Revisiting Performance Metrics for Multimodal mmWave Beam Prediction Using Deep Learning

Katarina Vuckovic*, Saba M. Hosseini*, and Nazanin Rahnavard
Department of Electrical and Computer Engineering
University of Central Florida, USA
Email: {katarina.vuckovic, saba.mhosseini, nazanin.rahnavard}@ucf.edu

***Abstract*—Directional millimeter-wave (mmWave) wireless communication systems with large antenna arrays face challenges in dynamically managing beams for mobile users. Recently, integrating sensors into communication frameworks has garnered interest for enhancing situational awareness. This paper introduces a novel end-to-end multimodal deep learning architecture for mmWave beam prediction. Our methodology incorporates multiple sensors, including camera, RADAR, LiDAR, and GPS, to improve beam prediction accuracy and efficiency across different scenarios. We conduct extensive comparisons between single and multimodal approaches, exploring various fusion methods such as early and late fusion. The proposed architecture employs both spatial and temporal feature correlation on the training dataset. Additionally, the current beam prediction performance evaluation metrics assume there is only one correct beam, but the nature of the communication problem suggests that multiple beams often perform similarly. Therefore, we propose new metrics that consider multiple beams as valid options and assess performance based on the beam strength. Our results indicate that our early fusion of features from multiple modalities consistently outperforms our late fusion models.**

***Index Terms*—mmWave beam predication, deep learning, multi-modal data fusion, power ratio, performance metrics**

I. Introduction

As the demand for mobile and wireless data increases, millimeter wave (mmWave) directional communication emerges as a solution due to its vast bandwidth [1]. The small wavelength of mmWave enables packing many antennas into a small area, forming an array that generates a narrow beam. Effective communication requires precise beam alignment between transmitter and receiver, involving a non-convex optimization problem solved through an *exhaustive search* over all possible beam pairs [2]. The large number of beams introduces high search overhead, complicating support for mobile and latency-sensitive applications.

Integrated Sensing and Communications (ISAC), a key enabler of next-generation wireless networks, supports various emerging applications by enhancing situational awareness through data integration from a variety of sensors. Initial research on sensing-aided beam prediction focused on position-based solutions using Global Positioning System (GPS) [3], [4], with later studies utilizing vision-based [5]–[10], Light Detection and Ranging (LiDAR) [11]–[14], and Radio Detection and Ranging (RADAR) [15] data for Deep Neural Network (DNN) beam prediction classifiers. Recognizing the limitations of using single sensory modalities, recent work has shifted towards multimodal fusion, leveraging the complementary strengths of these modalities to enhance prediction accuracy and robustness [16], [17]. Late fusion methods that combine visual and positional data are proposed in [18]–[20]. Several late fusion [21]–[24] and hybrid fusion [25] frameworks integrate some combination of LiDAR, RADAR, camera, and GPS data to enhance beam prediction for vehicular networks. Even though the proposed multimodal networks displayed improvement over single modalities, there is still considerable room for performance improvement and further exploration of new fusion models.

Previous beam prediction performance metrics, such as classification accuracy, were insufficient for accurately reflecting the performance of mmWave beam prediction models. These metrics primarily focused on the model's ability to identify *the single correct beam*. However, they did not consider the practical implications of beam selection in real-world scenarios. These metrics failed to account for scenarios where multiple beams may have nearly equivalent performance in terms of received power, resulting in misleading conclusions about model's effectiveness. Therefore, introducing new metrics like Power Ratio (PR) and top-K Beam (top-KB) provides a more holistic and practical evaluation of model performance, reflecting the true capabilities and limitations of these models in real-world applications.

The contributions of our work are summarized as follows:

- **Novel Framework:** We propose a novel end-to-end framework combining Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and Fully Connected (FC) layer. This framework extracts both spatial and temporal features from the input samples. Furthermore, this framework facilitates both single modality and multimodal data fusion.
- **Early Fusion Implementation:** To the best of our knowledge, we are the first to apply early fusion to mmWave beam prediction problem. Early fusion integrates data at the initial stages, improving model coherence and performance when handling diverse data types. Our early fusion deep learning framework outperforms the late fusion and single modalities.
- **Comprehensive Performance Metrics Comparison:** We evaluate various performance metrics—including

* Katarina Vuckovic and Saba Hosseini contributed equally to this work.

top-K accuracy, DBA score [26], precision, and recall—across different models. We analyze these metrics and introduce two new ones: PR and top-KB. This approach sets a thorough benchmark, revealing how each metric responds under different conditions. Our findings highlight that classification accuracy may not be suitable and metrics like PR should be considered.

