Challenges of Predictive Beamforming Using Geographical Positioning: Insights from the DeepSense Dataset

Rodion Vladimirov
University of Liechtenstein
rodion.vladimirov@uni.li

Pavel Laskov
University of Liechtenstein
pavel.laskov@uni.li

***Abstract*—The integration of machine learning (ML) into mobile networks is essential for enhancing their operational efficiency. A prominent application is predictive beamforming, where ML models demonstrated the ability to anticipate optimal beam directions using various inputs. The practical evaluation of such techniques relies on benchmark datasets that reflect diverse physical environments. DeepSense was the first dataset to provide such data for real-world assessment of predictive beamforming.**

We systematically evaluate several predictive beamforming techniques across multiple DeepSense scenarios, identifying key technical challenges. We focus on methods relying on the geographical position (GP) of user equipment (UE)—the most widely employed sensory input—as the primary feature. We investigate how (a) task formulation, (b) data normalization, (c) input modality, and (d) data drift influence beamforming accuracy in representative DeepSense settings. Our results pinpoint the challenges brought by the inherent imprecision of GP data in DeepSense. Notably, we find that the accuracy of current predictive beamforming approaches appears to have reached its plateau, regardless of the ML model used. Additionally, we reveal that data drift significantly degrades the accuracy—an issue not previously reported in the literature.

***Index Terms*—5G and beyond Networks, ML/AI for network management, Wireless technologies design and evaluation**

I. INTRODUCTION

Ongoing work to integrate machine learning (ML) algorithms into mobile networks uses geographical positioning (GP) data to enhance network performance. Today's mobile devices are equipped with satellite navigation modules that generate a valuable stream of data for network management. Despite the evolution of network-based positioning and introduction of new methods in 3GPP Rel. 18, such as bandwidth aggregation, carrier phase and device-to-device sidelink measurements [1]–[4], Global Navigation Satellite Systems (GNSS) remain an essential component of user positioning, particularly for outdoor and high-precision applications. 3GPP has defined several standards for collecting and exposing user data, including the geographical coordinates estimated by user equipment (UE), to be effectively utilized in mobile networks.

In 5G mobile networks relying on mmWave transmission technology, UE's GP is crucial for predictive beamforming. Due to the high path loss and directional nature of mmWave signals, beamforming is required to focus signal transmission in specific directions and to ensure that the optimal beam pair is selected for communication. Beam search is a fundamental process used to establish and maintain a reliable link between the base station (BS) and UE. The challenge lies in selecting the optimal beam pair from a large set of possible beam directions. In a typical codebook-based beam search process, the BS and the UE maintain a set of predefined angular vectors that correspond to different spatial directions. During beam alignment, the BS transmits reference signals using a subset of the codebook, while the UE measures the signal strength of each beam and selects the one with the highest signal-to-noise ratio (SNR) [5]. The UE then transmits its response, allowing the BS to refine the selection and establish the best beam pair. This alignment process induces extra latency that may be undesirable, especially for fast moving UE.

ML models can "shortcut" the alignment process by predicting the optimal beam pair based on UE's GP and, potentially, other data sources such as radar, LiDAR or optical images. Supervised ML methods learn which beams perform best for specific locations and signal conditions [6], [7]. Once trained, a ML model can be used to predict an optimal beam pair without a time-consuming beam alignment. Reinforcement learning methods have also been explored for this task, e.g., [8], [9].

While initial works on predictive beamforming, e.g., [10], [11], used simulated data to demonstrate the utility of ML for predictive beamforming, DeepSense was the first and most widely used real-world dataset created for this task [12]. It comprises multiple scenarios corresponding to different scenes and data sources. It enables continuous benchmarking of predictive beamforming techniques, potentially in multi-modal settings. The authors of DeepSense maintain a list of scientific results obtained on this data corpus [13].

The original intent of this work was to carry out a systematic benchmarking of various predictive beamforming techniques across several basic scenarios. This goal was motivated by gaps in the existing literature, where the new methods often compared with a subset of previous works or on a subset of DeepSense's scenarios. However, we have identified some new research gaps in the course of this work, which constitute the main contribution of our paper. The core problem is the natural inaccuracy of GP data, due to complex physical phenomena of measurements in GNSS [14], [15]. As a result, the accuracy

978-3-903176-77-5

of predictive beamforming techniques from the recent literature, when systematically evaluated on DeepSense, appears to have reached a plateau. Hence we explore if this effect can be overcome by extending existing methods with novel normalizations, learning task formulations and combination of various data modalities in DeepSense data. Furthermore, we have identified yet another challenge in using DeepSense for benchmarking of predictive beamforming. The data exhibits substantial temporal non-stationarity, which calls for different evaluation and potentially novel learning techniques.

The main contributions of this paper can be summarized as follows:

- We present a framework for a systematic validation of various design elements in ML pipelines used in predictive beamforming techniques.
- We assess the impact of two original techniques, using data normalization and the regression formulation, on the quality of predictive beamforming.
- We assess the impact of the temporal nature of positioning data on the quality of predictive beamforming.

