**Explainable AI based Path Loss Prediction for Rural Microenvironment**

### **Abstract**

This study presents a measurement campaign conducted in rural environment using a 5G nomadic platform to evaluate path loss characteristics in sub-6 GHz band. The measurements were performed in a vineyard environment, leveraging the nomadic platform to capture real world-data signal propagation data. Additionally, the study compares the performance of empirical models, ML, and Explainable AI models for path loss prediction. The results demonstrates that the explainable AI models provides better results in complex environments and also sheds light on the important explanatory features.

# Introduction

Smart farming (SF) involves a variety of digital technologies to increase the efficiency of farming operations. These technologies require high data rates and large bandwidths for stable and reliable wireless networks [1]. Farms are typically located in rural and sparsely populated areas, where the benefits of SF are constrained by the lack of reliable internet connectivity (Heikkilä et al., 2022). As Public LTE/5G cellular infrastructure is not universally available in rural regions., alternatively SF requires tailored wireless networks to meet high-speed broadband demand.

To address above mentioned challenges 5G Nomadic networks are considered a viable solution for smart Farming (SF) in rural environments [2]. These networks are designed to provide short-term on-demand connectivity solution for rural environments where local conditions change frequently.

Radio network planning plays a critical role in 5G deployments, and it is particularly important for rural environment, where terrain characteristics and landscapes strongly influence wireless signal propagation. Planning for nomadic 5G networks is even more challenging, as each site presents unique microenvironment that affect wireless signal interactions and received signal strength differently. As a result, signal performance may differ significantly between sites, even over short distances [3] [1].

In this context, the goal of this paper is to find the effect of rural microenvironment on the path loss behaviour of the 5G signals. To achieve this goal, we performed a comprehensive assessment of various models that would assist in predicting path loss. This work is part of NoLa project, which focuses on designing a nomadic 5G platform for rural applications, while evaluating it for real world use cases (e.g., smart farming, disaster relief, and road construction).

Traditionally, empirical path loss (PL) prediction models and machine learning (ML) models have been used for path loss (PL) prediction in different environments. Empirical models are transparent and computationally efficient, but their predictions in complex micro environments are less accurate [4]. On the other hand, ML models are accurate and handle with microenvironment well, but their lack of transparency makes it hard to trust their predictions. To balance transparency and accuracy, our previous work [5], used explainable boosting machines (EBM) to predict PL by using existing data of a university campus environment [6].

In contrast, this paper presents new path loss dataset obtained as a result of measurement campaign performed in a vineyard of Rhein-Mossel valley. Further, we compare the path loss values with predicted values from various models. In particular, we examined how well EBM Local explanations explain the reasoning of the model prediction. The main contributions of this work are summarized as follows.

* Systematic literature review of path loss modelling in rural environments.
* This work presents **a novel path loss dataset** for rural microenvironments, collected through an an extensive measurement campaign using ORAN compliant 5G equipment and spectrum analysers.
* Comparison of Bayesian Log-distance and XAI based NGB, EBM, and EBM2NGB path loss models in rural vineyard environment in terms of **accuracy**.
* **Global feature importance analysis** by comparing EBM and NGB-SHAP explanations for path loss prediction
* **Local feature importance analysis** by comparing EBM and NGB-SHAP explanations at individual prediction level.

Paper structure: Section 2 reviews the Related work, Section 3 describes the measurement environment (including study area, 5G measurement setup, path loss dataset, and evaluation models), Section 4 presents results, Section 5 concludes with a discussion on implication of results.

# Related Work

The path loss models are essential tools for designing and optimizing network coverage by estimating the path loss. To systematically review the existing body of knowledge, we classified the identified literature in two general topics: (i) 5G path loss modelling for rural environment, (ii) Explainable ML studies of PL prediction. Following subsection, explores the studies that focus on each of these topics.

## Path Loss modelling for Rural environment

To identify the relevant work regarding 5G path loss modelling in rural environment we conducted enriched research from the databases Scopus, Web of Science, IEEE Xplore, Springer, and Science direct.  To encompass a vast spectrum of pertinent papers, we defined the following search string.

"path loss model" and "5G" and "rural"

The initial search yielded a total of 313 papers. Since the search embodied several databases, we removed the duplicates of the papers, which reduced the number of papers to 299. The screening step involved a title search and we selected the papers that were applicable to 5G path loss modelling in rural environment, resulting in a total of 65 papers. The next step involved a thorough examination of the abstract and a rigorous review of all the publications to see if the findings were applicable to our area of interest and excluded path loss modelling studies involving drones and railway network in rural context. This scrutiny resulted in 11 papers for our study. The identified literature can be broadly classified into two categories. The first category of studies utilizes typical empirical models to predict the path loss in rural environment, while the second category of studies utilize advanced modelling techniques like ML and DL to predict PL.

Table 1 summarize the findings from first category of papers. These studies utilized empirical models to predict path loss across a wide range of frequencies (from 800 MHz to 75 GHz) in various rural scenarios. The table reveals that there is no universal empirical model that is valid for all rural scenarios. In low-frequency ranges (1-15 GHz), traditional models like WINNER II and Okumura-Hata remain relevant for path loss prediction (Schumacher et al., 2019) [7]. In contrast, for higher frequency ranges (16-60 GHz), there is more reliance on advance empirical models like Close-In (CI) model [8], and FI model [9] for path loss prediction.

Table 1 Empirical models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Paper | Frequency | Scenario | Rural Environment Details | Location | PL Model | Performance Indicator | Important Results |
| [10] | 3.4 GHz- 3.8 GHz | urban, suburban, and rural | Suburban slope and Rural village | Zurich, Ittigen, Meikrich, (Switzerland) | WINNER II , ECC-33 Model, SUI Terrain C Model, 3GPP RMa NLOS Model | RMSE, comparison with measured path loss | WINNER II D1 rural NLOS is best match for rural conditions; other models generally overestimate path loss. |
| [8] | 28 GHz, 38 GHz, and 73 GHz | urban, suburban, and rural | Rural area with large-scale path loss | Hong Kong (China) | Close-In (CI) model, The ABG model | Estimated path loss, MAPL, Cell Radius, Cell Area, Base Stations | Close-In (CI) model provided best fit for estimating cell radius and coverage in rural areas. |
| [7] | 15 GHz | urban, rural | General rural scenario with tri sector actenna| location not specified | Okumura-Hata model is used for rural scenario, Macro cell propagation model used for urban and suburban scenario | Throughput, Fairness index, Spectral Efficiency | Okumura-Hata model suitable for rural environments; statistical results prove that the 15  GHz spectrum is available for use |
| [9] | 40 GHz | Rural | Rural macrocell with minimal foliage | Tanjong Karang, Selangor (Malaysia) | Empirical FI (Floating Intercept) model, CI model | RMSE | The FI model is most effective and shows the lowest RMSE, The CI model is effective in Cross-Polarized Directional antennas under LOS conditions |
| [11] | 60 GHz | Green House (Rural) | Greenhouse (rural-like) environment | Universidad Nacional de Colombia, Bogotá campus (Colombia) | 3GPP Indoor Office model (InH-LOS), Weissberger model | Path loss | 3GPP InH-LOS model fit best in propagation parallel to furrow; Weissberger worked well for propagation perpendicular to furrow |
| [12] | 25.5, 26 GHz, 800 MHz | Rural | Rural Finnish Forest with dense coniferous trees, 40–700 m vegetation depth | Pornainen (Finland) | FITU-R, Weissberger, COST235, KAIST1, KAIST2, MED (Aalto1), MA (Aalto2) | RMSE | Aalto1 best overall fit; KAIST2 also suitable in high vegetation depth. |
| [1] | 28,38,60,75 GHz | Urban, Rural | Rural macro scenarios in simulations | 5GCM, 3GPP, METIS, and mmMAGIC | Path loss | 3GPP RMa model best-suited for rural macro environments; 5GCM not suitable in rural scenarios. |

Moreover, the Aalto1, KAIST2 (Saba et al., 2022) and Weissberger models [11] demonstrated better adaptation to Environment-specific factors like vegetation and terrain features. Similarly, for Rural macro scenarios 3GPP RMa model [1] and FI model [9] showed better performance. Although empirical models are less complex, they often fail to incorporate local environmental parameters and they are limited by environment. Due to these limitations empirical models provide less accurate path loss prediction [4].

The second group of studies explored advanced ML and DL models to overcome the limitations of empirical models. We identified only four studies in this category for rural environment. Table 2 provides summaries of studies that investigated ML/DL models for path loss prediction in rural environments. The identified studies are focused on Sub-6 GHz bands and covers a variety of rural scenarios including coastal and vegetation-specific. These studies show a paradigm shift toward a environment-aware, data aware path loss modelling. While [13] and [4] emphasize AI's role in extracting site-specific features from geospatial and visual data, [14] stress the influence of physical terrain properties (vegetation, coastal surfaces) on high-frequency signals. A frequent observation is that deep learning outperforms empirical models in complicated situations, with RNNs and hybrid techniques (such YOLO+3D mapping) showing promise. Further, [15] applied ANFIS (Adaptive Neuro-Fuzzy Inference System) to fringe areas in Uttarakhand, India, using 1800 MHz drive-test data. Their hybrid model achieves lower RMSE than traditional empirical models (Hata, COST-231) while offering simplicity and generalization.

Table 2 Advanced models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Paper | Frequency | Scenario | Propagation Environment | Location | PL Model | Input Features | Performance Indicator | Results |
| [13] | 800 MHz, 2 GHz | urban, suburban, and rural | The dominant path is influenced by various clutter types, such as buildings and terrain | rural location 80 km north of the central of Tokyo (Japan) | DL (feedforward neural network (FNN)) | Aerial Photo, building maps, Path profile | RMSE, Mean  Absolute Percentage Error (MAPE), Pearson correlation  coefficient (PCC), | Estimation accuracy improves by using building occupancy images and path profile B, estimation accuracy is the highest in rural scenario |
| [14] | 3.5 , 3.8, 4.2 GHz | Urban, Rural | Four different coastal terrains (small pebble, Air-dry sand, big pebble, Wet sand) and vegetation areas (Pine, Cherry, Orange, Walnut) | Various locations | Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN) | Distance, frequency, Coastal Terrain type / Vegetation type | R-squared, MAE, RMSE | RNN predicts better than LSTM for both scenarios, Path loss for coastal terrains is higher than vegetation areas |
| [4] | 850 MHz | Rural | Irregular terrains, buildings, and vegetation | Intelligent SPPL system based on Deep Learning and Computer Vision, FS, LN, Cost-231, MES | GPS location data, pitch, yaw, and roll information, 2D images, Satellite images | Mean Absolute Percentage Error (MAPE) | Proposed SPPL model outperformed empirical models after a distance of 33 meters. |
| [15] | 900, 1800, 2300 MHz | Urban, Suburban, Rural | Propagation path around a transmitter varies, different clutter type in buildings  hills, valleys, and a busy road| Dehradun, Uttarakhand (India) | Adaptive neuro-fuzzy (ANFIS)  Hata, COST 231, Ericsson, ECC-33, and Egli path loss models | ANFIS utilizes frequency, signal strength, distance | RMSE, MAE | ANFIS outperforms empirical models, achieved a RMSE of 11.2046, maximum RMSE 82.5012 observed in ECC-33 model |

## Application of XAI to PL Modelling

While the application of ML methods to PL modelling has received considerable attention in the recent signal propagation literature [16], extending interpretability (XAI) to ML-based PL prediction models remains a research gap in the wireless literature [17]. ML-based PL prediction models remain black-box tools that do not well represent the reasons for their predictions in the underlying propagation environment. Through a systematic literature review, [17] shows that only a subset of ML-based PL studies (XAI-based PL prediction models) apply XAI to clarify the logic behind the corresponding ML models for PL prediction.

XAI explanations can be divided into two categories: global and local explanations. Global explanations provide overall patterns of the effect of the input features across an entire dataset. Local explanations focus on the effect of the input features to generate a specific prediction within the dataset. Existing XAI-based PL prediction models often use global explanation techniques to uncover the overall logic behind their decisions, e.g., [18], [19], [20]. Local explanations to explain specific local PL prediction patterns within the dataset are partially addressed in [17] and [21], using LIME (Local Interpretable Model-agnostic Explanations) (to explain the feature importance of two randomly selected points) and EBM (to explain the marginal feature contributions at randomly selected points), respectively.



Focusing on local XAI is crucial because the global XAI approach can only explain the average PL behavior across the entire terrain. The reasons for abrupt changes in PL behavior at specific locations or signal behavior in sub-areas that deviates from global expectations are of great interest for efficient radio network planning. These can only be achieved by applying local explanations across the entire communication environment. To our knowledge, our work is the first to adequately apply local XAI to PL predictive modelling in rural environment.

# Measurement Environment

This section outlines the methodology employed in the study to measure 5G signal propagation in rural environment. The methodology is divided into three components. First the Study area is described with geographical context, second 5G Measurement setup is detailed. Finally, the process of compiling the path loss dataset is presented. and (3) path loss dataset.

## Study Area

The scenario selected for the rural path loss measurements was a vineyard located in the Middle Mosel Valley, a region renowned for producing some of the finest and most distinguished Riesling wines. The specific vineyard site, referred to as "Arena", is situated in the Bernkastel-Wittlich district of Germany.