II. System Model and Problem Formulation

A. Wireless Communication System Model

The system model consists of a BS equipped with an M-element Uniform Linear Array (ULA) mmWave antenna and four types of sensors: an RGB camera, a LiDAR, a RADAR, and a GPS. The BS provides service to a single mobile User Equipment (UE), which is fitted with an omni-directional antenna and a GPS sensor capable of gathering real-time location data. The communication system operates using Orthogonal Frequency Division Multiplexing (OFDM) with K subcarriers. The BS employs a pre-defined beamforming codebook $\mathbf{F} = \{\mathbf{f}_q\}_{q=1}^{Q}$ where $\mathbf{f}_q \in \mathbb{C}^{M\times 1}$ and Q is the total number of beamforming vectors. The received signal at the UE for the k-th subcarrier at time t can be written as

$$y_k[t] = \mathbf{h}_k^T[t]\mathbf{f}_{q[t]}x + v_k[t], \tag{1}$$

where $\mathbf{h}_k[t] \in \mathbb{C}^{M\times 1}$ is the channel between the BS and the UE, $q[t]$ is the beam index, $x \in \mathbb{C}$ transmitted complex symbol, and $v_k[t]$ is the noise with a complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$.

B. Beam Prediction Problem Formulation

The goal of beam prediction is to select the index $q[t]$ of the optimal beamforming vector from a set of candidate indices in the codebook, $\{1, 2, \cdots, Q\}$, in order to maximize the beamforming gain. This can be expressed mathematically as [26]:

$$q^*[t] = \underset{q\in\{1,\ldots,Q\}}{\text{argmax}} \frac{1}{K}\sum_{k=1}^{K}\left|\mathbf{h}_k^T[t]\mathbf{f}_q\right|^2. \tag{2}$$

Typically, the optimal beam index is identified by either leveraging explicit channel information, which is difficult to obtain in mmWave systems or by exhaustive search.

C. Evaluation Metrics

In mmWave beam prediction, the choice of performance metrics can significantly influence the conclusions drawn from the results. While some studies prefer metrics like classification accuracy [26], others focus on precision and recall [19] due to issues like data imbalance. To provide a comprehensive understanding of how these metrics perform under various conditions, this paper brings together different performance metrics types. By comparing these metrics against one another, we aim to establish a more robust benchmark to guide future research and ensure more effective solutions. Additionally, we offer a fresh perspective on beam prediction, suggesting that multiple beams can be considered as plausible solutions. Based on this approach, we introduce two additional metrics (power ratio (PR) and top-KB) to capture this viewpoint.

*1) **Top-K Accuracy**:* Top-K accuracy (top-KA) is defined as the percentage of test samples for which the ground truth beam index is within the K most likely predicted beams [26]. The K classes with highest probability in the softmax layer represent the top-K predicted beams. Furthermore, the authors in [26] define a Distance-based Accuracy (DBA) Score metric as

$$\text{DBA} = \frac{1}{K}\sum_{k=1}^{K}[1 - \frac{1}{N}\sum_{n=1}^{N}\underset{1\le k'\le k}{min}[min(\frac{|\hat{q}_{n,k'} - q_n|}{\Delta}, 1)]] \tag{3}$$

where q_n is the ground-truth beam index and $\hat{q}_{n,k'}$ is the k'th predicted beam index ($1 \le k' \le k$) for sample n. N is the total number of samples in the test set, Δ is a hyper-parameter set to 5, and K is set to 3 [26]. The DBA score measures how close the predicted beams are to the truth beam by assigning scores based on the difference between them.

*2) **Multi-class Precision and Recall**:* In classification tasks, *True Positive* for class i (TP_i) refers to the instances correctly classified as class i, *False Positive* for class i (FP_i) denotes instances incorrectly labeled as class i, and *False Negative* for class i (FN_i) refers to instances of class i that are incorrectly labeled as not belonging to class i. Precision and recall for each class i are calculated as $\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i}$ and $\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$, respectively.

The weighted-average method handles class imbalance present in the simulation dataset [27]. This technique computes metrics for each class and then weights them by class prevalence, assigning a weight to each class proportional to its representation in the dataset. This is represented as weight_i, which is calculated as the ratio of instances of class i (n_i) to the total number of instances across all classes, i.e., $\text{weight}_i = \frac{n_i}{\sum_{i=1}^{N} n_i}$. Here, n_i refers to the number of instances of class i and N is the total number of classes. This approach balances each class's contribution relative to its frequency, which is why some papers include it in their performance metrics [19]. The weighted precision (P_{weighted}) and recall (R_{weighted}) metrics for N number of classes are calculated as:

$$\text{P}_{\text{weighted}} = \frac{\sum_{i=1}^{N}(\text{weight}_i \times \text{Precision}_i)}{\sum_{i=1}^{N}\text{weight}_i} \tag{4}$$

$$\text{R}_{\text{weighted}} = \frac{\sum_{i=1}^{N}(\text{weight}_i \times \text{Recall}_i)}{\sum_{i=1}^{N}\text{weight}_i}. \tag{5}$$

