spatial correlation among parent and child beams in a certain environment. As an example, Fig. 1 shows the angular spatial correlation between the RSRPs of the parent and child beams, where each parent beam contains four child beams. Consequently, we assume that the $\mathrm{RSRP}^{\mathrm{c}}$ of the child beams is a function $f_1(\cdot)$ of the parental RSRP values, i.e.,

$$\mathrm{RSRP}^{\mathrm{c}} = f_1(\mathrm{RSRP}^{\mathrm{p}}). \quad (8)$$

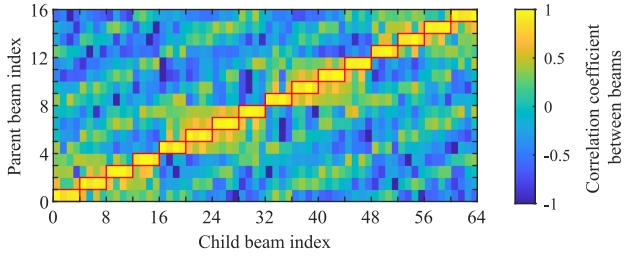


Fig. 1. Spatial correlation among RSRPs of parent and child beams.

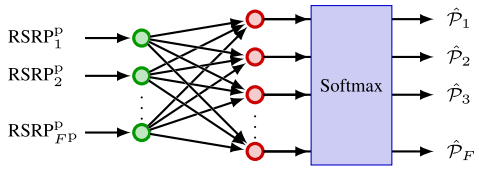


Fig. 2. Proposed low-complexity ML design for beam prediction.

In particular, we aim on probing the parent beams and obtaining their corresponding RSRPs and by intelligently merging these parent RSRPs with the strong correlation among parent and child beams, we propose to utilize an ML model that can predict the best child beam index $\hat{m}$. Due to the discrete number of candidate beams, the beam prediction problem can be formulated as a multiclass-classification problem as

$$\hat{m} = f_2(\mathrm{RSRP}^{\mathrm{p}}), \quad \hat{m} \in \{1, 2, \ldots, F\}, \tag{9}$$

where $f_2(\cdot)$ is the function that learns the correlation between parent and child RSRPs for optimal beam index prediction.

The benefits of our approach are multi-fold. First, due to the transmission of a smaller number of parent beams, it results in reduced RS overhead, latency, and RS transmit power. Second, since the proposed approach relies only on the RSRP measurements, which in the current 3GPP standardization are periodically reported to the BS [2], it does not incur any additional feedback overhead and is fully suitable for practical deployment in the existing mmWave communication system. Further, it can be extended to a multi-user scenario, where the beam prediction for each user can be performed separately. It can also be adopted for a UE with an antenna array, since the receive beam prediction problem can be formulated in a similar fashion. Furthermore, the proposed approach can also be deployed at multiple BSs, where each BS can train its own ML model for a specific propagation environment.

B. Model Design

In this section, we introduce our ML model and its corresponding inputs and outputs as shown in Fig. 2.

1) Input Layer: Based on our previous discussions, the RSRP$^{\mathrm{p}}$ of the parent beams obtained via the first level of traditional HBS is provided as an input to the model. This indicates that the input layer consists of F^{p} nodes. As an example, considering $F = 64$ beams and selecting $s_{\mathrm{T}} = 4$ results in $F^{\mathrm{p}} = 16$ parent beams, which indicates $1 - \frac{16}{64} = 75\,\%$ reduction in RS overhead as compared to the optimal EBS. This RS overhead reduction leads to reduced latency and RS transmit power and avoids frequent beam measurements at the UE resulting in reduced power consumption.