The selected vineyard as shown in Fig. 1 is characterized by steep slope, varied elevation, foliage, and supporting ferromagnetic materials (such as posts and wires). It is representative of typical steep-slope viticulture, where annually recurring tasks such as soil cultivation, defoliation, and spraying are among the most time-consuming and challenging. To maintain the financial sustainability, farmers are increasingly required to replace traditional practices with innovative smart technologies. A key enabler of this transition is the availability of data communication infrastructure, which must be present —at least temporarily—to support the management, and control of complex intelligent agricultural robotic machinery. In this context, 5G and beyond technologies can play a critical role.



*Figure1 Visualization of Arena vineyard in Bernkastel-Wittlich*

## 5G Measurement Setup

The 5G network utilized for this study was a customized 5G nomadic platform as shown in Figure 2. The presented system is O-RAN compliant and consist of the following components: an outdoor Radio unit (Benetel RAN650 ), an outdoor directional antenna (Alpha Wireless AW3924), a 3GPP Rel 16 compliant 5G Core (Genius core ), RAN Software (Airpuls RAN based on the OpenAirInterface project), a PTP Grandmaster Fronthaul switch (μFalcon-RX ), a Network Switch (MikroTik Cloud Router CRS312-4C+8XG-RM), and Edge Cloud. All equipment is housed in a ruggedized 19’’ rack along with power backup batteries. This entire setup is mounted on a mobile cart that also includes a mast for mounting the radio unit. For precise timing, an outdoor Global Positioning System (GPS) antenna is attached to the Fronthaul switch. Internet connectivity is provided by starlink satellite modem. The relevant configuration parameters for the measurement campaigns are summarized in Table I.

The 5G signal strength measurements were conducted using a the Keysight Nemo Handy solution installed on a smartphone (Samsung S23). The Nemo Handy solution supports 5G NR network measurements and logs various signal strength metrics to the smartphone’s internal storage. Since the log files are in Keysight’s proprietary format, we used the Nemo Outdoor software to post-process the logs.

|  |  |
| --- | --- |
| *Figure 2 Architecture of Nomadic platform* | *Figure 3 nomadic platform deployed in Arean* |

*Table 3 Configuration parameters for measurement campaign*

|  |  |
| --- | --- |
| Location | Arena |
| 5G NR Frequency | 3.75 GHz |
| SSB Frequency | 3.748 GHz |
| Transmit Power | 37 +/- 2.5 dBm (at antenna port) |
| Tx antenna Gain (dBi) | 12.8 dBi |
| Rx Antenna Gain (dBi) | ~ 0 dBi |
| Bandwidth | 100 MHz |
| Tx Height (m) | 4 |
| Rx Height (m) | 1.5 |

## Pathloss data set

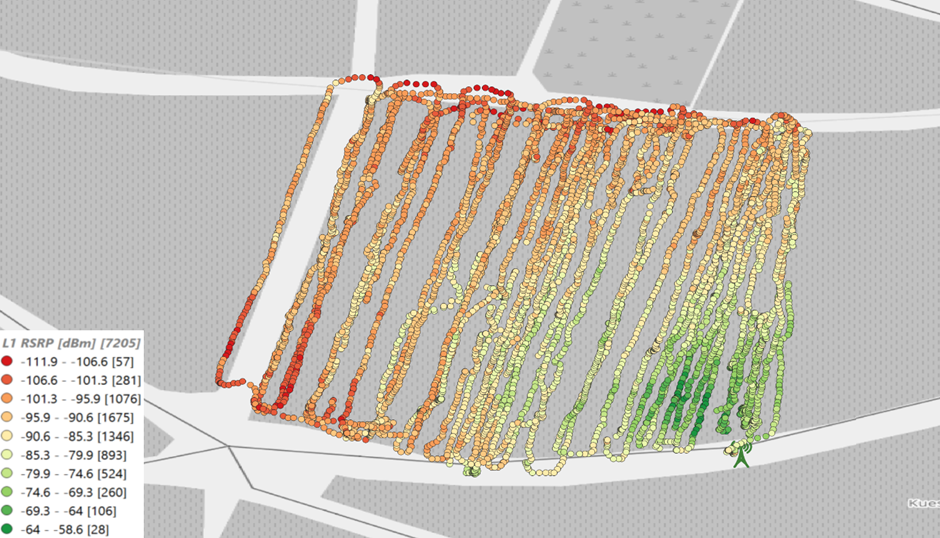
The signal strength measurements were conducted in the vineyard over a route of approximately 4700 meters (c.f. Figure 3). The route represents a realistic environment where intelligent agricultural robotic machinery is expected to operate. The measurement campaign involved traversing the vineyard rows twice using a Nemo Handy device, which enabled the collection of RSRP samples under various terrain conditions.

Figure 3 presents a heatmap of the RSRP (Reference Signal Received Power) values collected along the vineyard measurement route, with a total of 6246 signal strength samples visualized. The RSRP values are color-coded to represent signal quality, ranging from strong signals in green (e.g., -58 dBm) to weak and very poor signals in red (e.g., below -106 dBm). The bottom-right region of the vineyard shows predominantly green and yellow points, indicating strong to moderate RSRP values, typically between -58 dBm and -85 dBm. This suggests favorable conditions in this area, likely due to direct line-of-sight (LOS) with the base station and minimal obstruction. In contrast, the upper-left region of the vineyard is dominated by orange and red points, with RSRP values dropping below -100 dBm, indicating severe signal attenuation or possible non-line-of-sight (NLOS) conditions. This decline in signal quality across the route aligns with increasing distance from the antenna and the presence of potential obstructions such as vegetation, terrain variations, and structural elements within the vineyard.

Subsequent post processing of field measurements resulted in a comprehensive dataset for modeling path loss behavior in the vineyard environment. This dataset includes key variables such as longitude, latitude, elevation, altitude, clutter height, distance, and path loss values. The data set includes 7205 samples for most parameters and 6246 valid path loss values. A summary of dataset is shown in Table 4.

The spatial data—latitude, longitude, and elevation—provides high-resolution coverage across the vineyard, with elevation values ranging from 229 m to 259 m. The average clutter height is approximately 2.8 m, representing typical vineyard features like posts, wires and vegetation. The distance of measurement location from the base station varies from 0 m to over 103 m, covering both near-field and far-field propagation conditions. Corresponding path loss values range from 106.65 dB to 159.9 dB, capturing both line-of-sight (LOS) and non-line-of-sight (NLOS) scenarios. A standard deviation of 8.58 dB in path loss indicates considerable signal variation due to terrain and clutter.

The RSRP values and terrain information—including longitude, latitude, and elevation—were directly extracted from the Nemo Handy log files using the Nemo Outdoor post-processing software. In contrast, the clutter height, distance and path loss values were derived through post-processing techniques. Specifically, Clutter height was calculated in QGIS by computing the difference between the Digital Surface Model (SRTM DSM) and the Digital Terrain Model (EU-DEM), which provided estimates of above-ground object heights like vegetation and structures. The geodesic formula was used to calculate the distance between each measurement point and the base station, ensuring high accuracy over varying terrain. Finally, the path loss values were derived by combining the measured RSRP data with the Effective Isotropic Radiated Power (EIRP) of the transmitter, using the standard path loss equation.



*Figure 5 Measured RSRP values and route map in vineyard*

*Table 4 Descriptive statistics of data obtained from Arena*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Latitude (°)** | **Longitude (°)** | **Elevation (m)** | **Clutter (m)** | **Path Loss (dB)** | **Distance (m)** |
| **Mean** | 49.9132553 | 7.05010252 | 243.677307 | 2.78716242 | 137.296368 | 47.872833 |
| **Median** | 49.913239 | 7.050124 | 243 | 2.7 | 138.5 | 50.6577286 |
| **Standard Deviation** | 0.00024099 | 0.00031124 | 7.64354343 | 1.87998499 | 8.57765668 | 25.12818 |
| **Minimum** | 49.912846 | 7.049363 | 229 | 0 | 106.65 | 0 |
| **Maximum** | 49.913727 | 7.050567 | 259 | 6.10001 | 159.9 | 103.794726 |
| **Range** | 0.000881 | 0.001204 | 30 | 7 | 53.25 | 103.794726 |
| **Count** | 7205 | 7205 | 7205 | 7205 | 6246 | 7205 |

## Evaluation Models

## Bayesian Log-distance path loss model

Log-distance path loss model (Reference) is formally expressed as:

(1)

Where is the path loss (in decibels) at an arbitrary distance meters, is the path loss at a reference distance meters from the transmitter, is the environment specific path loss exponent that depends on the nature of the terrain, and is a normal (Gaussian) random variable with zero mean, reflecting the shadow fading with a mean of 0 and standard deviation.

To apply the Bayesian approach, we replace the in equation 1 with a Normal likelihood for the output with the expected value and , which is expressed in the equation 2:

(2)

Where:

(3)

Thus, we can rewrite the equation 2 as:

(4)

Equation 4 specifies the set of model parameters to be estimated comprising .

Hereby, we presume a normal distribution as the prior distribution for and . Thus:

(5)

(6)

The preferred choice for a prior on is a half-Cauchy distribution [22]:

(7)

In addition, we prefer the exponential distribution as a prior on is, as its concept is optimally defined for the first moment occurrences. Hereby, it describes the required the initial distance span from the transmitter until the signal weakening begins to occur:

(8)

is the mean distribution value and is computed through replacing the term in equation 1.

To extract the prior distribution values of the model parameters , we used the fitter package [[1]](#footnote-2) in python on the training data set.

The objective of the Bayesian approach is to determine the posterior probability distribution for the model parameters based on the above mentioned Normal likelihood of the output , the determined priors for the model parameters and the probability of observing i.e. .

(9)

As computing the exact posterior distribution is computationally intractable for continuous values, obtaining posterior distributions is accomplished via Markov Chain Monte Carlo (MCMC) algorithm to draw samples from the posterior. The sampling in our paper is done via PyMC3, which is an open source framework of probability distribution sampling [23] using MCMC methods to infer the model parameters .

## Natural Gradient Boosting (NGBoost)

In standard ML regression models, the prediction object is an estimate of a scalar function e.g. , where is comprising a vector of observed explanatory features influencing the as the prediction target. The NGBoost [24] is proposed to estimate the parameters of a probability distribution , where is a parameter vector describing the distribution. In our study, is containing the mean and the logarithm of the standard deviation of the distribution:

(10)

NGBoost incorporates three modules in a sequential way. First, it initiates two base learners in iteration , e.g. and to represent the mean and the standard deviation of the outcome, respectively. The majority of implementations use shallow decision trees as the base learners. Second, it uses Gradient boosting to sequentially training the base learners, in which each learner is optimized to minimize the current residual of the previous learners. Third, a scoring rule , which takes as input a forecasted probability distribution and one observation y (), and assigns a score to the forecast such that it minimizes the sum of the scores over the response variable from all training dataset. The distance between different distributions is typically evaluated using maximum likelihood estimator (MLE), which induces the Kullback–Leibler (KL) divergence between the true distribution and the predicted one. NGBoost’s main component comprises the leveraging of the natural gradient , which elaborates on the geometric structure of the parameter space [25]. In each iteration , and for each data , the algorithm computes the natural gradient of the with respect to the predicted parameters up to that iteration . The predicted outputs are scaled with stage-specific scaling factor , and a constant learning rate :

(11)

After iteration, the final prediction will be achieved through accumulation of the predictions resulted from all decision trees throughout the entire iteration steps.

The analysis of the data in our study based on the NGBoost python library [[2]](#footnote-3), which is built upon the Scikit-Learn package, and is designed to be flexible in terms of selecting proper scoring rules, distributions, and base learners.

## Tree-SHAP (SHapley Additive exPlanation)

SHAP (SHapley Additive exPlanations) [26] is a framework for explaining the ML models predictions. The SHAP basic principle is casting the original ML model by means of a surrogate model *g*, comprising the linear combination of the contributions of the model input features to generate *f* outcomes:

(12)

where, is the base value (intercept) of the model comprising the initial expectation about the mean model value without considering the contribution of the features, is the number of input features, is the contribution of each input feature , and is a binary value within the binary vector with regard to considering (1) or not considering (0) the contribution of the feature in the model output.

The contribution of each input feature is computed based on analyzing its impact on the model outcome by systematically modifying that input feature and observing the changes in the output. The basic formula to compute the SHAP value of the feature is:

(13)

where is the set of input features, is a subset of input features, is the expected values of the model when the feature is present, is the expected value of the model when the feature is not present, and is the weighted average of all possible subsets of in . The formula gets all possible subsets of features , which do not contain the feature, computes the contribution of adding the feature on the generated predictions in the all aforementioned subsets, and aggregate all resulted contributions to come up with the marginal contribution of that feature.

The notion of SHAP can be also extended to understand the interactive effect of multiple features beyond their isolated effects. Capturing pairwise interaction effects is introduced in [27]:

(14)

, where the represents the difference between SHAP values of feature in the presence of feature and the SHAP values of feature in the absence of feature .

Taking the interactive effects into account can be crucial in term of interpretation, as interaction between features can imply physical properties of the underlying environment. For example, the interaction effect of points distance from the transmitter in the terrain together with the transmission angle can include the Line of Sight or Non Line of Sight knowledge of the terrain, which for learning the PL behavior is of importance.