II. RELATED WORK

In the following, we briefly summarize the related work on predictive beamforming, the respective methods and metrics, as well as the technical features of the DeepSense dataset.

A. Data acquisition in ML-based beamforming systems

Real world data plays a crucial role in development and validation of ML techniques for wireless communications [16]. While synthetic datasets provide a controlled environment for algorithm development, they do not fully capture the complexity of real-world deployment. A ML model trained on synthetic data may include biases impairing the resulting solution. Recent studies address this problem by incorporating real-world data for calibration of simulated models of communication environments [17] [18] aiming to achieve high-fidelity synthetic representations.

DeepSense [19] is a multi-modal dataset collected from real-world wireless testbed deployments. Each deployment is equipped with communication and sensing hardware and designed to capture a set of measurements across different data modalities. The testbeds generated 44 datasets covering various scenarios. A significant amount of research in predictive beamforming relies on this dataset to demonstrate the viability of the proposed solutions.

B. Technical features of DeepSense

In this paper, we focus on *vehicle-to-infrastructure (V2I)* communication data (Scenarios 1-9) containing between 854 and 5964 samples, depending on the scenario. The technical setup of the V2I scenarios in DeepSense comprises two units: a static BS and a moving UE. The BS is equipped with a mmWave phased array, an RGB camera, a radar and a 2D LiDAR (in Scenarios 8-9). In most cases, the BS is positioned near the roadside, except in Scenario 6, where the antenna is placed at a greater distance from the road. The UE is installed on a vehicle that moves along one or several road lanes during the data collection. Further V2I scenarios (Scenarios 31-35) feature a variety of obstructions, such as cars, pedestrians and buildings, in the Line-of-Site (LoS) between the transmitter and receiver, emulating more complex field-of-view (FoV) geometry and wireless environments. These, as well as the remaining scenarios of DeepSense, covering vehicle-to-vehicle, pedestrian-to-infrastructure, static or indoor use cases, are beyond the scope of this work.

The transmitter and receiver operate at 60 GHz. The receiver is a 16-element Uniform Linear Array (ULA) configured with a codebook of 64 predefined beams. By sweeping through the codebook it captures signals across 64 angular directions. The beam sweep runs at 10 Hz [12]. The UE is equipped with a GPS receiver operating at 10 Hz for real-time positioning updates. Horizontal positional accuracy is 2.5 m without real-time kinematic correction (RTK) and 10 cm with RTK. The raw data collected using the testbeds undergoes post processing including synchronization and filtering to generate the final datasets.

C. Methods

The predictive beamforming problem, addressed in the majority of works, entails a communication system comprising a multi-element antenna array at the BS and a moving UE. The quality of the signal received at given time at the BS depends on the channel between the UE and the BS and the chosen beamforming vector. The beamforming task is to maximize the received signal power at the BS at each time step by selecting an optimal beamforming vector.

Conventional solutions for a beamforming problem are obtained by acquiring sufficient channel state information (CSI), or an exhaustive search across all beams, or combination of both. For large antenna arrays and scenarios with highly mobile UE, such methods lead to increasing complexity and a substantial beam selection overhead, which consumes network resources. Therefore, the proposed ML methods aim to solve the problem by utilizing side channel information generated by various sensors.

Such methods vary in data modalities and the processing pipeline architecture. Proposed solutions incorporate new techniques into a ML pipeline or evaluate diverse combinations of features extracted from different data representations. The most common approach is to predict the optimal *current beam*, which is accomplished by analyzing sensor measurements during the same time step as power readings. Another approach is to predict the optimal *future beam* at the next time step, which is denoted as beam tracking. The latter task involves the analysis of temporal dynamics and uses data sequences from a window of previous time steps.

A typical ML architecture for predictive beamforming comprises a feature extraction block in different variations, a feature fusion mechanism for merging different representations, and a final classifier. The feature extraction block depends on the data modality and is designed to reduce high-dimensional input to a compact set of features suitable for the downstream

tasks. This usually includes several stages, each processing the input with specific data modality, or the same data modality but with different algorithms.

D. Metrics

Assessment of beam prediction quality requires careful consideration of the appropriate metrics. The majority of previous ML-based approaches treat beam prediction as a multiclass classification problem (i.e., selecting the best index among several options). Hence the respective quality metrics, e.g., accuracy, precision, recall and F1-score [20], [21], can be naturally applied to assess the beam prediction performance. However, these metrics may not capture the physical semantics of beamforming. In practice, there may exist several "adjacent" channels with similarly high received power. To account for such an effect, distance-based accuracy (DBA) has been introduced [18], [22]–[24].

DBA counts wrong predictions if they fall within a specific range from the ground truth beam index. This range is defined by a parameter. While DBA indicates higher values when the predicted index is close to the ground truth, it lacks the information of the exact subset of beams which is required to further refine the selection. Another metric, *Top-K Accuracy*, often deployed in the previous work, indicates the ratio of data samples for which the ground truth index falls within the set of K indices with the highest probabilities output by the prediction block. This measure denotes the percentage of data samples for which the truly optimal beam can be found by searching over K most likely optimal beams suggested by the ML component. Another potential metric [23] measures how often the predicted index falls within K indices that exhibit the highest ground-truth power levels.