*3) **Power Ratio**:* The primary focus of related studies [4], [18], [26], [28] revolves around the beam prediction accuracy metric or the derived DBA-score. It should be noted that focusing on classification accuracy addresses the problem from a computer vision perspective where there is only one correct class. This contracts with the beam prediction problem in wireless communication systems, where multiple beams can have similar performance. In the context of top-KA, the model provide K beams, aiming to identify the ground truth beam characterized by maximum received power. If the

ground truth beam does not rank within the top-K predicted beams, the classification for that test sample is deemed incorrect. However, this approach overlooks the possibility that a predicted beam may have a power value very close to that of the ground truth beam, suggesting near-equivalent performance from a wireless channel perspective. To illustrate this point, Fig. 1 depicts the normalized received power distribution of beams in a test sample from the DeepSense dataset [27], where beam index 55 exhibits the highest power. Assume the model predicts beams 56, 57, and 58 as the top-3 beams. According to top-3A, this prediction is labeled as *incorrect*. Nonetheless, the powers of the predicted beams closely match that of the true beam 55. Therefore, we propose a metric that provides a more accurate depiction of the model's performance. We define PR as the ratio between the power of the predicted beam and the power of the ground truth beam, given by $PR = \frac{P_{predicted\ beam}}{P_{max}}$. Beams 56, 57, and 58 have a PR of 0.9978, 0.9954, and 0.9715, respectively. This indicates that the predicted beams possess nearly the same power as the true beam 55, resulting in a comparable performance.

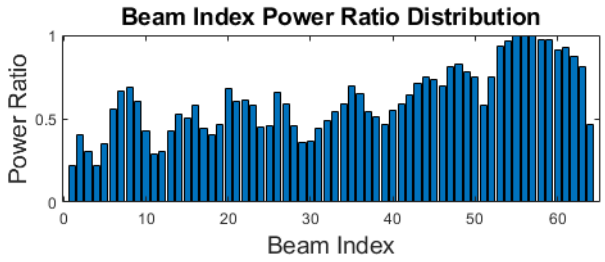

Fig. 1: PR distribution vs beam index for a test sample in the DeepSense dataset.

Additionally, we can evaluate the top-K PR by selecting the top 3 proposed beams and identifying the highest PR among them. For instance, the top-3 PR for beams 56, 57, and 58 would be 0.9978.

*4) **Top-K Beams and Top-K_1,K_2 Beams-Accuracy**:* In Fig. 1, we have shown that the classification accuracy metric based on a single ground truth can be misleading when it comes to the actual performance of the communication systems. It can be extremely difficult to predict the one true best beam from K predicted beams when the true beam powers are very close. This raises the concern whether it is necessary to predict the best (ground truth) beam at all or it is sufficient to utilize a beam with near equivalent power such as beam 56 instead of 55. Given this observation, we propose a new perspective on beam prediction that considers the "top-K beams" (top-KB) We focus on a single prediction, selecting the beam index with the highest probability, and then verify whether this beam is among the top K beam indices with the highest power. Referring back to Fig. 1, the three highest power beams are 55, 56 and 57. Since our prediction is 56, it falls within the top three beams (top-3B). We argue that top-3B in combination with the PR is a much more practical performance metric. Furthermore, we can combine the top-KA and top-KB metrics. The "top-K_1,K_2 beams-accuracy" involves predicting K_2 beams and then verifying whether they are among the K_1 best true beams. The coefficients K_1 and K_2 do not have to be the same. However, in in our analysis we chose to measure top-3,3BA. It is worth noting that top-KB is the special case of top K,1BA.

III. Network Structure

A. Single Modality

Fig. 2 shows the single-modality CNN+GRU+FC architecture, where $\boldsymbol{A}_V$, $\boldsymbol{A}_R$, $\boldsymbol{A}_L$, $\boldsymbol{A}_G$, and q represent the preprocessed data from vision, RADAR, LiDAR, and GPS, along with the predicted beam index, respectively. Vision, RADAR, and LiDAR datasets require sophisticated processing, while GPS data, consisting of simple 2D coordinates, does not need CNN feature extraction.

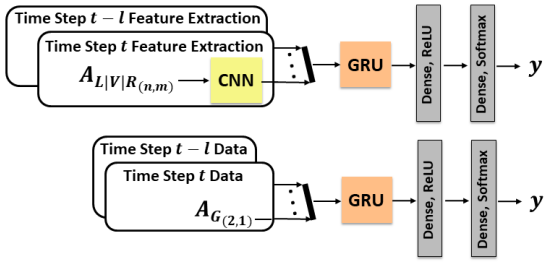

Fig. 2: Single modality DNN mmWave beam prediction architecture. The top graph corresponds to LiDAR, vision, and Radar, while the bottom one corresponds to GPS.

The models generally start with a CNN that extracts spatial features using layers with varying numbers of filters (ranging from 4 to 256) and ReLU activations, followed by batch normalization (momentum=0.9) and max pooling. While the GPS modality lacks this CNN component, it shares the rest of the network architecture. All modalities, utilize a GRU layer with 64 units to capture temporal dynamics, such as movement prediction and potential blockages. After the GRU, a dense layer (FC) with 64 units and a ReLU activation function and a 50% dropout prepares the data for the final FC output layer with 64 softmax units. Although the overall structure remains consistent, the models may differ slightly in the specific configurations of convolution layers, filters, and GRU units for each modality [1].