Since SHAP computation time increases exponentially with the number of features and the corresponding subsets , it is typically approximated from a number of subsets by regression models [28] or Monte Carlo methods [29] in practice. Though, based on the work in [30], Tree-based SHAP (by leveraging decision trees structures to disaggregate the contribution of each input in a decision tree) can achieve exact computation of SHAP values in polynomial instead of exponential time.

Computation of the SHAP values in our paper has been carried out based on the python SHAP package[[3]](#footnote-4).

## Explainable Boosting Machine

The basic idea behind EBM is analogous to generalized additive models (GAMs). GAM adopts a sum of arbitrary functions of variables (possibly nonlinear) that represent different features via splines, which altogether describe the magnitude and variability of the response variables [Reference]. For a set of multiple features e.g., and a univariate response variable e.g., , GAM is expressed by:

(15)

where, is intercept parameter, is representing independent variables in , is the dependent variable, and *h()* is the link function that relates the independent variables to the expected value of the dependent variable and represents a random variable. By using gradient-boosted ensembles of bagged trees for each feature function , EBM expands upon GAMs to preserve explainability while improving the predictive performance. Shallow decision tree generation, learning, and gradient updates in EBM are performed using a single predictor variable at a time in a round-robin fashion. The algorithm first builds a small tree with the input feature and computes the residuals (). It then fits the second tree with a different input feature to the residuals and goes on through all input features to complete the ongoing iteration. To mitigate the effect of each input feature’s order in the sequence, EBM incorporates a small learning rate. This renders the model to iterate through the training data over thousands of boosting iterations in which each tree only uses one predictor variable . Once the training of the ensemble of decision trees is completed, all trees associated with the single predictor variable will be summarized in a single function . Accordingly, all function associated with each predictor variable will be derived from the corresponding large set of shallow trees. In addition, EBMs take the combined impacts of two or more independent variables known as the interaction effect (GA2Ms). To compute the interactive effects, two-dimensional functions are learned to relate the response variable to pairs of predictor variables [31]. Hence, The EBM can be expressed by:

(16)

EBMs are highly interpretable, because the contribution of each independent variable or combination of independent variables to a final prediction can be visualized and understood by plotting and , respectively. The analysis of the data in our study based on EBM is done by the toolkit called InterpretML from Microsoft [32].

## Stacked EBM2NGB Model

Ensemble generalization ML Models [33] incorporate a number of standalone models (Base-learners) together to build a second-order model (Meta-learner), which exploits each individual Base-learner strengths to generate predictions with higher degrees of accuracy. In this paper, we propose a novel Ensemble generalization model EBM2NGB. EBM2NGB consists of two Base-learners: The first Base-learner is NGB, which delivers and as the mean prediction value and the standard deviation of the corresponding prediction, respectively. The second Base-learner is an EBM, which delivers as the expected prediction value given the set of input features . The Meta-learner is a second-order EBM model, which elaborates on the input feature vector comprising the outputs generated from the base EBM and the base NGB model, respectively:

(17)

## K-Means clustering

In this paper we use K-Means clustering [34] to figure out intrinsic grouping of data points based on their spatial location (comprising longitude, latitude) linked to the corresponding Path Loss values. Our goal is to identify distinct groups of spatially dispersed points in the dataset. Given a dataset, comprising the set of features , data points, and denoting the data point representing the values of the features in , which fits to one of clusters, the objective function of the K-means is to find:

(18)

, where is the sum of the squared error of all objects in database. and describe the clustering center matrix and the center of the cluster. describes the membership relations between the original dataset and the clusters, and is indicating the degree to which the data point fits in the cluster. The is the distance from point to cluster center . To optimize number of cluster on K-Mean clustering method, the elbow method can be used [35]. The elbow technique plots the variation of through increasing the number of clusters . The optimal is figured out at the point, where adding more clusters doesn't significantly reduce the .

# Experimental Results

## Model inputs, Model parametrization and Model Accuracy Results

The input feature of the Bayesian LD model only comprises the distance of separation between the corresponding transmitter and the receiver points. The input feature of the ML models include longitude, latitude, elevation, altitude, clutter height, and distance of separation between the corresponding transmitter and the receiver points. The output feature throughout our study is comprising the path loss values corresponding to each input data point. There are a total of 6244 data points along the measurement route. Though, based on decaying characteristics of the signal strength when it propagates from the base station towards different angles, we applied polar coordinate transformation to transform the geographical data into a polar coordinate system before training the model. Thereby, image axis-independent variables (longitude and latitude) are converted into polar coordination consisting of distance and the transmitting signal direction, i.e., Tx-Rx Angle. The Tx-Rx angle is computed by us through translating the longitude and latitude of the points to X-Y geographic points and setting the (X, Y) coordinate values of the transmitter location equal to the reference point (0, 0). The prediction accuracies of the models are evaluated based on splitting the data by a 1:4 training: testing ratio and by means of MAE (mean absolute error), MAPE (mean absolute percentage error), RMSE (root mean square error), and R squared. The overall performances of the 4 Models are presented in Table 3.

Table 4:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MAPE | RMSE | R² |
| Bayesian LD | 4.411 | 0.033 | 5.764 | 0.569 |
| EBM | 2.696 | 0.019 | 3.460 | 0.844 |
| NGB | 3.040 | 0.022 | 3.775 | 0.815 |
| EBM2NGB | 2.646 | 0.018 | 3.437 | 0.849 |

In order to inference the posterior distribution of the Bayesian LD model, we obtained the (mean, standard deviation) pair values (100.1589, 0.472) and (2.14013, 0.0268) via OLS regression to set as the prior distribution of the parameters and , respectively. Obtaining the residuals is done via subtracting the observed values in the dataset from the term as described in the equation (1). We then use the histogram of the overall s values (depicted in figure 1) to describe the s prior distribution. This results in choosing the normal distribution with (mean, standard deviation) values equal to (0.035, 5.659) as the best fit to s.

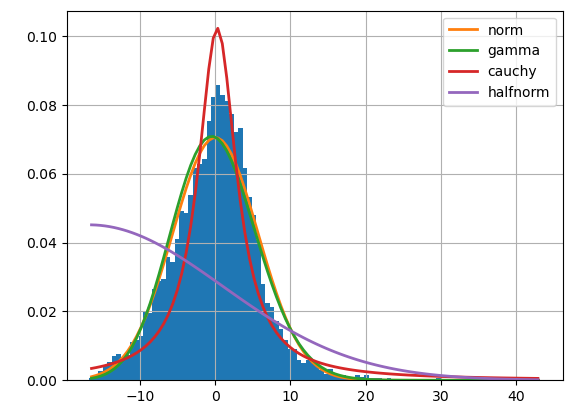


Figure 3

The resulting trace plots with regard to the posterior probability of the parameters via using No-U-turn sampler (NUTS) implemented in the probabilistic programming package for python PyMC3 are derived in accordance with the equations 1-9 and are depicted in figure 2. Each subplot in the left hand side panel comprises 4 different chains, each of them comprising 1000 draws (solid line: chain 1, dotted line: chain 2, dashed line: chain 3, and dot-dashed line: chain 4) from the posterior probability, which are depicted in the right hand side panels of the figure 2.

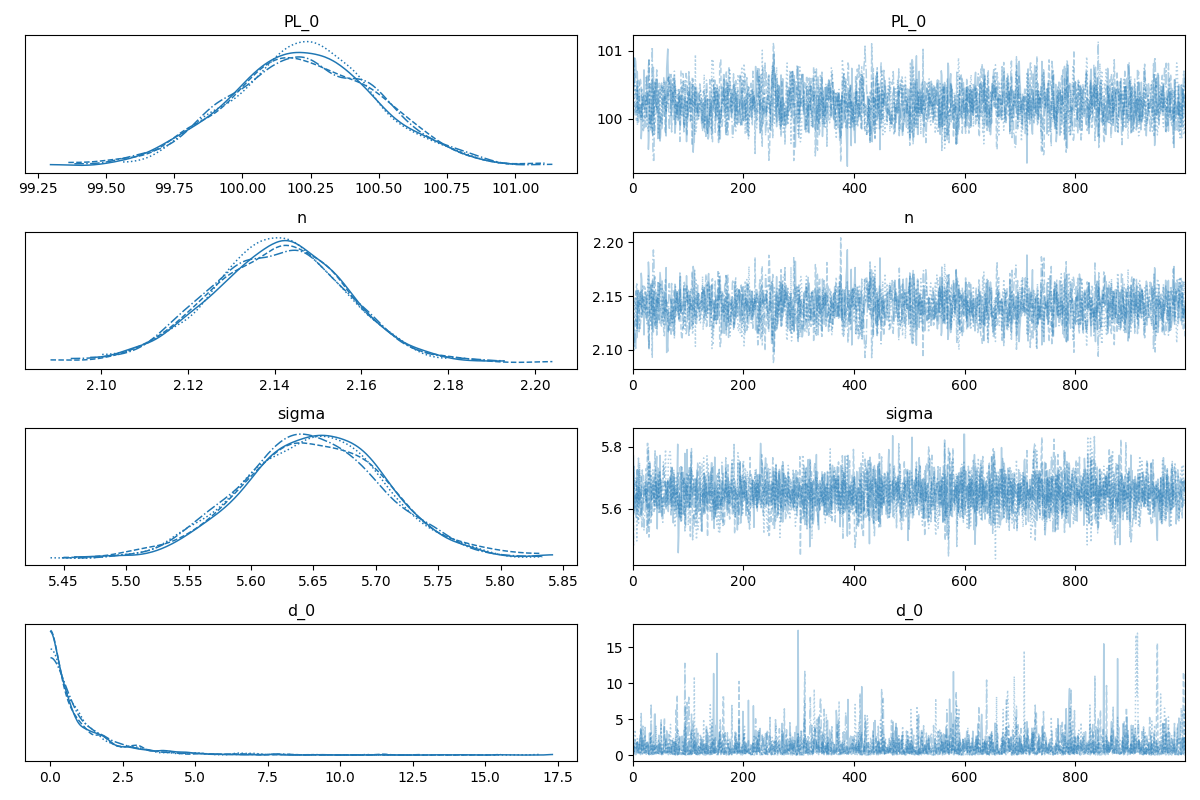


Figure 4

The fitted Bayesian LD model is tested by us on the 20 percent hold-out data set. As prompting the model to predict individual predictions given a specific distance, generates samples of 1000 posterior values, the results presented in figure 4 are comprising the expected mean of the predicted values together with the lower and the upper 95% confidence interval.

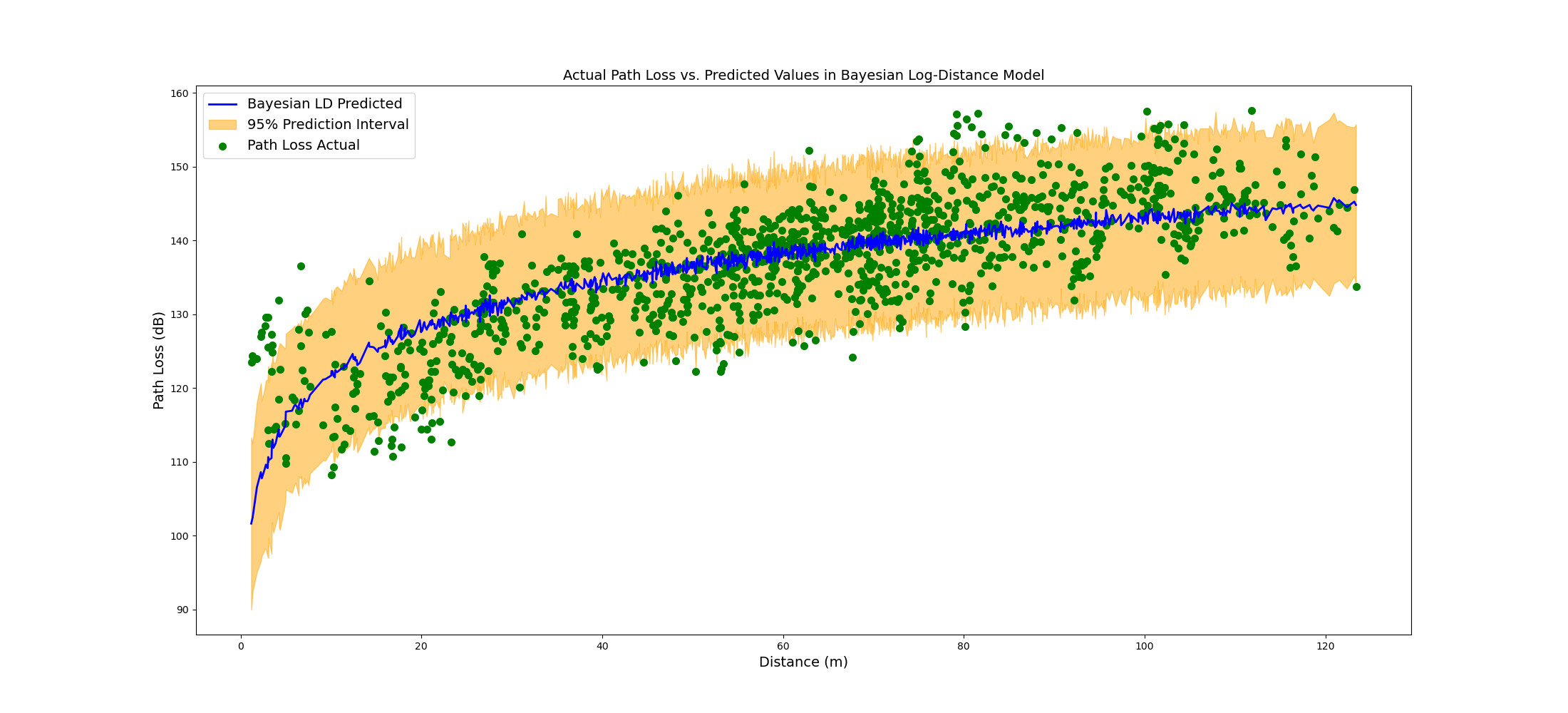


Figure 3:

The Bayesian LD model performs well in terms of covering the most points in the 95% of the confidence interval. Though, the suitability of the model to explain the expected predicted points in comparison with a horizontal line drawn at the mean *PL* value of the training dataset is not optimal. This is reflected in the R-squared value of the model, that represents the proportion of variance in the dependent variable (*PL*) that is explained by the independent variable (d), which is equal to 0.569. Indeed, only 56.9 percent of the variation in *PL* values is being explained by the factor distance via the Bayesian DL model. To elaborate more on the issue of un-explain-ability we apply ML models to the data.