The *Power Ratio* [23] metric aims to quantify the physical performance of the beam prediction by comparing the power level of the predicted, potentially suboptimal, beam with the power level of the optimal beam according to the ground truth. Previous works deployed several methods for computing the Power Ratio. For example, the work in [22], [25] accounts for noise by subtracting the minimum power value per sample or per scenario. In [18], [26], [27], the Power Ratio is defined on a logarithmic scale. Additionally, spectral efficiency of approaches based on reinforcement learning is estimated in [28]. The trade-off between the complexity of the model and performance is assessed with a special metric in [25].

III. DATA CHANNELS IN DEEPSENSE

Intended for investigation of multi-modal predictive beamforming, the DeepSense dataset comprises four types of different sensor data: GPS coordinates, radar, LiDAR data and RGB images, used as input for the prediction of the signal power received at the mmWave antenna. In this section, we elucidate the peculiarities of these data channels which are important for the design of our validation framework.

A. Positioning data

The UE position is determined by geographical coordinates collected with a *GPS-RTK module* and includes longitude and latitude in decimal degrees. Different transformations and normalizations have been used for data prepossessing in prior works. For example, conversion to the cartesian coordinate system can be performed using Universal Transverse Mercator (UTM) projection [23] [20] and can be further transformed into polar coordinates [24] [27]. In some works, the data points are normalized to absolute values [26] or centered relative to BS. Several studies quantize the coordinates by mapping them to a set of discrete values [21], [26].

The geographical position sensory data is known to be inaccurate due to various reasons such as interference, atmospheric conditions, number and geometry of connected satellites, receiver quality, deliberate distortion, and others. Figure 1 illustrates the inaccuracy of the GP data in the DeepSense dataset. It shows different trajectories reflecting the UE location for each "run" of the vehicle carrying the UE in Scenario 9. The setting of this scenario include a BS and UE with a position sensor placed in a vehicle. The movement trajectory is one-directional, on the road lane next to the BS with a speed limit ∼40 km/h. The location has vehicle and cycling traffic. The data collection takes place during daytime.

The color map reflects the index of the optimal beam provided as the ground truth. Despite a vehicle moving along a straight road stretch (slightly deviating from a north-south orientation) in the same direction—implying that it should be detected approximately at the same position when entering the antenna's FoV and leave the latter also approximately at the same point—the recorded coordinates vary greatly. It is hence clear that each "run" was affected by substantial systematic errors depending on unknown exogenous conditions.

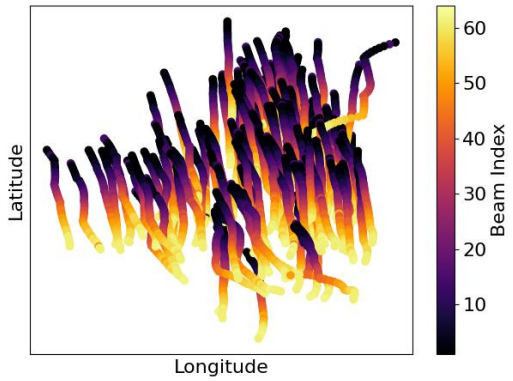


Fig. 1: Positioning data: Indices of best beams as a function of GPS longitude and latitude in Scenario 9.

B. RGB Camera data

The majority of prior works with the DeepSense camera data rely on pre-trained image classification and object detection models such as YOLO or MobileNet [29]. Preprocessing for model input can distort images or misalign, while high resolution data adds computation and latency. Further issues could concern the deployment of the cameras for data collection in view of regulatory restrictions and ethical

considerations. The crucial issue with using camera data as the main data source for predictive beamforming is, however, potential occlusion by other objects or the outright lack of LoS due to a specific scene geometry.

C. 2D LiDAR data

Scenarios 8-9 provide 2D LiDAR measurements in the form of point-cloud data in the horizontal slice of the environment. Each measurement represents a point defined by its position in space with respect to the LiDAR. The data is further processed and normalized; for example, [30], [31] apply static clutter removal method (SCR) to denoise the point cloud data by focusing on the moving objects and [32] utilizes the raw data and DL feature extraction to predict future beams.

D. FMCW radar data

FMCW radar data is processed through sequential FFTs on chirp samples and antenna channels to produce a Range-Angle matrix, which encodes the probability of object presence in the radar scene. Other representations, such a range-velocity maps or radar cube [33] have been employed in ML-solutions along with additional signal processing methods to extract various features from a noisy radar image.

E. Power of mmWave phased array

DeepSense scenarios feature *16-element antenna array (ULA)* with a codebook of 64 predefined beams. Each beam is determined by phase weights for 64 horizontal directions within 90° FoV. Beam steering is achieved by adjusting phase shifts across antenna elements ensuring constructive interference in the desired direction and destructive interference elsewhere. Beam sweeping is rapid sequential selection of beams from all positions in a codebook, or a certain subset of beams. During the sweeping time interval, received signal power levels at each beam are measured and recorded. These measurements are grouped into times steps. Resulting power vectors contain 64 values for each beam in a time step.