B. Early Fusion

Early fusion or feature-level fusion leverages correlations among multiple features early in the process. Each datapoint in the dataset is represented by the tuple $(\boldsymbol{A}_L, \boldsymbol{A}_V, \boldsymbol{A}_R, \boldsymbol{A}_G, q)$. In our approach, we combine features from different data sources before passing them to the CNN+GRU+FC framework (Fig. 3). All modalities are rescaled to the same height and width. The input sample data consists of a time sequence with five steps, each having a height of $n = 150$, a width of $m = 150$, and four channels representing the four modalities. The CNN employs six 2D

[1]The authors released their code: https://github.com/katarinavuckovic/MultiModal-Beam-Prediction-CNN-GRU-FCNN

convolution layers, starting with four filters and increasing to 256, using ReLU activation.

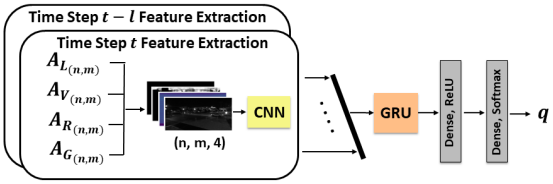

Fig. 3: Early fusion DNN mmWave beam prediction architecture.

C. Late Fusion

Separate DNNs extract low-dimensional feature vectors for each modality, which are then integrated using a late fusion network (Fig. 4). Individual CNN+GRU+FC classifiers generate specific class predictions. They are concatenated into a unified vector and passed to an FC network for final beam classification. This fusion technique benefits from its ability to integrate modalities of different sizes and dimensions, achieving a unified representation through local decisions. However, it is less efficient at leveraging correlations at the feature level across different modalities compared to early fusion.

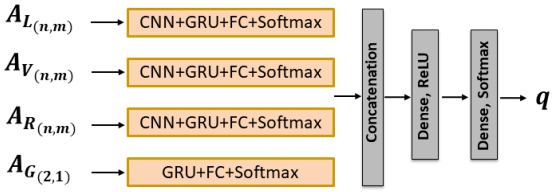

Fig. 4: Late fusion DNN mmWave beam prediction architecture.

IV. Simulation Results

A. Dataset

DeepSense6G [27] is a real-world multi-modal dataset with co-existing communication and sensing data. Scenarios 32-34 occur in a dynamic urban environment, both day and night, with constant vehicle and pedestrian movement, leading to changing RF propagation paths. The BS is equipped with a ULA mmWave antenna, RGB camera, RADAR, LiDAR, and GPS receiver, while the mobile user (a vehicle) has GPS and a mmWave receiver. Each dataset sample includes a sequence of sensory data: the camera, RADAR, and LiDAR have the last 5 samples, and the GPS user location has the last 2 samples, paired with the best beam index. Each sample also contains beam index power distribution information, with 64 beam indices in the codebook.

B. Labeling LOS vs. NLOS Scenarios

Obstacles greatly degrade channel quality in the mmWave band, leading to a reduction in service quality. Separately analyzing Line-Of-Sight (LOS) and Non-LOS (NLOS) cases can provide new insights into the model's performance. If the model cannot predict beams in NLOS cases, this imposes a limit on the accuracy. Unfortunately, the DeepSense 6G dataset does not explicitly label LOS and NLOS samples, necessitating a data labeling process. Therefore, we apply a methodology that classifies each sample by analyzing angular deviation from the expected beam direction, based on maximum and minimum GPS angles for each scenario. These angles are interpolated to predict beam direction for each beam index within a standardized range, with a tolerance threshold of 10 degrees. If a sample's actual beam angle falls within this tolerance, it is labeled as LOS; otherwise, it is classified as NLOS. The dataset contains a total of 10,777 LOS and 466 NLOS samples from Scenarios 32-34.

C. Data Preprocessing

Properly preprocessed data can significantly enhance the accuracy and efficiency of a DNN. Next, we discuss the preprocessing techniques applied to each dataset modality.

1) LiDAR: LiDAR captures precise 3D coordinates (x, y, z) and intensity values, resulting in a comprehensive 360° point cloud. Preprocessing involves standardizing the number of points in each point cloud by downsampling or upsampling to a fixed number, ensuring consistency for reliable analysis. With frequent horizontal target movement, focus is on range, azimuth angle, and intensity. Each point (x, y, z) is processed to compute the Euclidean distance $d = \sqrt{x^2 + y^2 + z^2}$ and the azimuth angle α, calculated as:

$$\alpha = \begin{cases} \arctan 2(y, x) \times \frac{180}{\pi}, \text{if } \arctan 2(y, x) \times \frac{180}{\pi} \geq 0 \\ \arctan 2(y, x) \times \frac{180}{\pi} + 360, \quad \text{otherwise} \end{cases} \tag{6}$$

To ensure all angles fall within 0 to 360 degrees, we add 360 degrees to any negative azimuth angles. Points that are negligibly distant from the sensor, indicating obstructions, are removed. Points are then organized by their computed azimuth angle for efficient data processing. The data is converted to a 2D format by quantizing angles and distances, assigning intensity values to pixels based on their front-view positions. Finally, a bilateral filter is applied to enhance data quality [29].