The EBM model trained in this paper is hyper-parametrized by using a grid search through the parameter space: 0.01<“learning\_rate” (with 0.01 increments)<0.05, 1<“max\_leaves” (integer with 1 increments)<4, 2<“min\_samples\_leaf ” (integer with 1 increments)<6, “early\_stopping\_rounds”[100, 200], and “early\_stopping\_tolerance”[1e-6, 1e-5]. It results in utilizing 0.003 as the “learning\_rate”, incorporating the “max\_leaves” of the trees to be 2, setting the “min\_samples\_leaf ” equal to 4,“early\_stopping\_rounds” equal to 100, and the parameter “early\_stopping\_tolerance” equal to 1e-6 . The results of applying the EBM model on the 20 percent hold-out data based on the distance of separation between the corresponding transmitter and the receiver points as input and the predicted path loss as output are shown in figure 5.

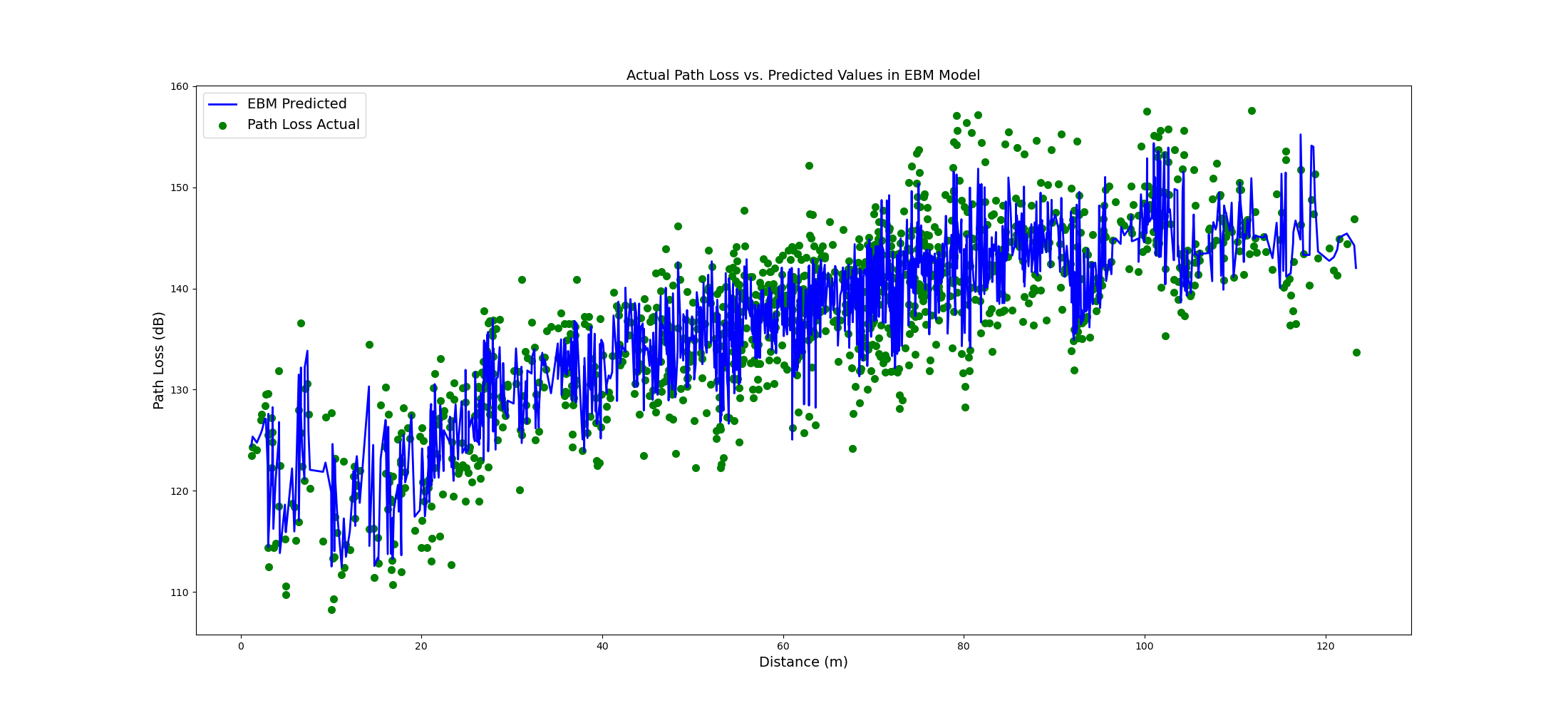


Figure 5:

In comparison with the Bayesian LD model, the EBM model performs well in terms of decreasing the MAE, MAPE, and the RMSE metrics to the lower corresponding error values. Especially, the R-squared metrics value is increased up to 0.844. However, the EBM point predictions might be exposed to the over-fitting. To elaborate more on the issue of uncertainty we apply the probabilistic NGB model to the data.

The NGB model trained in this paper is hyper-parametrized by using a grid search through the parameter space: 0.01<“learning\_rate” (with 0.01 increments)<0.05, 1<“max\_depth” (integer with 1 increments)<4, “minibatch frac“[0.5, 1.0], 1<“early\_stopping\_rounds“(integer with 1 increments)<11, and “distribution“[Normal]. It results in utilizing 0.001 as the “learning\_rate”, incorporating the “max\_depth” of the trees to be 2, setting the “minibatch frac“ equal to 0.5,“early\_stopping\_rounds” equal to 10, and the “distribution“ to be Normal. The results of applying the NGB model on the 20 percent hold-out data based on the distance of separation between the corresponding transmitter and the receiver points as input and the predicted path loss as output are shown in figure 6.

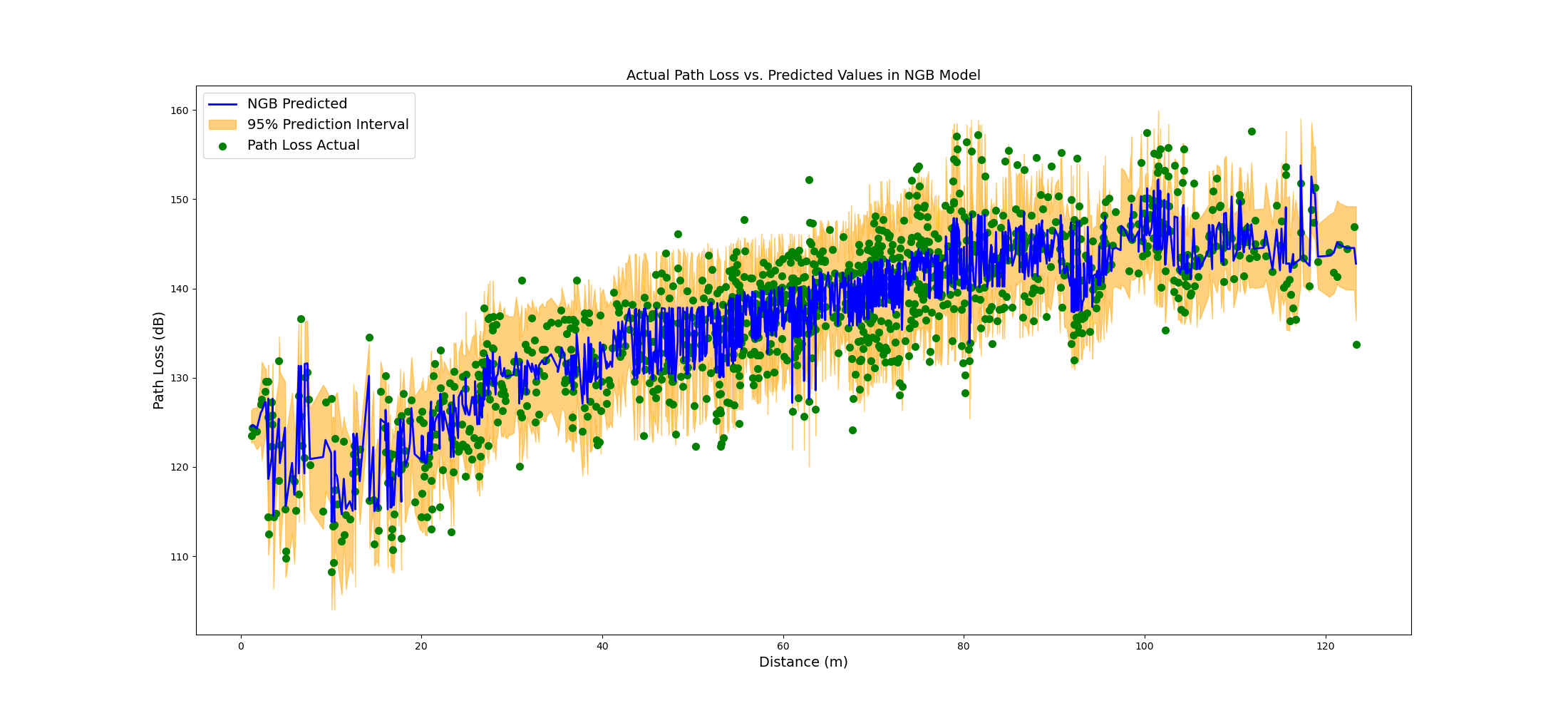


Figure 6:

In comparison with the Bayesian LD model, the NGB model generate more accurate predictions while preserving the probabilistic nature of the predictions. However, as presented in table 4, the delivered MAE, MAPE, and the RMSE metrics are slightly deteriorated. To integrate the probabilistic outcome of the NGB model with the more accurate EBM model, we further evaluate the probabilistic-informed ensemble EBM2NGB model on the hold-out dataset. As presented in table 4 and illustrated in figure 8, the EBM2NGB model metrics are slightly improved in comparison to the EBM model.

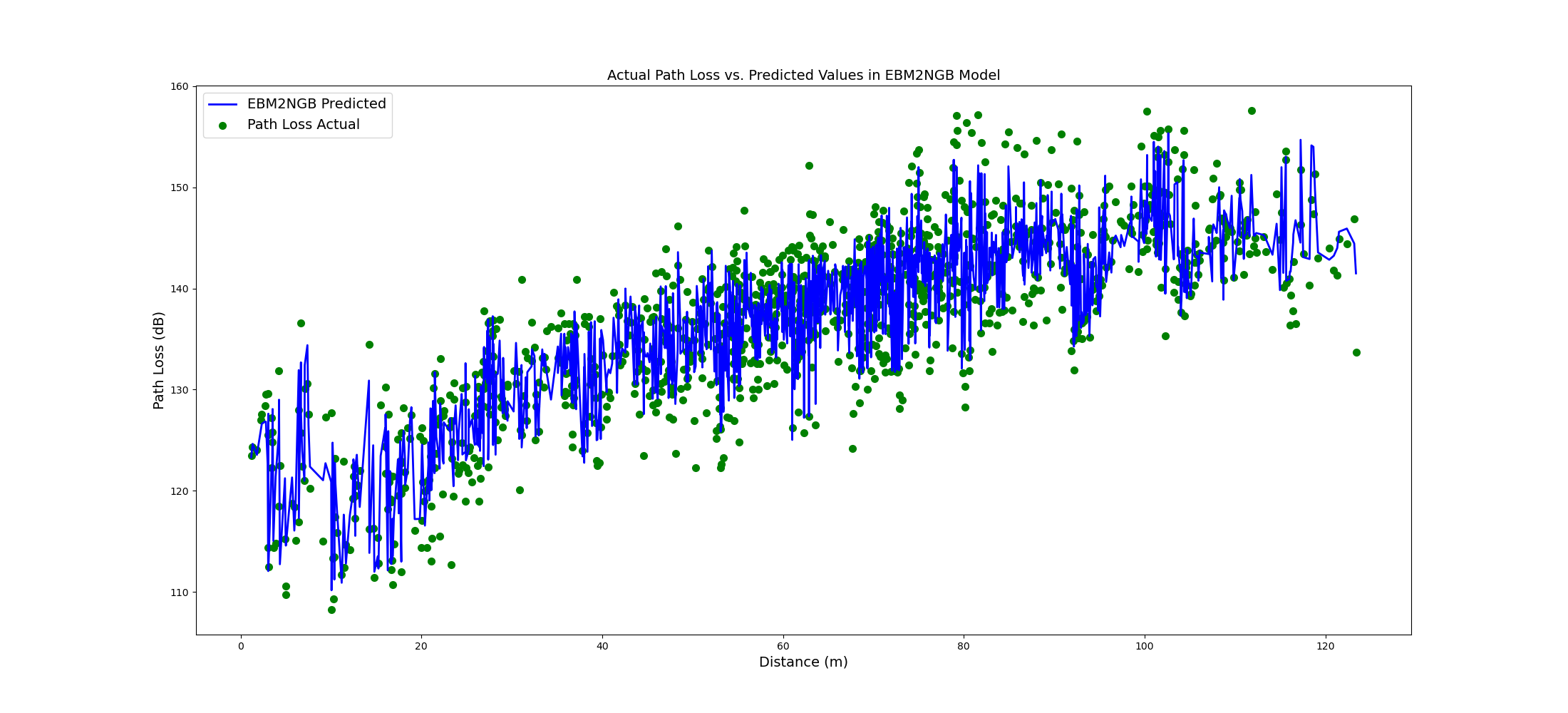


Figure 7:

## Model Global Explanations

This section elaborates on the rationale behind the EBM and NGB models to generate their predictions. From a global explanation point of view, we explain how both ML models decide to generate predictions over the entire dataset. The EBM model is a self-explainable, which does not require a surrogate explanation due to its glass box nature. In contrast, the explanations for the decisions behind the NGB model are drawn via the SHAP analysis.