2) Output Layer: For prediction of the optimal beam from all the candidate child beams, a fully-connected (FC) layer, consisting of F nodes is introduced, which learns the spatial correlation between RSRP$^{\mathrm{p}}$ and RSRP$^{\mathrm{c}}$ and transforms it to the candidate child beams. The output of this FC layer $\boldsymbol{\gamma} \in \mathbb{C}^{F\times 1}$ can be written as

$$\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_F]^T = \mathbf{A}^T \mathrm{RSRP}^{\mathrm{p}} + \mathbf{b}, \tag{10}$$

TABLE I
LIST OF SIMULATION PARAMETERS

Parameters		Values
No. of BS antennas	N_{T}	64
No. of UE antennas	N_{R}	8
BS codebook size	F	64
BS parent codebook size	F^{p}	16
Transmit power	P	30 dBm
BS antenna gain [18]		8 dBi
UE codebook size	W	8
UE noise figure	N_{F}	6 dB
Center frequency	f_{c}	28 GHz
Bandwidth	B	100 MHz
Sub-carrier spacing		120 kHz
Cell radius		200 m
Noise power	σ^2	$(-174 + 10\log_{10}B + N_{\mathrm{F}})$ dBm
Path loss at distance r		$(20\log_{10}r + 20\log_{10}f_{\mathrm{c}} - 147.56)$ dB

where $\mathbf{A} \in \mathbb{R}^{F^{\mathrm{p}}\times F}$ and $\mathbf{b} \in \mathbb{R}^{F\times 1}$ represents the model weights and the biases, respectively. Then a softmax activation function converts the output of FC layer to a probability measurement as follows

$$\hat{\mathcal{P}}_m = \frac{e^{\gamma_m}}{\sum_{\mu=1}^{F} e^{\gamma_\mu}}, \quad m \in \{1, 2, \ldots, F\}. \tag{11}$$

The output of the softmax layer represents the predicted probability of the m-th beam being the best. Finally, the child beam with maximum probability $\hat{\mathcal{P}}_m$ is selected, i.e.,

$$\hat{m} = \underset{m\in\{1,2,\ldots,F\}}{\mathrm{argmax}} \ \hat{\mathcal{P}}_m. \tag{12}$$

IV. PERFORMANCE EVALUATION

This section evaluates the performance of our proposed model. For reproducibility of results and extension, our simulation dataset and source code is publicly available [19].

A. Dataset Generation and Model Training

Our dataset consists of parent RSRP measurements, i.e., RSRP$^{\mathrm{p}}$, which are provided as input features to the proposed model. In addition, offline training labels, i.e., optimal beam indices for training are obtained via the optimal EBS [17]. Table I lists default simulation parameters for the dataset collection. The location of the UE is randomly drawn uniformly from the cell coverage area. Finally, the channel model is considered as a clustered delay line (CDL) model with a default delay spread of 100 ns [18]. The dataset consists of 25 000 samples, where the training, validation, and testing data split is 70 %, 10 %, and 20 %, respectively. Further, the ML model is trained for $n_{\mathrm{e}} = 100$ epochs and the model parameters are optimized by the Adam optimizer [20] with the mean square error as the loss function.

TABLE II
COMPUTATIONAL COMPLEXITY COMPARISON

	No. of Trainable Parameters	Model Size (Mbits)	No. of FLOPs
Proposed model	1,088	0.04	1.09×10^3
FC-NN in [6]	17,728	0.5	1.77×10^4
CNN in [9]	67,008	2.1	3.32×10^5
CNN in [10]	263,044	8.5	1.02×10^6
CNN in [8]	352,034	11.2	1.37×10^6
CNN in [11]	595,289	19.04	3.70×10^7
CNN in [16]	739,073	23.6	4.73×10^7

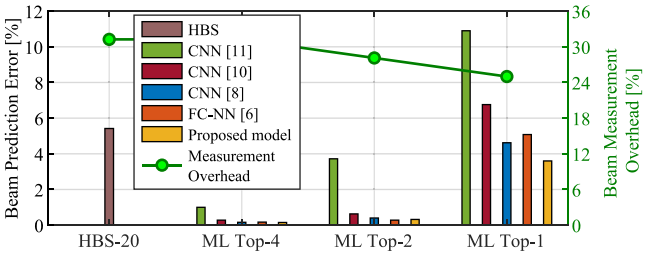


Fig. 3. Comparison of beam prediction error and RS overhead.