IV. VALIDATION DESIGN

Heterogeneous prior work on predictive beamforming, based on the DeepSense dataset, uses various modalities, methods and specific model configurations. The main goal of our validation, to be presented in the remaining part of this paper, is to investigate the impact of various design decisions in such methods on the overall performance of the resulting system. We focus on the GP data as the input, as it is used in most of the previous works as input—either as a sole input or as one of the inputs for multimodal systems—due to its independence of occlusion. For simplicity, we focus on Scenarios 1–9, where line-of-sight conditions are met. More advanced scenarios introduce additional problem complexity, which makes investigation of general features of various predictive beamforming techniques more challenging. Specifically, we are interested in such features as the impact of the data normalization, the ML task solved by the prediction model (classification or regression), multimodality and temporal dependency in the evaluation protocol.

A. Datasets

The data collection took place across six different geographical locations. The dataset for each scenario contains a different number of samples recorded under daytime or nighttime conditions. The BS and sensors are installed on the side of the road with the FoV bisector nearly perpendicular to the movement of the vehicle in most settings, except Scenario 7, where it is directed slightly along the road at a small angle. The road width differs from two to six lanes. In Scenario 6, the antenna is positioned further from the road resulting in a larger FoV.

B. Related benchmarks

Among previous studies focusing on GP-aided beam prediction, the closest work related to our setup is [26], covering the same Scenarios 1-9 and considering accuracy and power ratio metrics. It is also consistent with the top scores from the DeepSense repository.

Direct comparison with other results on the GP-aided beam prediction is rather difficult. Different studies used different scenarios, dataset modifications and performance metrics. For example, [34] is mostly focused on image-based solutions and does not specify the details of the position-based classification model. Moreover, it uses a different dataset from the drone communication Scenario 23. The same scenario is used by [35], [36] where the comparison is further complicated by the downsampling of the number of beams from 64 to 32.

The downsampling of a codebook is investigated in detail in [20] with the data from Scenarios 1-9. This study similarly provides the results only for 32 classes or less. While downsampling of classes was shown to benefit the accuracy, this method might not scale efficiently for systems with narrow beams and therefore large codebooks. An alternative modification of the dataset from Scenario 9 was used in [21], where label grouping was applied. A common power vector was assigned to the specific areas on the grid, which reduces the power noise associated with different UE locations.

Other works consider datasets from Scenarios 31-34 that feature more complex vehicle trajectories and crowded scenes. The DBA, Top-1,2 Accuracy and Power Ratio are presented in [27], while [24] provides only DBA as an assessment metric. The work in [23] covers various performance metrics and introduces an additional filtering method.

C. Methods

We consider the following four methods throughout our validation protocol:

- M0: baseline method re-implemented as described in [26].
- M1: same input and architecture as M0 except that the final layer of the network implements the regression problem, as described later in Section IV-E.
- M2: classification problem with a single normalized feature, as described later in Section IV-D.
- M3: same as M2 but extended with the second input feature of the GP data (longitude).

Besides the comparison between these four generic methods, we are also interested in whether such a comparison is affected by using the physically motivated metric (Power Ratio) and by the sampling of data points in the ML experimental protocol (random or sequential).

D. Normalizations

As mentioned in Section III-A, different normalization techniques have been used in previous work to transform the positioning data into a convenient form for processing in the ML pipeline. Nevertheless, the noisy character of GP data, as evidenced in Figure 1, suggests that the underlying "semantics" of the position communicated to the model may be poorly represented, even after the standard normalization. To address this issue, we introduce a new normalization method based on the underlying movement pattern observed in the DeepSense data. Notably, in all scenarios of our interest, the UE transmitter moves through the FoV of the antenna along the longitude. Hence it can be expected that under the LoS conditions, which hold for Scenarios 1-9, the maximum received power moves from the first to the last beam index during the UE movement. Since the UE moves on an almost straight line, the maximum received power should move at the same speed as the speed of the vehicle (most likely, constant) across all beam indices if the antenna is calibrated accordingly.

The intuition of the constant speed movement of the transmitter across the receiver's FoV is further illustrated in Figure 2, based on a subsample of data from Scenario 9. For each K, corresponding to separate "runs" of UE, it shows on the vertical axis the latitude values scaled between 0 and 1, based on the smallest and the largest values *received in that "run"*. The color map represents, similar to Figure 1, the beam index with the optimal ground-truth power. It can be clearly seen that, unlike the raw GPS data displayed in Figure 1, the latitude data alone, normalized between 0 and 1 for individual "runs", exhibits a more consistent, albeit still not fully uniform, dependency between the geographical position and an optimal beam index.