As the result of training, the intercept of the EBM model (the parameter in equation 15 and equation 16) in our study amounts 137.329 dB. Likewise, the intercept of the NGB-SHAP model (the parameter in equation 12), which is the *PL* mean values in the training dataset is equal to 137.340 dB. The intercept for the parameter (equation 10), which is an indicator for measuring the mean uncertainty of the model is equal to 1.218, which is associated with 3.380 dB.

Figure 9 illustrates the average absolute additive (main and interactive) effect, which each feature in equation 16 can induce to change the outcome of the EBM model beyond the EBM model intercept. Hence, the values in figure 9 can be perceived as the Global expression of the features importance in the model predictions. Among all the features, the transmitter to receiver distance plays the most significant role in generating prediction. The interaction of the transmitter to receiver angle with distance, as well as the main individual effect of the transmitter to receiver angle, are the second and the third most important features globally contributing in *PL* predictions, respectively.

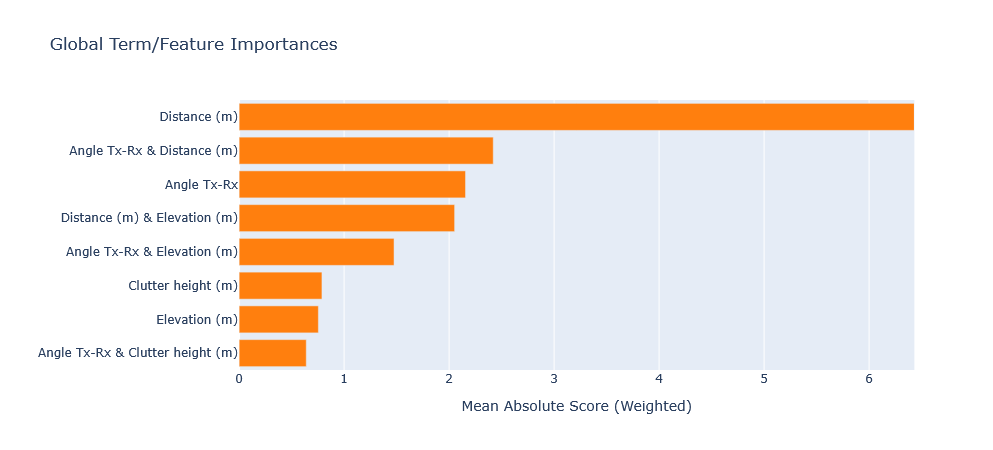


Figure 8

Figure 10 illustrates the average absolute additive (main and interactive) effect, that each feature in equation 16 can contribute to change the outcome of the NGB-SHAP model beyond the model’s intercept. In this figure, the values on the diagonal show the main features effect, and the other values show the SHAP interaction for pairs of features. Thereby, the main effect of the transmitter to receiver distance, the transmitter to receiver angle, and the elevation play the most important roles by contributing 5.87 dB, 1.1 dB, and 0.87 dB to the overall predictions, respectively. The interaction of the transmitter to receiver angle with distance is the next significant factor.

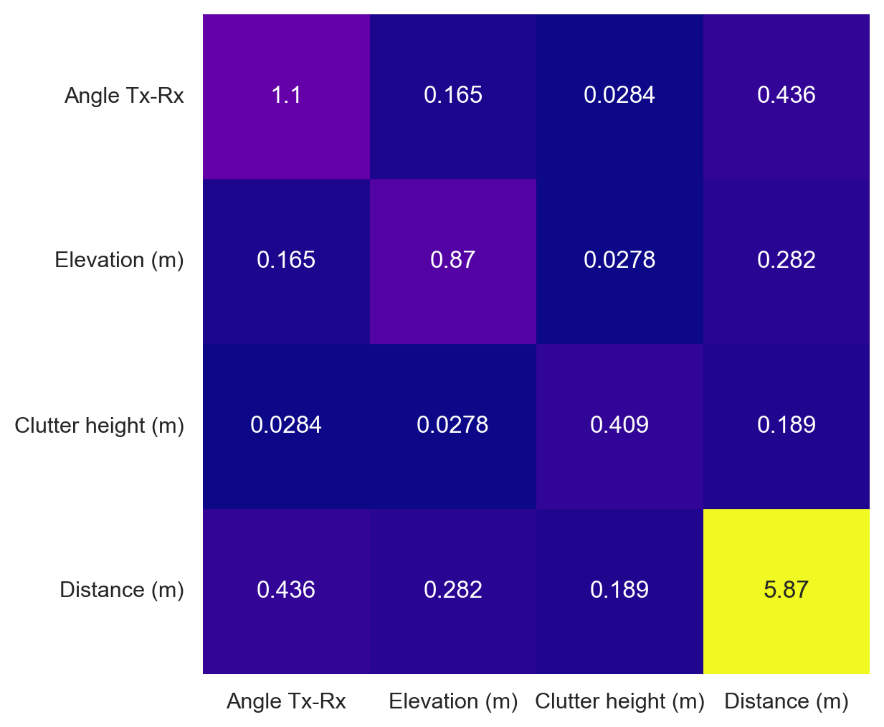


Figure 10

Figure 11 illustrates the average absolute additive (main and interactive) NGB-SHAP effects, with respect to the generated standard deviation in the NGB model. As shown in figure, the transmitter to receiver angle emerges as the main source of the uncertainty in NGB predictions. This phenomenon can be attributed to the directional antenna pattern in the study terrain, which likely divide the receiving points into separate signal coverage regimes within the vineyard.

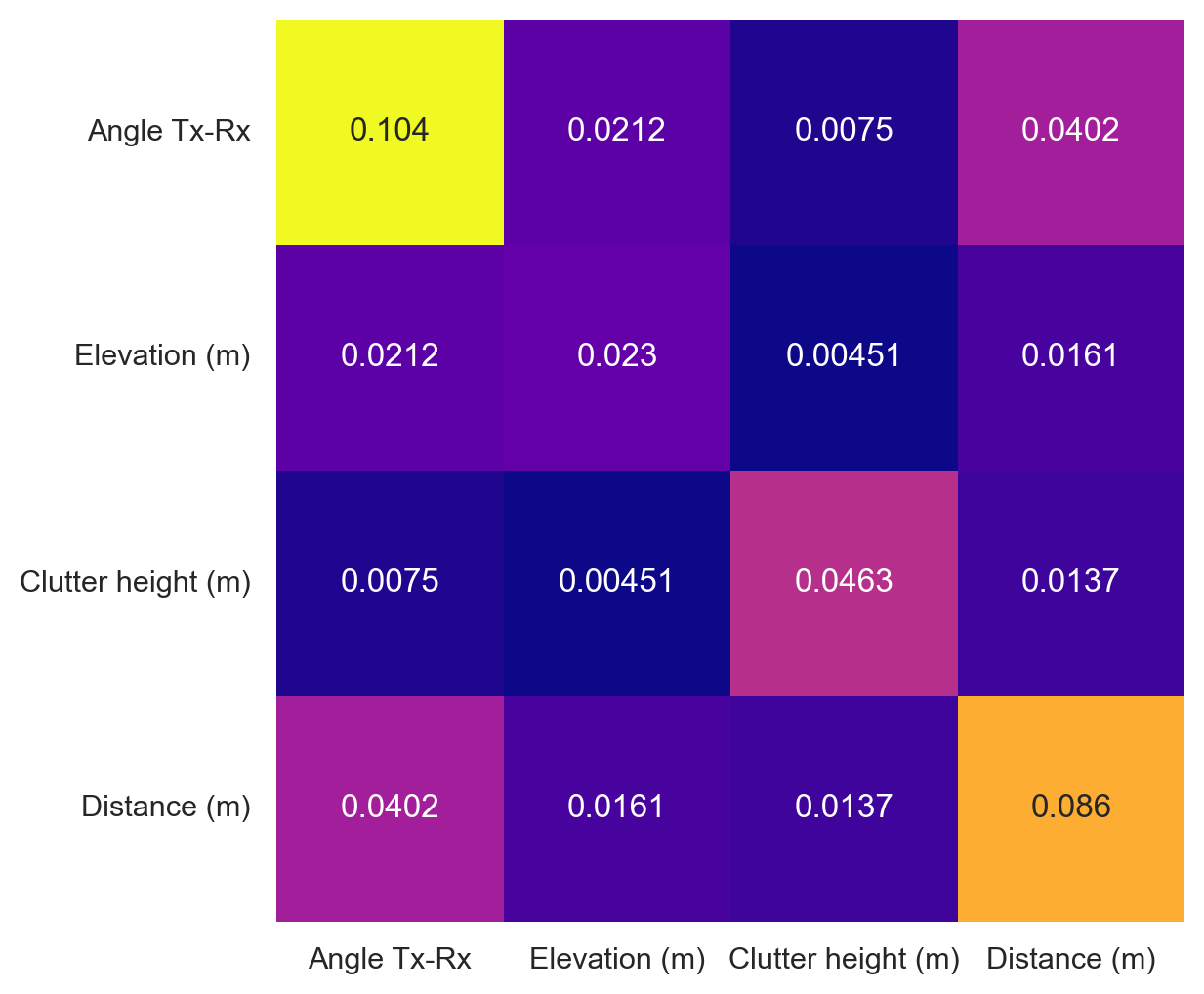


Figure 11 describes the features marginal contributions in the EBM model. Feature marginal contribution maps each possible value of an input feature to a corresponding contribution (in a look-up table manner) to generate PL predictions. The overall effect of the feature distance from the transmitter is shown in the upper left-hand panel in Figure 11. The effect of the feature distance in the vineyard reveals a linear characteristic, with a 45-degree rise from the minimum distance up to around 100 m and then with some near to zero but fluctuating overall slope from that point onward.

The upper right-hand panel shows the overall effect of the feature transmitter to receiver angle. The center point in the upper right-hand panel of the Figure 11 represents the location of the transmitter. The ring around the center point represents the PL strengthening or weakening when it transmits from the base station towards a specific direction. The effect of the Tx-Rx angle on the path loss in our study terrain can be understood based on the directional antenna pattern. Especially, the receiver points located between 270o and 315o have the best chance of establishing a line of sight (LoS) link to the transmitter. In contrast, the receiver points located between 45o and 225o fall outside the antenna’s main coverage area.

The overall elevation’s effect is shown in the lower left-hand panel in Figure 11. In general, higher elevations are associated with higher losses in received signal. However, this pattern becomes less consistent between elevations of approximately 245 meters and 260 m.

The overall effect of the feature clutter height (lower right-hand panel in figure 11) can be easily divided into two parts. For clutter heights up to 3 m, path loss tends to increase, for clutter higher than 3 m, path loss begins to decrease.