B. Key Performance Indicators

For performance evaluation in terms of overhead, the RS overhead reduction (%) is defined as $1 - \frac{F^{\mathrm{P}}}{F}$. Similarly, latency reduction (%) is defined as $1 - \frac{T_{\mathrm{FP}}}{T_{\mathrm{F}}}$, where T_{FP} and T_{F} represent the transmission time of F^{P} and F SSBs, respectively. For beam prediction accuracy, the KPI Top-K (%) is defined as the percentage that the truly optimal transmit beam is among the K best beams predicted by the ML model and the beam prediction error (%) is calculated as 1 − beam prediction accuracy. Here, the Top-1 optimal transmit beam is obtained via EBS [17]. Further, the beam prediction accuracy is also evaluated in terms of achieved average RSRP. Finally, to highlight the practicality of our approach, we analyze model complexity, inference latency, and its generalization behavior across various scenarios.

C. Complexity Analysis

An important measure of ML model complexity is the number of trainable parameters (n_{l}), which for an FC-neural network (NN) layer with n_{i} inputs and n_{o} outputs can be computed as $n_{\mathrm{l}} = (n_{\mathrm{i}}+1)n_{\mathrm{o}}$. Further, for a convolutional layer this number can be obtained as $n_{\mathrm{l}} = n_{\mathrm{f}}(f_{\mathrm{h}}f_{\mathrm{w}}f_{\mathrm{d}} + 1)$, where $n_{\mathrm{f}}, f_{\mathrm{h}}, f_{\mathrm{w}}$, and f_{d} indicate the number of filters, filter height, width, and depth, respectively. We evaluate the complexity in terms of model size with 32-bit precision. Table II indicates that due to a smaller number of trainable parameters and smaller model size, the proposed design benefits from lower power consumption as compared to the state of the art.

The time complexity of our proposed ML model is compared in terms of number of required floating-point operations (FLOPs) using big-$\mathcal{O}$ notation. During training, the ML model performs forward and backward pass and it is useful to analyze the training and inference time complexity. In both directions, the trainable parameters of a layer with w nodes are updated by a matrix-vector multiplication resulting in a time complexity of $\mathcal{O}(w^2)$ FLOPs. Furthermore, considering an NN with l layers, w nodes per layer, and training the network with n_{d} data samples, and for n_{e} epochs requires $\mathcal{O}(n_{\mathrm{e}}n_{\mathrm{d}}lw^2)$ FLOPs during training, while the inference requires only $\mathcal{O}(n_{\mathrm{d}}lw^2)$ FLOPs as only forward pass is performed during inference. Similarly, the time complexity of a CNN, with l_{c} convolutional and l FC layers during training is $\mathcal{O}(n_{\mathrm{e}}n_{\mathrm{d}}(n_{\mathrm{f}}l_{\mathrm{c}}i_{\mathrm{h}}i_{\mathrm{w}}(f_{\mathrm{h}}f_{\mathrm{w}}f_{\mathrm{d}})) + lw^2)$. Here, in addition to the parameters defined above, i_{h} and i_{w} indicate input height and width, respectively. Further, the inference time complexity is then given as $\mathcal{O}(n_{\mathrm{d}}(n_{\mathrm{f}}l_{\mathrm{c}}i_{\mathrm{h}}i_{\mathrm{w}}(f_{\mathrm{h}}f_{\mathrm{w}}f_{\mathrm{d}})) + lw^2)$.

Table II summarizes the complexity comparison with the state of the art. For a fair comparison, the number of estimated FLOPs are for one epoch and one data sample, i.e., $n_{\mathrm{e}} = n_{\mathrm{d}} = 1$. Here, it can be noticed that the proposed ML model achieves significantly lower computational complexity and benefits from lower power consumption. Further, the execution of the proposed ML model on an Intel i7-1185G7 processor with 4 cores and a clock speed of 3 GHz indicates that the training time per epoch and per data sample is 9 µs, which allows efficient and fast model retraining. Besides, the execution time for each prediction is around 2 µs, which is quite marginal as compared to the SSB duration of 35.6 µs resulting in fast beam prediction.