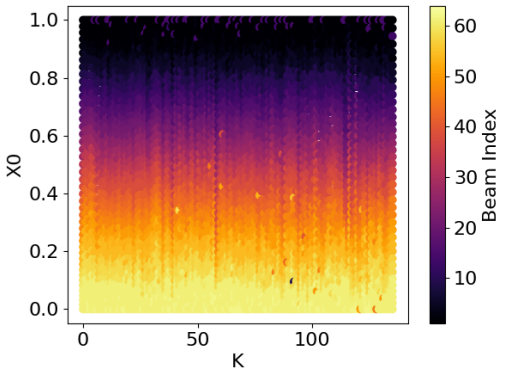


Fig. 2: Normalized Latitude

In real operation, such a per-run normalization is obviously not possible. For benchmarking based on the DeepSense dataset, it represents an assumption that the GPS receiver is consistently calibrated so that the values at the extreme points of the antenna's FoV are approximately the same.

E. Regression vs. Classification

In previous work on predictive beamforming, the last layer in the ML pipeline implements a classification task, i.e., a decision which beam index is expected to have the best signal power. Most classifications methods have a "soft-max" output layer which can assess the probability of each class given the input. This enables more flexible decisions than simply selecting the most likely index, e.g., choosing K most likely indices and using other criteria to decide among them.

Despite such flexibility, the model is inherently trained to predict a *single best beam index* since it is given a single index as the ground truth at the training stage. For beam selection, the difference between different reasonably good indices, may not be very significant in practice, and providing the "winner-takes-all" ground-truth may already exclude some valuable information from the training.

As an alternative, we propose in this study to consider a regression formulation of the beam prediction problem. In this formulation, the model is given the power values at all 64 beams as ground truth at training and is requested to predict all such values for each new data point. A decision on the beam index selection during the model deployment can be made, similar to classification with a soft-max output, based on some rule involving the ordered set of predicted power values. Albeit the regression formulation may appear largely underdetermined at the first glance (since the size of the output space by far exceeds the size of the input space), the intuition is that, similar to multiclass classification, the model weights "learn" intrinsic dependencies between input values and the expected distribution of the output power values from the large corpus of training data.

Formally, a regression model M predicting the power vector P_M is trained by minimizing the mean square error loss (MSE) between the predicted values P and the ground truth values $\hat{P}$:

$$L = \|P_M(x_0, x_1) - \hat{P}(x_0, x_1)\|^2$$

The model uses the same MLP architecture as described in Section IV-C. It consists of an input layer with either one or two units, depending on the input dimensionality, three fully connected hidden layers, each with 128 units and Leaky ReLU activation function. The output layer is a 64-dimensional vector, with each element representing power level on one of the beams. The MLP is trained minimizing MSE loss with a batch size of 32. Initial learning rate is set to 0.01 and decreases by a factor of 0.2 when no improvement in validation loss is observed for 5 consecutive epochs.

F. Validation metrics

To allow the comparison of results across various datasets and validation criteria, we use the two commonly used metrics for predictive beamforming, introduced in Section II-D:

$Accuracy@K$ and $PowerRatio@K$. For the regression formulation presented in Section IV-E, the metrics are defined as follows.

$Accuracy@K$ measures the ratio of samples for which the ground-truth beam index falls into the set of indices defined by the top K predicted signal powers:

$$Accuracy@K = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left(k_n^{\text{opt}} \in \text{Top-K}(P_n)\right)$$

where k^{opt} is the index of the ground truth beam, P is predicted power vector, N is the number of samples.

$PowerRatio@K$ is a physical metric that measures power loss from wrong predictions. For the regression formulation, it reflects the ratio between the best among the top K predicted power values and the ground-truth power of the best beam:

$$PowerRatio@K = \frac{1}{N}\sum_{n=1}^{N} \frac{P_n^{\max}}{P_n}$$

where P is the maximum received power in the optimal direction and $P^{\max}$ is the power at the chosen beam from top K predictions.

V. EXPERIMENTAL RESULTS

The experiments to be presented in this section aim to shed light on the utility of the essential design parameters of various approaches for predictive beamforming, to the extent this is possible on the basis of the DeepSense dataset and its Scenarios 1-9. In particular we are interested in the following research questions:

1) Can beam index prediction quality be improved by novel elements of a ML pipeline?
2) Can data quality be improved by appropriate preprocessing and/or normalization?
3) Can beam index prediction model be reliable over a prolonged interval of time?

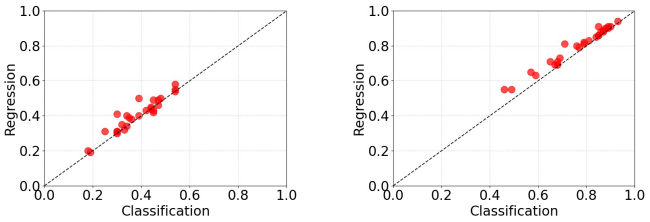
A. Model performance across datasets

Our first set of experiments compares the beam prediction performance for the four methods M0—M3 described in Section IV-C. Performance metrics for Scenarios 1-9 are presented in Table I, respectively.