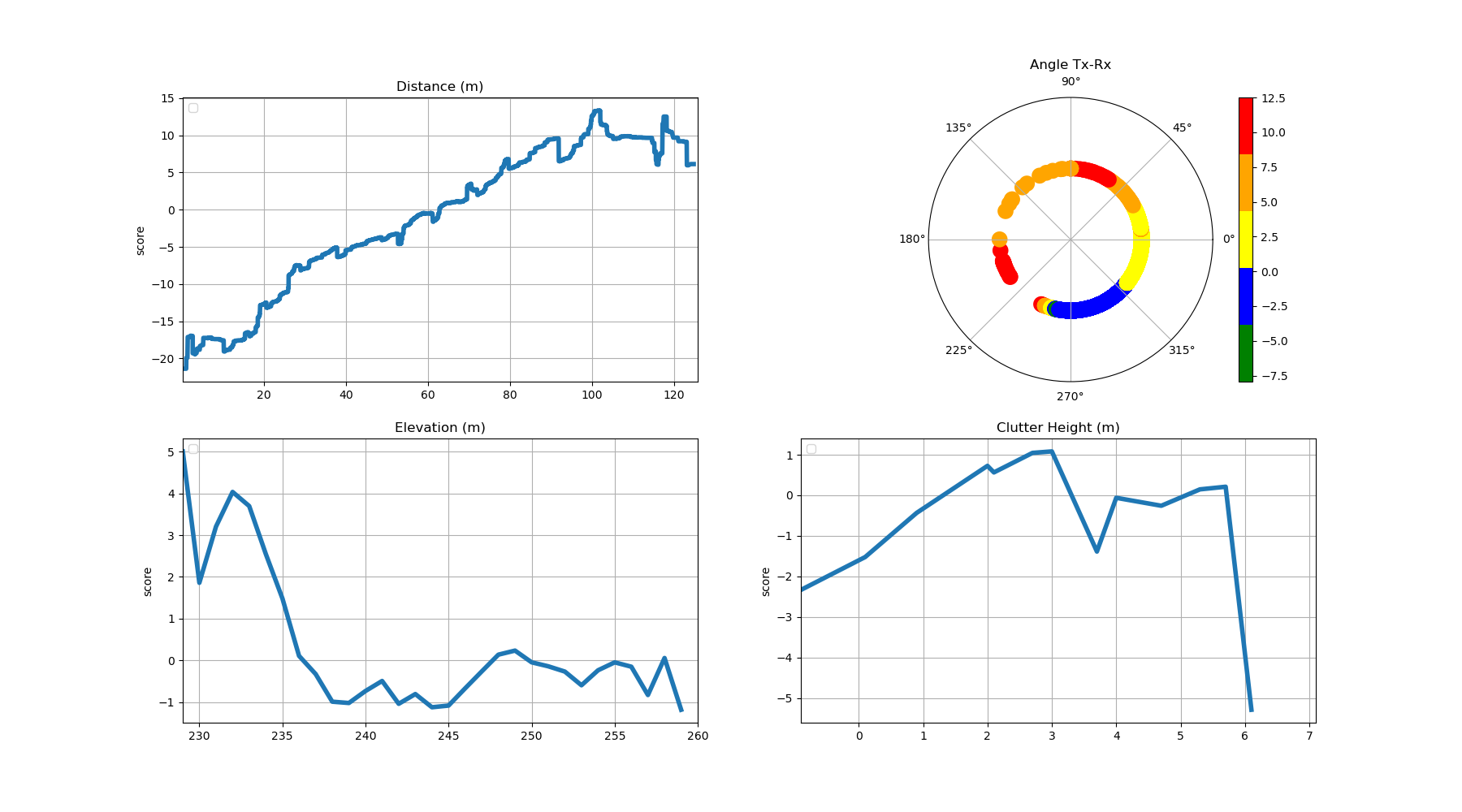


Figure 11

Figure 12 describes the features marginal contributions in the NGB-SHAP model. The sub-panels can be understood analogous to the sub-panels described in figure 11. Indeed, the overall patterns generated through the NGB-SHAP model comply to a large extent with those from the EBM. SHAP local explanations can generate different features marginal contributions values for the same feature value in each data point. This stochastic effect is reflected in the features marginal contributions in figure 12. In addition, while there are partial disagreements between the XAI generated based on the NGB-SHAP and the XAI generated based on the EBM mode. For example, the upper right-hand side of the figure 12 (based on the NGB-SHAP model) explains most of the receiver points located between 0o and 45o as the points having a chance to receive negative PL contributions from the transmitter to the receiver angle point of view. In contrast, the EBM based contribution (cf. Figure 11) associate same points with higher PL contributions. Additional trivial inconsistencies between the model explanation patterns can be seen by comparing each sub-panel in the figure 11 with its corresponding counterpart in figure 12.

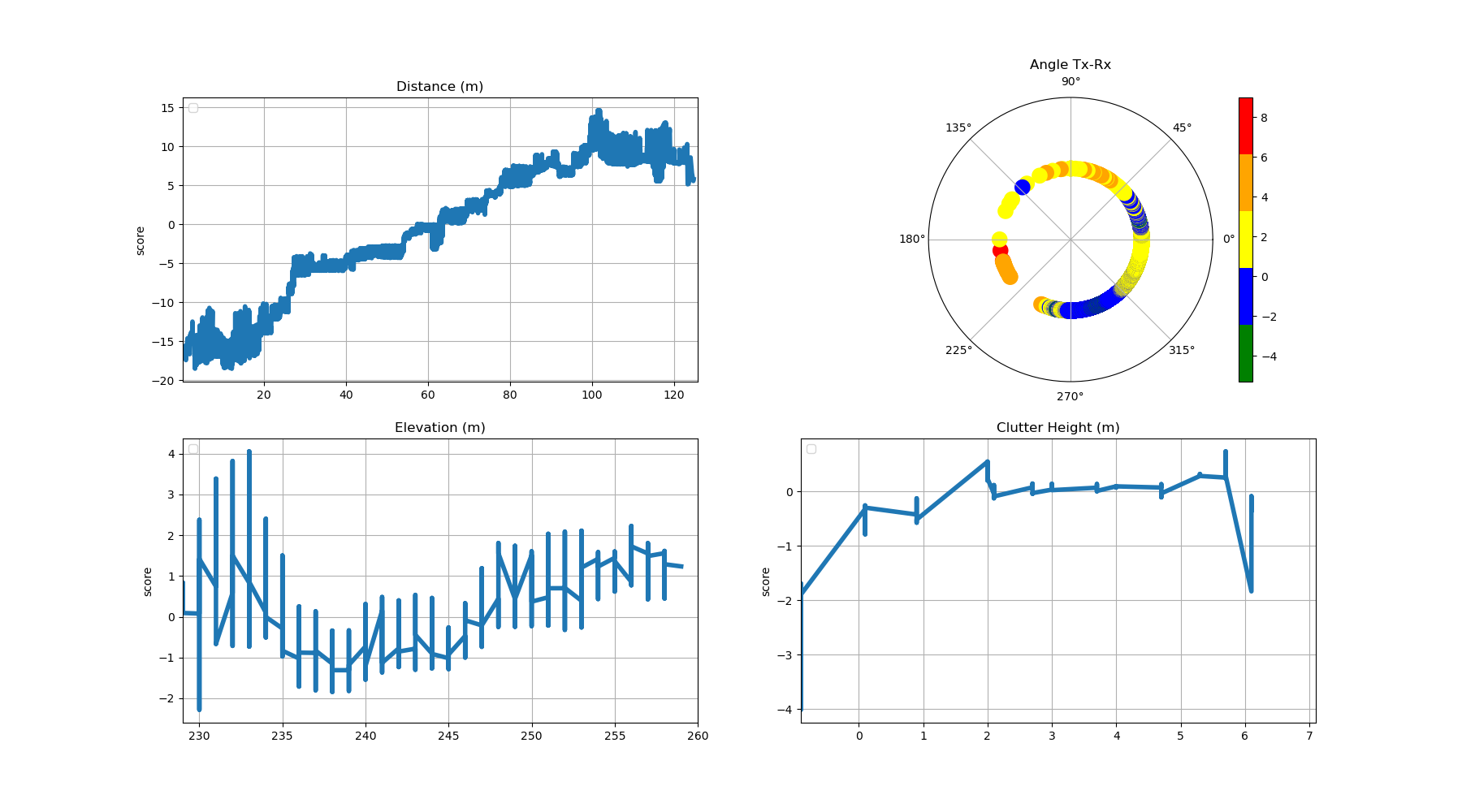


Figure 12

Figure 13 demonstrates the contribution of the interaction effects in the NGB-SHAP model (right-hand panel) and the EBM model (left-hand panel) explained by the equations 14 and 16, respectively. A thorough interpretation of bilateral feature interaction (as well as possible higher orders of interactions) would require a detailed spatial analysis of the area between the transmitter and the receiver, along with specific characteristics of the receiver location. Here, we focus on the interaction of distance and the Tx-Rx angle, which emerges as the most significant interaction in both models.

The mean absolute interaction effect between distance and Tx-Rx angle is approximately 2.5 dB in the EBM model and 0.5 dB in the NGB-SHAP model. In the right-hand side of the figure 13, the blue colored points are indicating on average near to zero contribution of the combination of the corresponding feature values in dB units to the NGB-SHAP model’s PL predictions. Green colored points, located within 0 to 20 meters from the transmitter and between transmitting angles 0o to -100o , exhibit negative contributions (i.e., reduced path loss) of up to -2 dB. The points outside the aforementioned angle are suffering from up to +4dB path loss.

In contrast, the trained EBM model identifies a subset of the receiver points located within 0 to 80 meter and transmission angle between -50o and -100o as benefitting from significant reductions in path loss, ranging from -5 dB to -10 dB. However, increasing the values of the distance within the transmitting angles higher than the -50o can cause high path loss effects ranging from +5dB to +10 dB. Analogous to the NGB-SHAP explanation, the EBM model reveals a distinct spatial zone comprising of points between 0 to 20 meters distance and outside the transmitting angle in-between the 0o to -100o suffering from up to 10dB path losses can be detected.

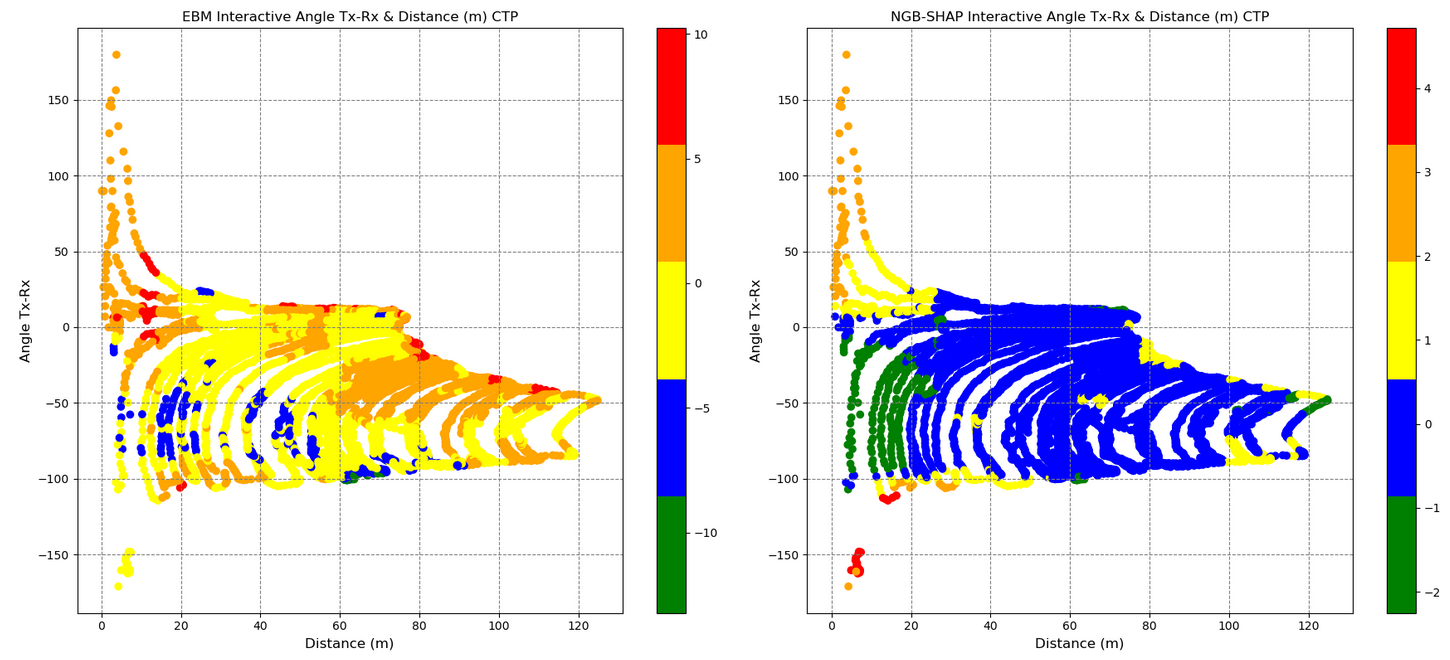


Figure 13

## Model Local Explanations

Global XAI in the previous subsection sheds light on the overall importance of input features and the corresponding marginal contributions in the signal attenuation within our study terrain. However, there are questionable observations with regard to the behavior of path losses in specific locations of the terrain. These include, for example, abrupt changes in PL patterns at some points in the underlying study area or PL behavior deviating from the global XAI view within specific sub-area of the terrain. This necessitates more in-depth look at the reasons behind the ML model predictions at the local level. To support our local XAI analysis based on a consistent view of the terrain, we first apply K-Means clustering based on the spatial locations of data points (comprising longitude, latitude) linked to their corresponding Path Loss values. This serves to figure out intrinsic grouping of data points across the entire vineyard map. The optimal resulted number of clusters through applying the elbow method is 4. The resulting four clusters are visualized on the vineyard map in figure 14, with distinct colors distinct colors used to represent each cluster on the right hand panel. The left hand side panel in figure 14 is plotting the observed path loss values sorted by means of distinct colors. The observed path loss values map and the clustered path loss values map comply with each other. Inspecting the *PL* values corresponding to the points belonging to each of the 4 clusters detected by the K-Means clustering, shows a distinctive partitioning of the data points across the vineyard based on the PL values, in which the green, blue, yellow, and red cluster points Pl values in dB units are lying in the (min, max) ranges equal to (106.65, 125.95), (126.0 , 135.25), (135.3, 143.1), and (143.15, 159.9).