D. Simulation Results

For performance evaluation, in addition to the two-level HBS, and ML solutions from [6], [8], [10], [11], the traditional EBS is selected as a baseline for comparison [17]. In terms of RS overhead, the baseline EBS is set to 100 %. HBS requires 16 parent and 4 child beam measurements, resulting in an RS overhead of 32 %. For ML models, during inference, the RS overhead and latency depends on $K \in \{4, 2, 1\}$, reflecting the necessity of probing the remaining K beams for final selection. Consequently, the proposed approach results in an RS overhead of around 32 %, 28 %, and 25 %, respectively, as shown in Fig. 3. Similarly, due to the reduced number of transmitted SSBs, the latency is reduced by around 69 %, 72 %, and 75 %, respectively. It is worth emphasizing that due to RS overhead reduction, the proposed approach significantly reduces the RS transmit power consumption. Furthermore, due to less frequent beam measurements and beam quality reporting, the UE also lowers its power consumption.

Fig. 3 showcase the beam prediction accuracy, where it can be observed that the proposed design achieves the lowest beam prediction error as compared to the state of the art. It is worth emphasizing that the approaches in [6] and [8] can only achieve a comparable accuracy at the cost of increased computational complexity. Similar observations can be made from Fig. 4, where the performance is compared in terms of the average RSRP. Here, it can be observed that for $K = 1$, the mean RSRP achieved by HBS, [6], [8], and the proposed design is well within a 0.2 dB margin of the optimal (EBS) transmit beam. Consequently, the proposed approach avoids the need of an additional beam scan, which further highlights its efficiency. It is interesting to note that only the approaches in [10], [11] achieve a lower beam prediction accuracy and lower RSRP. It is due the fact that their performance is sensitive to the selection of the subset of child beams.

For the evaluation of ML model generalization, we consider several scenarios as listed in Table III, where the dataset size for all scenarios is kept fixed. Fig. 5(a) demonstrates that

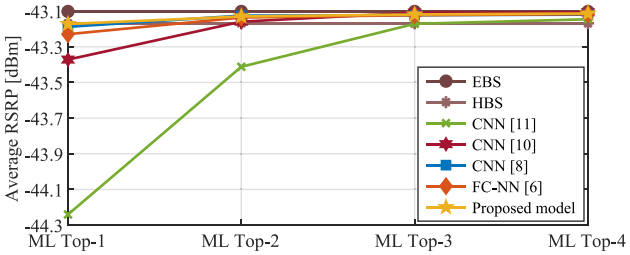
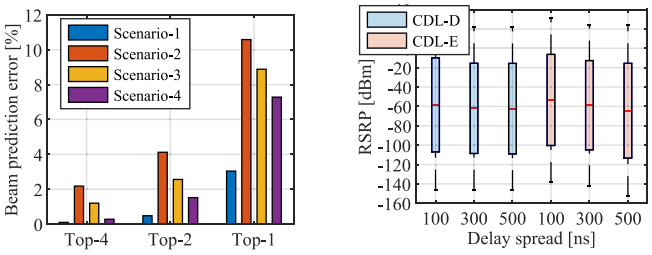


Fig. 4. Comparison in terms of achieved average RSRP [dBm].

TABLE III
LIST OF SCENARIOS FOR ML MODEL GENERALIZATION

Scenario	Model Training		Model Inference	
	Channel Profile	Delay Spread (ns)	Channel Profile	Delay Spread (ns)
1	CDL-D	100	CDL-D	100
2	CDL-D	100	CDL-D	300, 500
3	CDL-D	100, 300, 500	CDL-E	100, 300, 500
4	CDL-D, CDL-E	100, 300, 500	CDL-D, CDL-E	100, 300, 500



(a) Beam prediction error. (b) Dataset distribution.