2) RADAR: Raw RADAR measurements $X \in \mathbb{C}^{M_r \times S \times A}$, where M_r is the number of RADAR antennas, S is the number of samples per chirp, and A is the number of chirps per frame, allowing extraction of range, angles, and speed of moving objects. To convert to Range-Angle (RA) maps, the Range Fast Fourier Transform (FFT) on time samples shifts chirp signals into the frequency domain, revealing distance information based on the signal's round-trip time. Clutter is reduced by averaging fluctuations in chirp samples. The Angle FFT on the receive antenna samples captures angular details, with enhanced resolution achieved using a larger FFT, denoted as M_F, with zero-padding for finer angle sampling. The resulting comprehensive RA map is produced by merging data per chirp sample [15]:

$$H_{RA} = \Psi^P_{RA}(X) = \sum_{a=1}^{A} |F_{2D}(X, :, a)|, \tag{7}$$

where $\Psi^P_{RA}(\cdot)$ is the function that converts raw RADAR data X to RA map and $F_{2D}(\cdot)$ is a 2D Fourier transformation.

3) Camera: The vision sample is an RGB image with dimensions (540, 960, 3). To reduce unnecessary information and reduce computational resources, the images are converted to grayscale and downsized to (150, 150).

4) GPS: The GPS data, consisting of latitude and longitude coordinates for both the user and the BS, is converted into Cartesian-XY coordinates using the Universal Transverse Mercator (UTM) projection [30]. This centers the BS in the coordinate system and normalizes user locations relative to it. The GPS dataset initially includes only the first two instances in the sequence, which are expanded to five for uniform dimensionality across all modalities. Before early fusion, each user's position is transformed into a 2D array format, with one cell containing latitude and longitude, and other cells zeroed to maintain consistent dimensions among modalities.

D. Data Normalization

We employ normalization techniques to address disparities in scale and variance across our dataset, specifically and only for single modality vision and the early fusion models. For the single vision modality, each sample is normalized by dividing pixel values by 255, while we apply L2 normalization on the early-fused multimodal data. For the early fused modalities, normalization ensures that all data from diverse modalities are maintained within a specific range, preventing any single feature with a broader range from disproportionately influencing the learning process. This technique scales each vector in the dataset to have a Euclidean norm of 1, promoting uniformity and mitigating undue influence of features from different scenarios.

E. Noise Addition

Although we are working with real-world data, adding appropriately scaled white noise to each modality before fusion has enhanced both the performance and robustness of our model. We apply Gaussian white noise by adding random values drawn from a normal distribution with zero mean and a specified noise level, ensuring the noise respects the dimensionality of the input data for each modality. This process simulates potential measurement errors or environmental variability, helping the model generalize better to unseen data, reducing overfitting, and improving overall robustness during training.

F. Training

The proposed models are trained end-to-end. The dataset is split into 90% training and 10% testing using a stratified split for balanced class distribution. During training, 20% of the training data is used for validation. We set the learning rate to 0.01 or 0.001, depending on the model. The number of epochs is set to 300 and batch size to 30. Early stopping is employed to prevent overfitting, halting training if validation performance does not improve for 12 consecutive epochs. Additionally, we use a learning rate reduction strategy, decreasing the learning rate by a factor of 10 if the validation loss plateaus for 8 epochs. The results are averaged over 10 training sessions, each using different train/test dataset splits generated from 10 distinct random seed values.

G. Results

This study evaluates the performance of mmWave beam prediction using individual sensor modalities such as Vision (V), LiDAR (L), RADAR (R), and GPS (G), along with various combinations of these modalities. The results, detailed in Table I, include metrics such as top-1A, top-3A, DBA score, top-3B, top-3,3BA, top-1 PR, top-3 PR, and recall/precision.

In both LOS and NLOS conditions, early fusion models consistently outperform single-modality and late-fusion models across all metrics. Among individual modalities, GPS performs best in LOS scenarios but struggles in NLOS conditions due to its inability to detect dynamic blockers. In NLOS conditions, LiDAR, RADAR, and Vision models perform better because they can detect features indicating blockage and recommend an appropriate NLOS beam index.

The early fusion models exhibit the best performance in LOS scenarios, with all early fusion models delivering similar results. This questions the necessity of LVRG processing, given its added complexity. Moreover, the comparable performance of LVR to LVRG and VG shows that GPS is not solely responsible for early fusion model success. While LVR is slightly inferior to LVRG and VG, it still outperforms any single modality, further emphasizing that GPS alone does not drive early fusion model effectiveness. In NLOS scenarios, LVR performs best, as GPS generally hinders performance in these conditions. Although overall accuracy is lower across all modalities in NLOS, likely due to fewer NLOS samples in the training dataset, the models show moderate success in top-3,3BA selections and consistently maintain a PR score of 0.9 or higher. Future work may focus on data augmentation to increase the NLOS sample size.