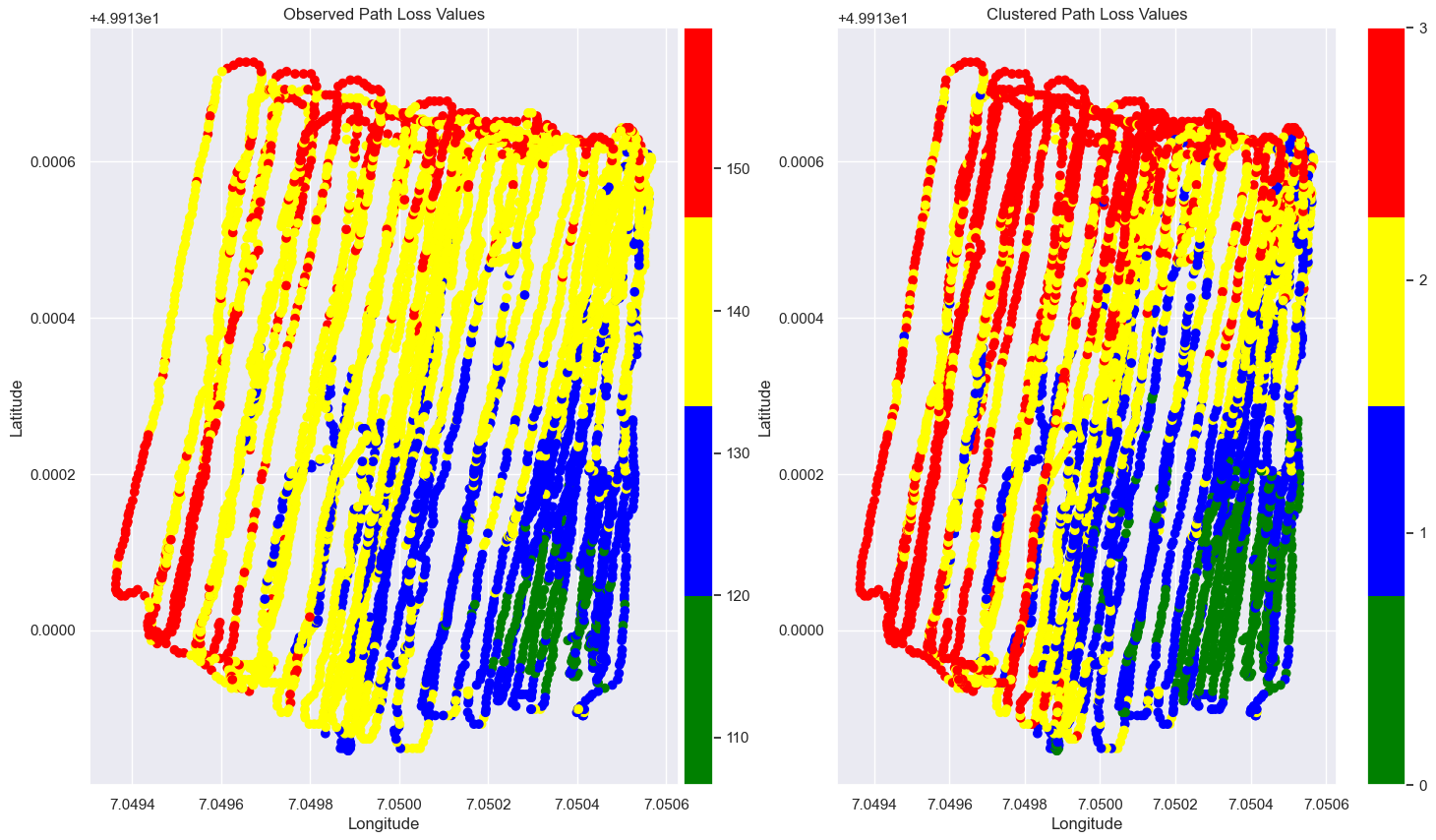


Figure 14

Two local sub areas of the terrain, presented in figure 18 and figure 21, have been identified as regions showing deviating path loss behavior from the global XAI view and demonstrating abrupt changes in the path loss pattern, respectively. The square in the center of the Figure 18 is encircling measured points with latitude and longitude in the (min, max) ranges equal to (49.913046, 49.913096) and (7.049713, 7.049763) , respectively. The points in this sub area are seemingly of similar geospatial profile in terms of distance and transmitting angle relative the base station. Though, the blue points are comprising less signal attenuation values even when the red points are slightly less distanced from the base station.

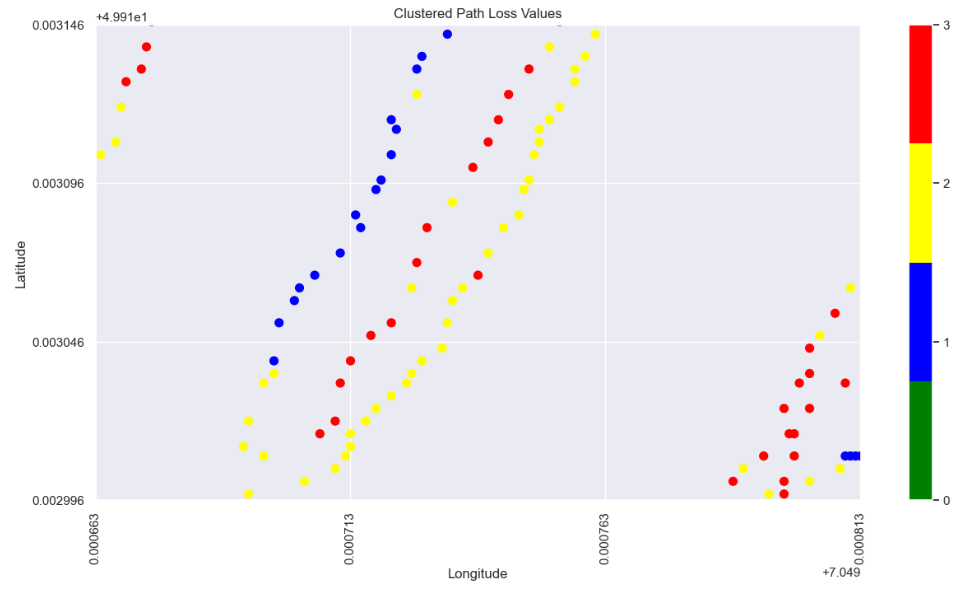


Figure 9

The difference between the red points data and the blue points data can be shown from an overview of the points profile, mean observed values as well as mean predicted values by ML models conveyed through the table 5.

Table 5

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean  Angle Tx-Rx | Mean  Elevation (m) | Mean  Clutter height (m) | Mean Distance (m) | Mean Observed PL | Mean EBM Predicted | Mean NGB Predicted (mean, std) |
| Blue points | 3 | -79.47 | 236.0 | -0.89 | 80.82 | 133.55 | 133.78 | (136.44, 4.96) |
| Red points | 5 | -81.14 | 235.40 | 1.30 | 79.26 | 144.79 | 140.88 | (140.13, 4.20) |

The difference between the interpretation of the blue points in contrast to the interpretation of the red points from the ML models can be seen in figure 19 and figure 20, respectively. Figure 19 presents the the interpretation of the red points, featuring NGB-SHAP explanation on the left-hand side panel and the EBM explanation on the right-hand side panel. Similarly, Figure 20 provides the interpretation of the blue points, again NGB-SHAP explanation on the left-hand side panel and the EBM explanation on the right-hand side panel.

In both figures, presented bars are describing the amount of change in the model output beyond the model intercepts (for NGB-SHAP) and (for EBM ), as defined in equations 12 and equation 16, respectively. Teal coloured bars denote features that contribute to a decrease in the *PL* values. The decreasing effects are sorted based on their importance from left to the right. In contrast, the maroon colored bars denote features that contribute to an increase in predicted PL values. The increasing effects are sorted based on their importance from right to the left.

Comparing the explanation of the red points and the explanation of the blue points via the EBM and NGB-SHAP explanations in figure 19 and figure 20, shows how the decreasing impact of the factor clutter height as the foremost decreasing factor and the third most decreasing factor in the NGB-SHAP model and the EBM model, respectively, changes the balance of the influencing forces to come up with lower path loss values by the blue points. The decreasing effect of the factor clutter height on decreasing the PL values can be checked with reference to the lower right-hand side panels in figure 11 and figure 12.

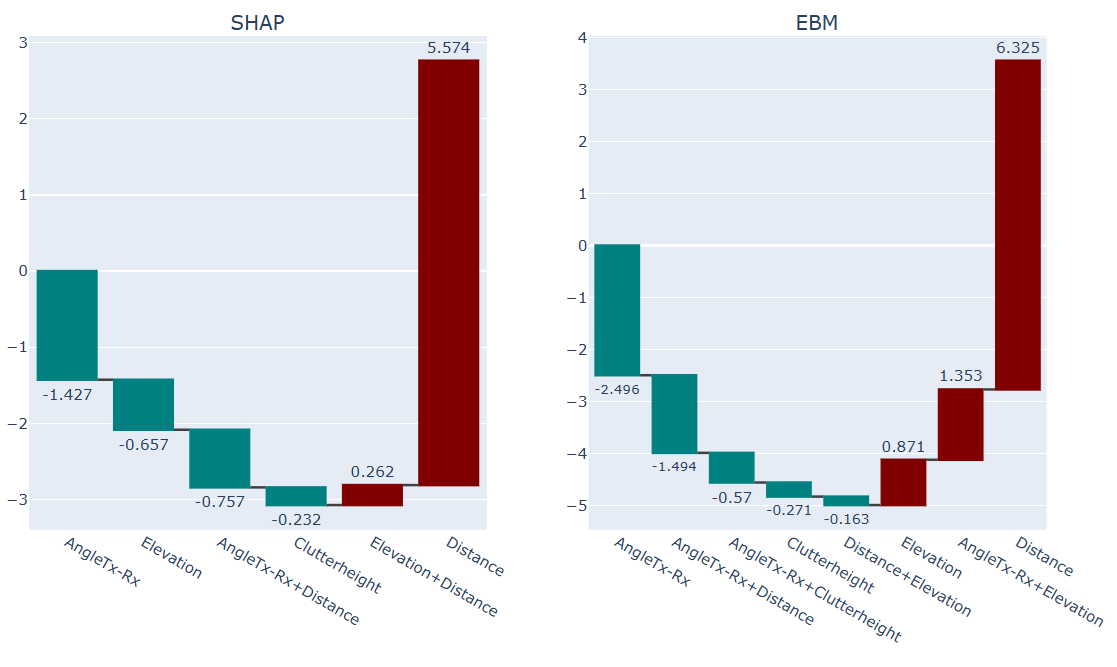
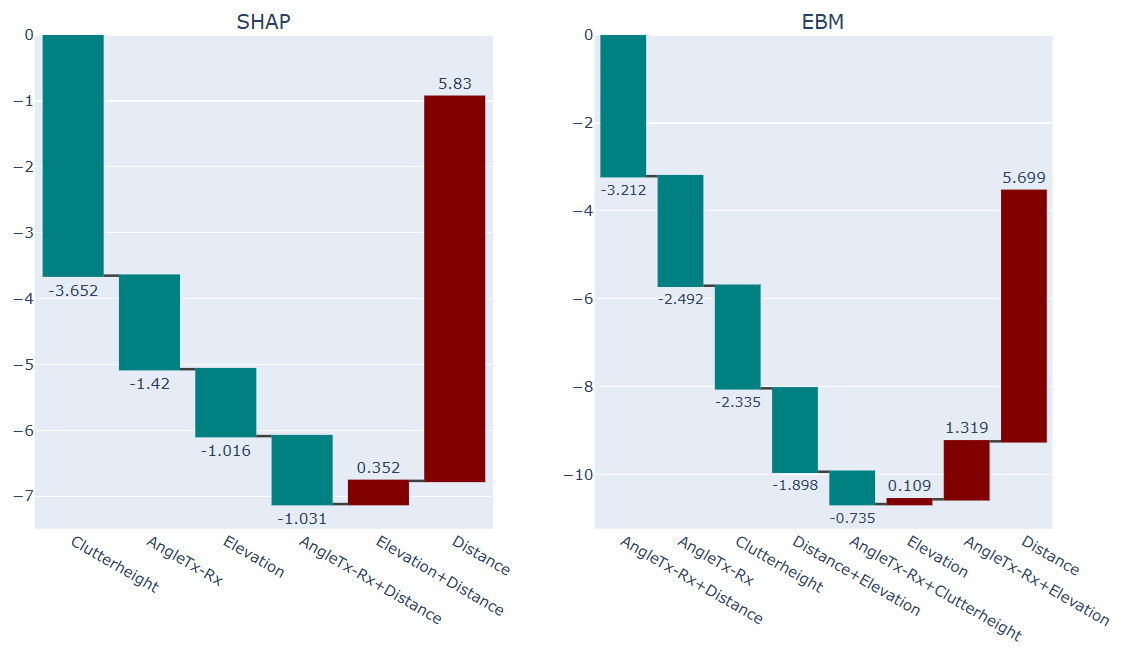


Figure 10



Figure

Figure 21 is comprising the next investigated sub-region in our study in terms of application of local XAI. The square in the center of the Figure 21 is entailing measured points with latitude and longitude in the (min, max) ranges equal to (49.912496, 49.912996) and (7.050413, 7.050463) , respectively. The points in this sub area are seemingly of similar geospatial profile with regard to their proximity to the base station. Though, the blue points are comprising higher signal attenuation values even when they are lying next to the base station (with 3.08 meter distance on average from Tx).