Fig. 5. Generalization behavior of proposed beam prediction model.

when compared to scenario-1 for $K = 1$, the ML model suffers from a generalization error of around 7.5 percentage points in scenario-2. This generalization error can be attributed to the difference in the dataset distribution of the channel profile CDL-D with different delay spreads as shown in Fig. 5(b). Similar behavior can be observed in scenario-3, where the difference in the dataset distribution of the channel profiles CDL-D and CDL-E leads to a generalization error of around 6 percentage points. Further, the generalization error can be reduced when the model is trained on a mixed data set, i.e., scenario-4. However, owing to the different individual distributions in the mixed dataset, the prediction accuracy error in scenario-4 is still around 4 percentage points higher as compared to scenario-1. An important observation made here is that training a model for a large number of scenarios results in reduced inference performance for a specific scenario. Consequently, there exists a trade-off between ML model accuracy performance and its generalization capabilities.

V. CONCLUSION

In this letter, we focused on spatial domain beam prediction and proposed a low-complexity ML design for reduction in RS overhead, latency, and power consumption. It has been demonstrated that the proposed design can achieve near optimal beam prediction accuracy with significantly lower RS overhead and computational complexity as compared to the state of the art, making it suitable for efficient and faster beam prediction. Further, through simulation results, we showed that there exists a trade-off between ML model accuracy and its generalization capabilities. These 3GPP compliant evaluations indicate the feasibility of ML-based mmWave beam prediction for 5G-Advanced NR and beyond 5G communication networks. As future work, we will extend our framework to the second use case of ML-based temporal domain mmWave beam prediction.

REFERENCES

[1] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[2] "Study on new radio access technology; Version 2.0.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.802, Mar. 2017.
[3] M. Q. 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, 2023.
[4] Q. Li et al., "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, no. 6, pp. 1787–1809, Jun. 2023.
[5] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in Millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.
[6] Y. Heng and J. G. Andrews, "Machine learning-assisted beam alignment for mmWave systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 4, pp. 1142–1155, May 2021.
[7] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 212–216, Feb. 2023.
[8] 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, no. 1, pp. 682–691, Jan. 2021.
[9] K. Ma, D. He, H. Sun, Z. Wang, and S. Chen, "Deep learning assisted calibrated beam training for millimeter-wave communication systems," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6706–6721, Oct. 2021.
[10] G. Jiang and C. Qi, "Near-field beam training based on deep learning for extremely large-scale MIMO," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 2063–2067, Aug. 2023.
[11] Y. Liu, M. Li, J. Zhang, M. Wu, and L. Li, "Deep learning aided two-stage multi-finger beam training in millimeter-wave communication," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 26–30, Jan. 2023.
[12] "New study item: Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," 3GPP, Gothenberg, Sweden, document TSG RAN Meeting #94e, 3GPP RP-213599, Dec. 2021.
[13] "RAN1 chair's notes," 3GPP, Gothenberg, Sweden, document TSG RAN WG1 #109-e, May 2022.
[14] "RAN1 chair's notes," 3GPP, Gothenberg, Sweden, document TSG RAN WG1 #114, Aug. 2023.
[15] "Evaluation on AI/ML for beam management," Huawei, Shenzhen, China, 3GPP, Sophia Antipolis, France, document R1-2203142, May 2023.
[16] "Evaluation methodology and results on AI/ML for beam management," Keysight Electron. Co., Santa Rosa, CA, USA, 3GPP, Sophia Antipolis, France, document R1-2306420, Aug. 2023.
[17] "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface; Version 2.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.843, Dec. 2023.
[18] "Study on channel model for frequencies from 0.5 to 100 GHz; Version 17.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.901, 2022.
[19] "A low-complexity machine learning design for mmWave beam prediction." February 2, 2024. [Online]. Available: https://github.com/MuhammadKhan86/Low-Complexity-Beam-Prediction
[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.