In terms of model complexity, the single V model contains 182,768 learnable parameters, making it more efficient but less effective in complex scenarios compared to early fusion models. The early VG model has 410, 540 parameters, while the early LVRG model has a similar count of 410, 612. Despite the nearly identical number of parameters between VG and LVRG, LVRG requires more computation due to processing larger data volumes during each forward pass.

1) Rethinking of Performance Metrics: When multiple beams nearly achieve a PR of 1, predicting the accurate beam becomes difficult, especially in imbalanced datasets. If the correct beam index is in a less common class, the model typically favors more frequent classes, rarely choosing the minority class beam index. Therefore, focusing exclusively on top-K accuracy is problematic and misdirected. This mindset suggests a singular correct solution and overlooks alternative viable beams with comparable performance. It seems that researchers are dedicating significant effort to optimizing a process that is unnecessarily complex and does not substantially improve communication performance. The DBA score attempts to address this issue by evaluating the proximity between the true and predicted beam. However,

TABLE I: Performance metrics for different single and multimodal models separated by LOS and NLOS test dataset samples. The top-1 and top-3 accuracy is labeled as top-[1A, 3A], the DBA score as DBA, the top-3 beam as top-3B, the top-3 beams-accuracy as top-3,3BA, the top-1 and top-3 Power Ratio as top-[1, 3] PR, and weighted Recall and weighted Precision are reported as R and P, respectively. Results show the mean and standard deviation for 10 different train/test trials.

Modality	top-[1A, 3A]	DBA	top-3B	top-3,3BA	top-[1, 3] PR	Weighted R, P
LOS						
L	[0.37 ± 0.01, 0.71 ± 0.02]	0.81 ± 0.01	0.70 ± 0.02	0.88 ± 0.01	[0.94 ± 0.00, 0.97 ± 0.00]	0.37 ± 0.01, 0.33 ± 0.02
V	[0.44 ± 0.01, 0.81 ± 0.01]	0.88 ± 0.01	0.81 ± 0.01	0.96 ± 0.01	[0.97 ± 0.00, 0.99 ± 0.00]	0.44 ± 0.01, 0.38 ± 0.01
R	[0.38 ± 0.01, 0.75 ± 0.02]	0.84 ± 0.01	0.75 ± 0.01	0.92 ± 0.01	[0.95 ± 0.00, 0.98 ± 0.00]	0.38 ± 0.01, 0.34 ± 0.01
G	[0.45 ± 0.01, 0.81 ± 0.01]	0.88 ± 0.01	0.81 ± 0.02	0.96 ± 0.01	[0.97 ± 0.00, 0.99 ± 0.00]	0.45 ± 0.01, 0.35 ± 0.01
Early VG	**[0.46 ± 0.02, 0.83 ± 0.01]**	**0.89 ± 0.01**	**0.85 ± 0.01**	**0.97 ± 0.01**	**[0.98 ± 0.00, 0.99 ± 0.00]**	0.46 ± 0.02, 0.42 ± 0.03
Early LVR	**[0.46 ± 0.02, 0.83 ± 0.01]**	**0.89 ± 0.00**	0.84 ± 0.01	**0.97 ± 0.00**	**[0.98 ± 0.00, 0.99 ± 0.00]**	0.46 ± 0.02, 0.41 ± 0.02
Early LVRG	**[0.46 ± 0.01, 0.83 ± 0.01]**	**0.89 ± 0.01**	0.84 ± 0.01	**0.97 ± 0.00**	**[0.98 ± 0.00, 0.99 ± 0.00]**	**0.46 ± 0.01, 0.42 ± 0.02**
Late VG	[0.39 ± 0.05, 0.74 ± 0.06]	0.84 ± 0.04	0.73 ± 0.08	0.93 ± 0.04	[0.95 ± 0.02, 0.99 ± 0.00]	0.39 ± 0.05, 0.32 ± 0.07
Late LVRG	[0.42 ± 0.01, 0.79 ± 0.01]	0.87 ± 0.01	0.78 ± 0.01	0.96 ± 0.00	[0.97 ± 0.00, 0.99 ± 0.00]	0.42 ± 0.01, 0.37 ± 0.01
NLOS						
L	[0.03 ± 0.02, **0.14 ± 0.03**]	**0.18** ± 0.04	0.25 ± 0.04	0.43 ± 0.05	[0.89 ± 0.01, 0.93 ± 0.01]	0.03 ± 0.02, 0.06 ± 0.06
V	[0.05 ± 0.04, 0.12 ± 0.05]	0.17 ± 0.05	**0.28 ± 0.08**	0.48 ± 0.09	**[0.90 ± 0.02, 0.93 ± 0.01]**	0.05 ± 0.04, **0.09 ± 0.08**
R	[**0.06 ± 0.03**, 0.10 ± 0.06]	0.16 ± 0.05	0.25 ± 0.06	0.40 ± 0.09	[0.89 ± 0.01, 0.92 ± 0.01]	**0.06 ± 0.03**, 0.08 ± 0.05
G	[0.01 ± 0.01, 0.06 ± 0.03]	0.08 ± 0.01	**0.28 ± 0.05**	0.41 ± 0.05	[0.89 ± 0.00, 0.92 ± 0.01]	0.01 ± 0.01, 0.01 ± 0.01
Early VG	[0.04 ± 0.05, 0.12 ± 0.05]	0.16 ± 0.05	0.27 ± 0.09	0.47 ± 0.07	**[0.90 ± 0.01, 0.93 ± 0.01]**	0.03 ± 0.05, 0.07 ± 0.09
Early LVR	[0.04 ± 0.01, **0.13 ± 0.07**]	0.17 ± 0.04	**0.28 ± 0.07**	**0.49 ± 0.07**	**[0.90 ± 0.01, 0.93 ± 0.01]**	0.04 ± 0.01, 0.09 ± 0.05
Early LVRG	[0.04 ± 0.04, 0.11 ± 0.07]	0.17 ± 0.05	0.25 ± 0.08	**0.49 ± 0.06**	**[0.90 ± 0.01, 0.93 ± 0.01]**	0.04 ± 0.04, 0.06 ± 0.07
Late VG	[0.02 ± 0.02, 0.05 ± 0.03]	0.09 ± 0.04	0.26 ± 0.05	0.41 ± 0.06	[0.88 ± 0.02, 0.92 ± 0.01]	0.02 ± 0.02, 0.02 ± 0.02
Late LVRG	[0.02 ± 0.03, 0.09 ± 0.06]	0.14 ± 0.04	0.23 ± 0.06	0.42 ± 0.05	**[0.90 ± 0.01, 0.93 ± 0.01]**	0.02 ± 0.03, 0.04 ± 0.07