Figure 21

The difference between the green points data and the blue points data can be shown from an overview of the points profile, mean observed values as well as the mean predicted values by ML models conveyed through the table 6.

Table 6

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean  Angle Tx-Rx | Mean  Elevation (m) | Mean  Clutter height (m) | Mean Distance (m) | Mean Observed PL | Mean EBM Predicted | Mean NGB Predicted (mean, std) |
| Green points | 55 | 17.99 | 236.12 | 0.89 | 2.80 | 121.00 | 121.71 | (121.98, 2.82) |
| Blue points | 23 | 56.84 | 235.56 | 0.89 | 3.08 | 128.13 | 127.03 | (126.27, 2.81) |

Comparing the explanation of the blue points and the explanation of the green points via the EBM and NGB-SHAP explanations in figure 22 and figure 23, shows how the increasing impact of the factor transmitting angle as the foremost increasing factor both in the NGB-SHAP model and the EBM model, respectively, changes the balance of the influencing forces to come up with higher path loss values by the blue points. The difference between effect of the transmitting angle being 56.84 degree and 17.99 degree on increasing and decreasing the *PL* values, respectively, can be checked with reference to the upper right-hand side panels in figure 11 and figure 12.

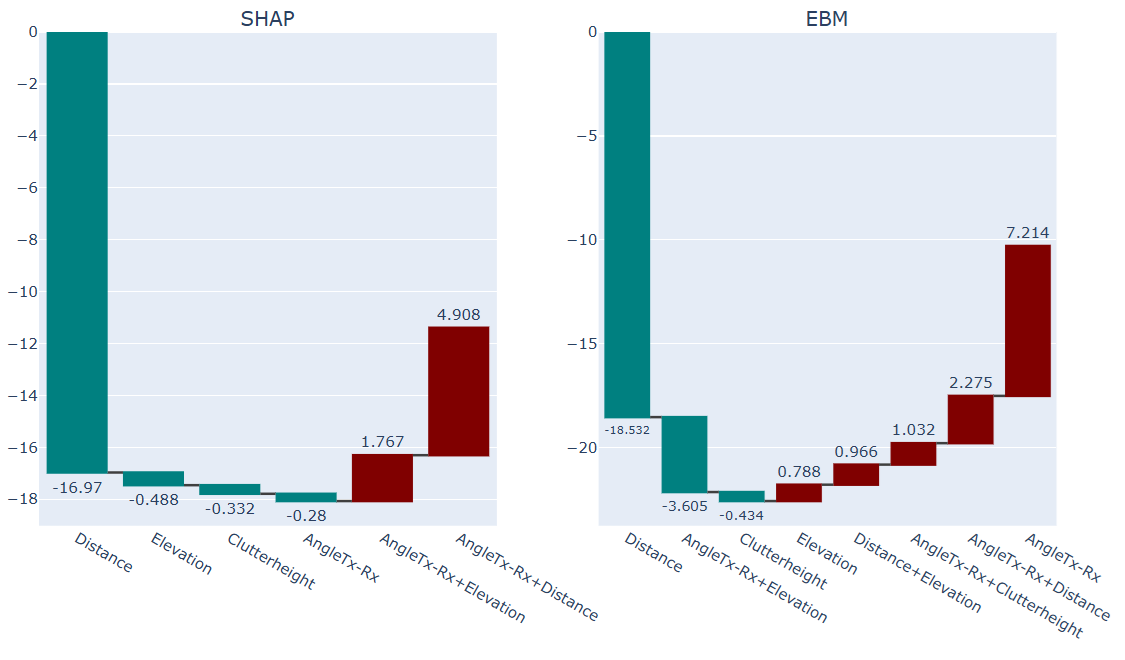


Figure 22



Figure 23

# Discussion

* Interpret how the open field and vineyard setting impacts model behavior and feature influence.
* Trade-offs between accuracy and interpretability.
* Generalizability of models across environments.

# Conclusion

* Summarize findings (e.g. performance, key influencing features)
* Highlight potential applications (e.g., deployment planning in agriculture).
* Future work (e.g. more diverse terrains)

## References

[1] C. Sudhamani, M. Roslee, L. L. Chuan, A. Waseem, A. F. Osman, and M. H. Jusoh, “Performance Analysis of a Millimeter Wave Communication System in Urban Micro, Urban Macro, and Rural Macro Environments,” *Energies (Basel)*, vol. 16, no. 14, Jul. 2023, doi: 10.3390/en16145358.

[2] D. Lindenschmitt *et al.*, “Nomadic Non-Public Networks for 6G: Use Cases and Key Performance Indicators.”

[3] A. Krause, W. Anwar, A. B. Martinez, D. Stachorra, G. Fettweis, and N. Franchi, “Network Planning and Coverage Optimization for Mobile Campus Networks,” in *Proceedings - 2021 IEEE 4th 5G World Forum, 5GWF 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 305–310. doi: 10.1109/5GWF52925.2021.00060.

[4] Y. Egi and E. Eyceyurt, “Classified 3D mapping and deep learning-aided signal power estimation architecture for the deployment of wireless communication systems,” *EURASIP J Wirel Commun Netw*, vol. 2022, no. 1, Dec. 2022, doi: 10.1186/s13638-022-02188-2.

[5] H. Khalili, H. Frey, and M. A. Wimmer, “Balancing Prediction Accuracy and Explanation Power of Path Loss Modeling in a University Campus Environment via Explainable AI,” *Future Internet*, vol. 17, no. 4, p. 155, Mar. 2025, doi: 10.3390/fi17040155.

[6] S. I. Popoola, A. A. Atayero, O. D. Arausi, and V. O. Matthews, “Path loss dataset for modeling radio wave propagation in smart campus environment,” *Data Brief*, vol. 17, pp. 1062–1073, Apr. 2018, doi: 10.1016/j.dib.2018.02.026.

[7] F. Qamar, A. Talib, H. Mhd Nour, D. Kaharudin Bin, B. N. Kamarul Ariffin, and A. Irfan, “Characterization of MIMO Propagation Channel at 15 GHz for the 5G Spectrum,” in *13th Malaysia International Conference on Communications*, Johor Bahru, Johor, Malaysia: Institute of Electrical and Electronics Engineers (IEEE), Nov. 2017, pp. 265–270.

[8] C. Kaitatzis, A. Boursianis, S. K. , Goudos, and D. Panagiotis I., “A Preliminary Coverage Study in Millimeter Wave Bands for 5G Communication Networks,” in *MOCAST : 6th International Conference on Modern Circuits and Systems Technologies*, Thessaloniki, Greece: IEEE, May 2017, p. 326.

[9] M. A. M. Nordin and H. A. M. Ramli, “Performance analysis of 5G path loss models for rural macrocell environment,” *IIUM Engineering Journal*, vol. 21, no. 1, pp. 85–99, 2020, doi: 10.31436/iiumej.v21i1.1247.

[10] A. Schumacher, R. Merz, and A. Burg, “3.5 GHz Coverage Assessment with a 5G Testbed,” in *89th Vehicular Technology Conference (VTC2019-Spring)*, IEEE, 2019.

[11] C. J. Furnieles, J. A. Castro, D. S. Lopez, J. E. Arevalo, and J. L. Araque, “Path Loss Measurements in the 60 GHz Frequency Band in a Greenhouse,” in *LACAP 2024 - 1st Latin American Conference on Antennas and Propagation, Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/LACAP63752.2024.10876224.

[12] N. Saba, L. Mela, K. Ruttik, J. Salo, and R. Jantti, “Millimeter-Wave Radio Link Analysis for 5G FWA by Combining Measurements and Geospatial Data,” in *Proceedings - 2022 IEEE Future Networks World Forum, FNWF 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 433–438. doi: 10.1109/FNWF55208.2022.00083.

[13] T. Hayashi and K. Ichige, “A Deep-Learning Method for Path Loss Prediction Using Geospatial Information and Path Profiles,” *IEEE Trans Antennas Propag*, vol. 71, no. 9, pp. 7523–7537, Sep. 2023, doi: 10.1109/TAP.2023.3295890.

[14] K. Kayaalp, S. Metlek, A. Genc, H. Dogan, and İ. B. Basyigit, “Prediction of path loss in coastal and vegetative environments with deep learning at 5G sub-6 GHz,” *Wireless Networks*, vol. 29, no. 6, pp. 2471–2480, Aug. 2023, doi: 10.1007/s11276-023-03285-w.

[15] A. Gupta, K. Ghanshala, and R. C. Joshi, “Path loss predictions for fringe areas using adaptive neuro-fuzzy inference system,” *International Journal of System Assurance Engineering and Management*, vol. 13, pp. 866–879, Jun. 2022, doi: 10.1007/s13198-021-01196-7.

[16] F. E. Shaibu, E. N. Onwuka, N. Salawu, S. S. Oyewobi, K. Djouani, and A. M. Abu-Mahfouz, “Performance of Path Loss Models over Mid-Band and High-Band Channels for 5G Communication Networks: A Review,” Nov. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/fi15110362.

[17] H. Khalili, H. Frey, and M. A. Wimmer, “Balancing Prediction Accuracy and Explanation Power of Path Loss Modeling in a University Campus Environment via Explainable AI,” *Future Internet*, vol. 17, no. 4, p. 155, Mar. 2025, doi: 10.3390/fi17040155.

[18] Y. Nuñez, L. Lovisolo, L. da Silva Mello, and C. Orihuela, “On the interpretability of machine learning regression for path-loss prediction of millimeter-wave links[Formula presented],” *Expert Syst Appl*, vol. 215, Apr. 2023, doi: 10.1016/j.eswa.2022.119324.

[19] Y. Sun *et al.*, “Environment Features-Based Model for Path Loss Prediction,” *IEEE Wireless Communications Letters*, vol. 11, no. 9, pp. 2010–2014, Sep. 2022, doi: 10.1109/LWC.2022.3192516.

[20] X. Liao, P. Zhou, Y. Wang, and J. Zhang, “Path Loss Modeling and Environment Features Powered Prediction for Sub-THz Communication,” *IEEE Open Journal of Antennas and Propagation*, 2024, doi: 10.1109/OJAP.2024.3454120.

[21] I. Yazici, E. Özkan, and E. Gures, “Enhancing Path Loss Prediction Through Explainable Machine Learning Approach,” Institute of Electrical and Electronics Engineers (IEEE), Sep. 2024, pp. 1–5. doi: 10.1109/wincom62286.2024.10655363.

[22] A. Gelman, “Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper),” 2006. [Online]. Available: http://www.stat.columbia.edu/~gelman/

[23] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, “Probabilistic programming in Python using PyMC3,” *PeerJ Comput Sci*, vol. 2016, no. 4, 2016, doi: 10.7717/peerj-cs.55.

[24] T. Duan *et al.*, “NGBoost: Natural Gradient Boosting for Probabilistic Prediction,” 2020.

[25] S. Amari, R. Karakida, and M. Oizumi, “Fisher Information and Natural Gradient Learning of Random Deep Networks,” Aug. 2018, [Online]. Available: http://arxiv.org/abs/1808.07172

[26] Messalas Andreas, Kanellopoulos Yiannis, and Makris Christos, “Model-Agnostic Interpretability with Shapley Values,” in *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, IEEE, 2019, pp. 1–7. doi: doi:10.1109/IISA.2019.8900669.

[27] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[28] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Adv Neural Inf Process Syst*, 2017, [Online]. Available: https://github.com/slundberg/shap

[29] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowl Inf Syst*, vol. 41, no. 3, pp. 647–665, Nov. 2014, doi: 10.1007/s10115-013-0679-x.

[30] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[31] A. E. Maxwell, M. Sharma, and K. A. Donaldson, “Explainable boosting machines for slope failure spatial predictive modeling,” *Remote Sens (Basel)*, vol. 13, no. 24, Dec. 2021, doi: 10.3390/rs13244991.

[32] H. Nori, S. Jenkins, P. Koch, and R. Caruana, “InterpretML: A Unified Framework for Machine Learning Interpretability,” *arXiv preprint arXiv:1909.09223*, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.09223

[33] C. Zhang and M. Yunqian, *Ensemble Machine Learning: Methods and Applications*. Springer publishing company, 2012.

[34] X. Yang, W. Zhao, Y. Xu, C. D. Wang, B. Li, and F. Nie, “Sparse K-means clustering algorithm with anchor graph regularization,” *Inf Sci (N Y)*, vol. 667, May 2024, doi: 10.1016/j.ins.2024.120504.

[35] D. Marutho, S. Hendra Handaka, and E. Wijaya, “The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News,” in *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, 2018, pp. 533–538.

1. (https://github.com/cokelaer/fitter, accessed on 18 June 2025) [↑](#footnote-ref-2)
2. GitHub - stanfordmlgroup/ngboost: Natural Gradient Boosting for Probabilistic Prediction [↑](#footnote-ref-3)
3. [Welcome to the SHAP documentation — SHAP latest documentation](https://shap.readthedocs.io/en/latest/index.html) [↑](#footnote-ref-4)