In general, it can be observed that the results obtained by novel methods do not provide a substantial improvement over the baseline method M0. The regression technique M1 has a slight advantage over M0 in terms of $Accuracy@K$ but remains largely on the par with M0 in terms of $PowerRatio@K$ (more detailed analysis is provided in Section V-B). The two normalization techniques incorporated in methods M2 and M3 provide substantial improvements over M0 in some scenarios, e.g., 8 and 9, but also suffer substantial quality setbacks in other scenarios, e.g., 1, 6 and 7. Among the two normalization techniques, M3 has a slight advantage over M2 in terms of accuracy, apparently because the longitude does enable to differentiate between vehicles traveling on different lanes in some scenarios.

B. Comparison of classification and regression formulations

A better understanding of the difference in quality between the classification and regression formulations can be obtained by considering the scatterplots of the performance metrics for these methods reported in Table I.



(a) $Accuracy@1$ (b) $PowerRatio@1$

Fig. 3: Regression vs. Classification

These scatterplots display the pair of corresponding results for regression and classification as a point on 2D surface; if this point lies above the dotted line $y = x$, the regression method outperforms the classification, and vice versa. It can be observed in the scatterplots that regression outperforms classification in $Accuracy@1$ metric in the majority of considered scenarios. This is even more prominent if $PowerRatio@1$ metric is concerned, where regression outperforms classification in all experiments. However with more "relaxed" metrics, such as Top-3 and Top-5, classification outperforms regression.

C. Temporal data shift

Stable performance of an ML-system in real-world relies on rigorous validation with the available dataset. A common procedure applied in previous work is to randomly divide the dataset into three parts in various proportions. For example, the work in [34]–[36] reserve 70% of the data for training and 30% for validation, while [20] tests 3 different proportions. The work in [26], [27] uses half of the validation set during training and another half to compute the final metrics. The same procedure is applied in [23] but with 10% of the unseen samples reserved for final validation. Overall, [20], [26] conclude that the proportions of the training and validation subsets in Scenarios 1-9 have minimal impact on the performance.

However, most of the studies do not provide additional details on the specific procedures used to split the data. Except for the study in [23] that applies stratified split to preserve the balance of classes and [20] mentions the application of the hold-out partitioning method. Due to the use of different datasets and the downsampling method, a direct comparison is difficult. To extend the previous analyses, we conduct experiments to compare the performance of the GP-based models with 2 basic types of data splits.

In line with prior work we adopt 70:10:20 ratio with 80% for training and intermediate measurements and 20% for final assessment. In the first experiment we randomly assign data points to each of the subsets, and in the second experiment the data points are split according to the timeline, in a sequential

TABLE I: Accuracy@K—PowerRatio@K

		Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
Top-1	M0	0.54—0.93	0.47—0.89	0.30—0.59	0.25—0.57	0.44—0.84	0.39—0.85	0.30—0.67	0.42—0.79	0.39—0.77
	M1	0.58—0.94	0.46—0.90	0.30—0.63	0.31—0.65	0.44—0.85	0.50—0.91	0.31—0.69	0.43—0.81	0.40—0.79
	M2	0.36—0.79	0.44—0.87	0.34—0.65	0.30—0.68	0.47—0.86	0.34—0.76	0.18—0.46	0.54—0.90	0.45—0.85
	M3	0.35—0.81	0.45—0.88	0.32—0.69	0.33—0.68	0.48—0.87	0.30—0.71	0.19—0.49	0.54—0.89	0.45—0.85
Top-3	M0	0.80—0.99	0.70—0.98	0.40—0.84	0.38—0.82	0.65—0.94	0.63—0.97	0.49—0.87	0.64—0.93	0.60—0.93
	M1	0.81—0.99	0.68—0.96	0.42—0.77	0.44—0.79	0.63—0.92	0.72—0.98	0.46—0.83	0.63—0.91	0.57—0.90
	M2	0.57—0.93	0.68—0.97	0.46—0.86	0.45—0.88	0.68—0.97	0.52—0.86	0.29—0.76	0.78—0.97	0.66—0.96
	M3	0.59—0.95	0.68—0.97	0.46—0.86	0.47—0.89	0.68—0.97	0.42—0.83	0.32—0.75	0.75—0.97	0.67—0.96
Top-5	M0	0.92—1.00	0.84—0.99	0.51—0.92	0.46—0.90	0.78—0.97	0.74—0.99	0.58—0.93	0.76—0.96	0.71—0.96
	M1	0.90—1.00	0.78—0.99	0.52—0.84	0.55—0.85	0.75—0.95	0.88—1.00	0.57—0.88	0.73—0.94	0.67—0.94
	M2	0.69—0.98	0.83—0.98	0.59—0.92	0.58—0.93	0.81—0.99	0.64—0.88	0.39—0.83	0.87—0.99	0.78—0.99
	M3	0.75—0.97	0.82—0.98	0.56—0.91	0.59—0.94	0.82—0.99	0.50—0.88	0.44—0.84	0.87—0.99	0.78—0.99