while this is an improvement, it overlooks the possibility that a nearby beam could have a low PR, whereas a more distant beam might have a PR close to 1. Instead, we propose that evaluating channel efficiency through the analysis of top-KB and PR would be more beneficial.

For instance, as shown in Table I, the early fusion model achieves a top-3A score of 0.83 in LOS, indicating that there is still considerable room for improvement. However, with a top-3 PR score of 0.99, it becomes clear that from the PR standpoint, there is minimal room for further enhancement. Moreover, the top-3,3BA for early fusion in LOS scenarios is 0.97, which may be a more realistic metric given the very high PR score. In contrast, NLOS scenarios show that top-KB values can significantly exceed top-KA. For example, in early LVR NLOS scenarios, top-3B at 0.28 outperforms top-3A at 0.13. The combined top-3,3BA score is 0.49 with an impressive average top-3 PR of 0.93. This example illustrates that despite the deceptively low top-3A, the performance of the model may be better than expected due to the high PR.

2) Power Ratio: We observe high top-1 PR values, often reaching 0.98 in LOS and above 0.90 in NLOS. However, model accuracy remains low, particularly in NLOS scenarios, indicating concerns about misclassified samples. We explore this by considering scenarios where a beam is incorrectly predicted but maintains a high PR, indicating either i) multiple high beams or ii) all beam indices have similar power due to LOS blockage. The latter can produce misleadingly inflated PR values. Our goal is to show that the model performs better than random selection in misclassified cases.

The expected value (mean) is a fundamental measure of central tendency in probability distributions. Now, consider the scenario where a beam index is randomly selected repeatedly, without employing the proposed beam prediction algorithm. In such a case, the resulting expected (mean) PR is obtained.Fig. 5 shows a Complimentary Cumulative Distribution Function (CCDF) of both the mean and predicted PR, plotted exclusively for the misclassified samples in the early fusion LVRG model. As is shown, the predicted beams PR outperform the mean PR. For example, in top-3, more than 90% of LOS and 80% of NLOS samples have a PR greater than 0.9, compared to the mean PR, where only 20% of LOS and 50% of NLOS samples have PR above 0.9. This demonstrates that consistently selecting a beam index with a PR exceeding the expected value implies superior performance of our algorithm compared to random beam index selection.

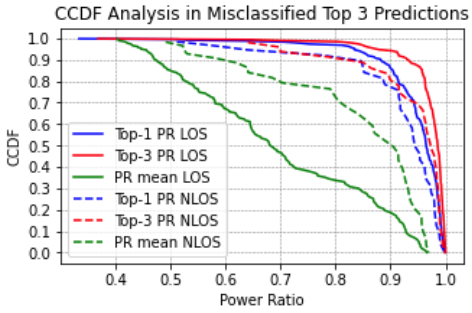

Fig. 5: PR CCDF for early LVRG Fusion Model.

V. Conclusion

This study demonstrates the effectiveness of using a multimodal deep learning architecture for mmWave beam prediction, combining data from camera, RADAR, LiDAR, and GPS to improve the performance of the model. Our proposed CNN+GRU+FC architecture aims to extract both spatial and temporal correlations from the input samples. Our experiments demonstrate that early fusion models outperform single

modality and late fusion approaches. The findings recommend using a broader set of performance metrics, beyond traditional accuracy, to more thoroughly evaluate beam prediction capabilities. Accordingly, we propose incorporating new metrics to provide a more comprehensive view of the model's true performance. By combining existing and proposed metrics, we outline a new way for model evaluation based on more realistic performance metrics.

REFERENCES

[1] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 432–445.

[2] K. Vuckovic, M. B. Mashhadi, F. Hejazi, N. Rahnavard, and A. Alkhateeb, "Paramount: Towards generalizable deep learning for mmWave beam selection using sub-6GHz channel measurements," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[3] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Position-aided millimeter wave v2i beam alignment: A learning-to-rank approach," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–5.

[4] J. Morais, A. Bchboodi, H. Pezeshki, and A. Alkhateeb, "Position-aided beam prediction in the real world: How useful GPS locations actually are?" in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 1824–1829.

[5] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *2020 IEEE 91st vehicular technology conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[6] W. Xu, F. Gao, S. Jin, and A. Alkhateeb, "3D scene-based beam selection for mmWave communications," *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1850–1854, 2020.

[7] M. A. L. Sarker, I. Orikumhi, J. Kang, H.-K. Jwa, J.-H. Na, and S. Kim, "Vision-aided beam allocation for indoor mmWave communications," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 1403–1408.

[8] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5G," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 76–83, 2021.

[9] H. Ahn, I. Orikumhi, J. Kang, H. Park, H. Jwa, J. Na, and S. Kim, "Machine learning-based vision-aided beam selection for mmWave multi-user miso system," *IEEE Wireless Communications Letters*, 2022.

[10] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 193–10 208, 2021.

[11] A. Klautau, N. González-Prelcic, and R. W. Heath, "LIDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 909–912, 2019.

[12] M. B. Mashhadi, M. Jankowski, T.-Y. Tung, S. Kobus, and D. Gündüz, "Federated mmWave beam selection utilizing LIDAR data," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2269–2273, 2021.

[13] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LIDAR and position-aided mmWave beam selection with non-local CNNs and curriculum training," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2979–2990, 2022.

[14] S. Wu, C. Chakrabarti, and A. Alkhateeb, "LiDAR-aided mobile blockage prediction in real-world millimeter wave systems," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 2631–2636.

[15] U. Demirhan and A. Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 2655–2660.

[16] S. Y. Nathan Gaw and M. R. Gahrooei, "Multimodal data fusion for systems improvement: A review," *IISE Transactions*, vol. 54, no. 11, pp. 1098–1116, 2022. [Online]. Available: https://doi.org/10.1080/24725854.2021.1987593

[17] J. Summaira, X. Li, A. M. Shoib, S. Li, and J. Abdul, "Recent advances and trends in multimodal deep learning: A review," 2021.

[18] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multi-modal beam prediction using real millimeter wave datasets," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2727–2731.

[19] G. Reus-Muns, B. Salehi, D. Roy, T. Jian, Z. Wang, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on visual and location data for v2i mmWave beamforming," in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, 2021, pp. 559–566.

[20] J. Nie, Y. Cui, T. Yu, J. Mu, W. Yuan, and X. Jing, "An efficient nocturnal scenarios beamforming based on multi-modal enhanced by object detection," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 515–520.

[21] B. Salehi, J. Gu, D. Roy, and K. Chowdhury, "Flash: Federated learning for automated selection of high-band mmWave sectors," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 1719–1728.

[22] J. Gu, B. Salehi, S. Pimple, D. Roy, and K. R. Chowdhury, "Tune: Transfer learning in unseen environments for v2x mmWave beam selection," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 1658–1663.

[23] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639–7655, 2022.

[24] M. Arnold, G. Charan, U. Demirhan, A. Alkhateeb, and M. Alloulah, "Analysis of multi-modal beam prediction under distribution shift," 2022. [Online]. Available: https://github.com/ITU-AI-ML-in-5G-Challenge/BeamBench

[25] Y. Tian, Q. Zhao, Z. el abidine Kherroubi, F. Boukhalfa, K. Wu, and F. Bader, "Multimodal transformers for wireless communications: A case study in beam prediction," 2023.

[26] G. Charan, U. Demirhan, J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Multi-modal beam prediction challenge 2022: Towards generalization," 2022.

[27] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "Deepsense 6G: A large-scale real-world multi-modal sensing and communication dataset," 2023.

[28] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 212–216, 2023.

[29] G. Melotti, A. Asvadi, and C. Premebida, "CNN-LIDAR pedestrian classification: combining range and reflectance data," in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2018, pp. 1–6.

[30] Turbo87, "utm: Bidirectional UTM-WGS84 converter for python," https://github.com/Turbo87/utm.