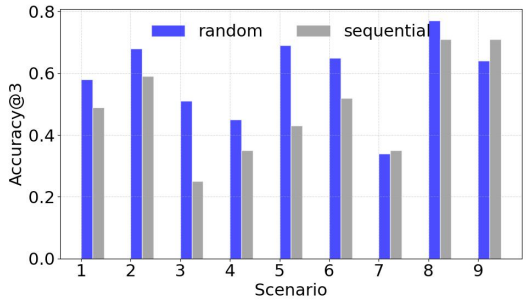


Fig. 4: Perfromance of M2 with different data splits across all scenarios

manner. We repeat training and validation for all models and datasets and observe that performance metrics with random split correspond to the previously obtained values and the benchmark. However, with the sequential split the performance is noticeably worse in most of the scenarios, except Scenario 7. In Figure 4 we provide the accuracy comparison for M2.

These results demonstrate the importance of a temporal component in the GPS error. While we describe the known drawbacks of GPS receivers in Section 3.1, our results show that in Scenarios 1-9 the performance of the ML-models still could be significantly improved by specific data permutations. However, to ensure stable performance of these systems additional details could be considered during the validation stage, e.g., the time windows of test and validation datasets, if the data exhibits a temporal non-stationarity. Better understanding of the nature of the GPS error, its development over time and predictability would also benefit the stability of GPS-based beamforming systems.

D. Multi-modal techniques

Finally, we study the impact of multi-modal techniques, combining various data sources. In this experiment, besides the GP data, we extract the coordinates of the maximum intensity points from a sequence of radar-angle maps. In each frame we identify the maximum value indicating a radar-angle bin where reflected signal is the strongest. A resulting sequence corresponds to the positions of the dominant object in a polar space. In Table II we compare the performance of the GP-based solution M3 extended with a radar feature in Scenario 9. The results show the potential increase in $Accuracy@K$ and $PowerRatio@K$. However radar data has limitations that affect its quality. The noise caused by environmental factors, reflections, interference and clutter can result in artifacts leading to target misidentification.

TABLE II: Multi-modal Accuracy@K—PowerRatio@K

Method	Top-1	Top-3	Top-4
GPS	0.45—0.85	0.67—0.96	0.78—0.99
Radar	0.39—0.76	0.59—0.92	0.72—0.96
GPS + Radar	0.48—0.88	0.70—0.97	0.82—0.99

VI. DISCUSSION AND CONCLUSIONS

Our validation of benchmark results on predictive beamforming using the DeepSense dataset reveals that the hitherto proposed methods have reached a performance plateau that cannot be substantially improved on by means of new methods alone. The main problem lies in the quality of the geographic positioning data which is a crucial input for predictive beamforming techniques. Various kinds of interference inherent for GP data present a substantial challenge to the downstream ML pipelines, especially for the multi-dimensional predictive beamforming task. Multi-modal techniques, which involve additional data sources such as images, radar or LiDAR, may be helpful, especially in more complex scenes and scenarios, yet their susceptibility to occlusions and other interference sources might be a serious obstacle to their practical utility. Our investigation of a straightforward combination of GP and radar data reveals only a minor improvement of the beamforming accuracy. A better understanding of technical issues causing the unpredictability of the GPS data quality is highly desirable for future progress in predictive beamforming.

The formulation of predictive beamforming as a regression task in the ML pipeline, proposed in our work, seems to have an advantage over the previous methods deploying classification as the learning task. Prediction of expected signal powers for different beams conveys valuable information for the

respective beamforming decisions. The utility of this approach should be further verified on better quality data.

Serious attention should be paid in future work to the non-stationarity of the data involved in predictive beamforming. Our experimental evaluation has revealed that the quality of predictive beamforming strongly deteriorates if the temporal nature of the positioning data is taken into account and validation is performed on the data collected *strictly after* the model has been trained. Previous results have used random data sampling ignoring potential non-stationarity, which might have a profound negative impact on transferability of ML-driven results into practice.

References

[1] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari, and F. Tufvesson, "5g mmwave positioning for vehicular networks," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 80–86, 2017.

[2] L. Chen, X. Zhou, F. Chen, L.-L. Yang, and R. Chen, "Carrier phase ranging for indoor positioning with 5g nr signals," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10 908–10 919, 2021.

[3] G. Torsoli, M. Z. Win, and A. Conti, "Beyond 5g localization via sidelinks in industrial iot scenarios," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2024, pp. 920–925.

[4] J. Kaur, M. Shawky, M. S. Mollel, O. R. Popoola, M. A. Imran, Q. H. Abbasi, and H. T. Abbas, "Ai-enabled csi fingerprinting for indoor localisation towards context-aware networking in 6g," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–5.

[5] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3gpp nr at mmwave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[6] 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 Journal of the Communications Society*, vol. 5, pp. 38–50, 2023.

[7] M. Q. Khan, A. Gaber, P. Schulz, and G. Fettweis, "Machine learning for millimeter wave and terahertz beam management: A survey and open challenges," pp. 11 880–11 902, 2023.

[8] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "Optimal beam association for high mobility mmwave vehicular networks: Lightweight parallel reinforcement learning approach," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5948–5961, 2021.

[9] C. Laskos, S. Dimce, A. Zubow, and F. Dressler, "Towards virtual to real-world transfer learning for mobile mmwave beam tracking."

[10] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep learning power allocation in massive mimo," in *2018 52nd Asilomar conference on signals, systems, and computers*. IEEE, 2018, pp. 1257–1261.

[11] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE access*, vol. 6, pp. 37 328–37 348, 2018.

[12] 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," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023.

[13] Wireless Intelligence Lab, "DeepSense 6G Challenges," https://www.deepsense6g.net/challenges/, 2023, accessed: 2025-03-19.

[14] V. Hamza, B. Stopar, O. Sterle, and P. Pavlovčič-Prešeren, "Observations and positioning quality of low-cost gnss receivers: a review," *GPS solutions*, vol. 28, no. 3, p. 149, 2024.

[15] X. Li, J.-P. Barriot, Y. Lou, W. Zhang, P. Li, and C. Shi, "Towards millimeter-level accuracy in gnss-based space geodesy: A review of error budget for gnss precise point positioning," *Surveys in Geophysics*, vol. 44, no. 6, pp. 1691–1780, 2023.

[16] G. Apruzzese, R. Vladimirov, A. Tastemirova, and P. Laskov, "Wild networks: Exposure of 5g network infrastructures to adversarial examples," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 5312–5332, 2022.

[17] S. Jiang and A. Alkhateeb, "Digital twin based beam prediction: Can we train in the digital world and deploy in reality?" in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2023, pp. 36–41.

[18] M. Arnold, B. Major, F. V. Massoli, J. B. Soriaga, and A. Behboodi, "Vision-assisted digital twin creation for mmwave beam management," *arXiv preprint arXiv:2401.17781*, 2024.

[19] DeepSense 6G, "DeepSense 6G: A Multi-Modal Dataset for Sensing-Driven Wireless Communication," https://www.deepsense6g.net/, 2024, accessed: 2025-01-31.

[20] K. K. Biliaminu, S. A. Busari, J. Rodriguez, and F. Gil-Castiñeira, "Beam prediction for mmwave v2i communication using ml-based multiclass classification algorithms," *Electronics*, vol. 13, no. 13, p. 2656, 2024.

[21] L. Marenco, L. E. Hupalo, N. F. Andrade, and F. A. de Figueiredo, "Machine-learning-aided method for optimizing beam selection and update period in 5g networks and beyond," *Scientific Reports*, vol. 14, no. 1, p. 20103, 2024.

[22] G. Charan, U. Demirhan, J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Multi-modal beam prediction challenge 2022: Towards generalization," *arXiv preprint arXiv:2209.07519*, 2022.

[23] K. Vuckovic, S. M. Hosseini, and N. Rahnavard, "Revisiting performance metrics for multimodal mmwave beam prediction using deep learning," in *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*. IEEE, 2024, pp. 881–887.

[24] Y. Tian, Q. Zhao, F. Boukhalfa, K. Wu, F. Bader *et al.*, "Multimodal transformers for wireless communications: A case study in beam prediction," *arXiv preprint arXiv:2309.11811*, 2023.

[25] A. D. Raha, K. Kim, A. Adhikary, M. Gain, Z. Han, and C. S. Hong, "Advancing ultra-reliable 6g: Transformer and semantic localization empowered robust beamforming in millimeter-wave communications," *arXiv preprint arXiv:2406.02000*, 2024.

[26] 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*. IEEE, 2023, pp. 1824–1829.

[27] S. Tariq, B. E. Arfeto, U. Khalid, S. Kim, T. Q. Duong, and H. Shin, "Deep quantum-transformer networks for multi-modal beam prediction in isac systems," *IEEE Internet of Things Journal*, 2024.

[28] M. Ghassemi, H. Zhang, A. Afana, A. B. Sediq, and M. Erol-Kantarci, "Multi-modal transformer and reinforcement learning-based beam management," *IEEE Networking Letters*, 2024.

[29] S. Imran, G. Charan, and A. Alkhateeb, "Environment semantic aided communication: A real world demonstration for beam prediction," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2023, pp. 48–53.

[30] 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.

[31] O. Rinchi, A. Alsharoa, and I. Shatnawi, "Deep-learning-based accurate beamforming prediction using lidar-assisted network," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–5.

[32] 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, 2022.

[33] 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.

[34] I. Ahmad, A. R. Khan, R. N. B. Rais, A. Zoha, M. A. Imran, and S. Hussain, "Vision-assisted beam prediction for real world 6g drone communication," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–7.

[35] G. Charan, A. Hredzak, C. Stoddard, B. Berrey, M. Seth, H. Nunez, and A. Alkhateeb, "Towards real-world 6g drone communication: Position and camera aided beam prediction," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 2951–2956.

[36] 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)*. IEEE, 2022, pp. 2727–2731.