Progress In Electromagnetics Research C, Vol. 119, 65–79, 2022

Large Intelligent Surface-Assisted Wireless Communication and Path Loss Prediction Model Based on Electromagnetics and Machine Learning Algorithms

Wael Elshennawy*

Abstract—This paper presents the application of machine learning-based approach toward prediction of path loss for the large intelligent surface-assisted wireless communication in smart radio environment. Two bagging ensemble methods, namely K-nearest neighbor and random forest, are exploited to build the path loss prediction models by using the training dataset. To generate the data samples without having to run measurement campaign, a path loss model is developed owning to the similarity between the large intelligent surface-assisted wireless communication and the reflector antenna system. Simple path loss expression is deduced from the system gain of the reflector antenna system, and it is used to generate the data samples. Simulation results are presented to verify the prediction accuracy of the path loss predictions models. The prediction performances of the trained path loss models are assessed based on the complexity and accuracy metrics, including R^2 score, mean absolute error, and root mean square error. It is demonstrated that the machine learning-based models can provide high prediction accuracy and acceptable complexity. The K-nearest neighbor algorithm outperforms random forest algorithm, and it has smaller prediction errors.

1. INTRODUCTION

Large intelligent surface (LIS) system lately has been proposed as a promising solution with low cost and energy efficient. LIS can support a diverse variety of applications, including ultra-reliability low latency communication (uRLLC), massive machine-type communications (mMTC) [1], etc. The architecture of LIS is a spatially continuous surface, which comprises a collection of closely spaced tiny antenna elements, e.g., programmable meta-surface deployed into a limited aperture [2]. LIS exhibits great ability to manipulate electromagnetic waves with a limited power consumption. For instance, LIS can be controlled via external signals such as backhaul control link fed from the base station (BS) to manipulate in real-time the reflected wave phase and magnitude [3]. This property allows to use LIS in wireless communication systems as a reflecting surface between the BS and user equipment (UE). This application of the LIS as a reflecting surface is known as LIS-assisted wireless communication [4]. LIS can be densely deployed around devices and terminals, which makes the propagation channel more line-of-sight (LoS) favorable.

LIS-assisted wireless communications system can provide many advantages including the capability to provide a reliable and space-intensive communication by effectively establishing LoS channels between LIS and users [4]. Though LIS-assisted wireless communication system can afford low-cost and simplistic architecture, it brings a difficulty of including two wireless communication channels between the BS and a user. Besides, the channel estimation overhead scales with the number of antenna elements [5, 6]. Since LIS is typically envisioned for including many antenna elements, this overhead burden could potentially

Received 30 January 2022, Accepted 3 March 2022, Scheduled 14 March 2022
* Corresponding author: Wael S. Elshennawy (wael.elshennawy@orange.com).
The author is with the Orange Business Services Co., Cairo, Egypt.

be a challenge in LIS implementation. To sum up, channel estimation overhead can pose a significant bottleneck for the deployment of LIS into a smart radio environment.

The existing works for path loss model estimation are mainly based on either analytical models [3, 7, 8] or empirical models [9] which rely on data collected in specific propagation scenarios. Although the empirical models are computationally efficient and easy to implement, the actual path loss at a specific location cannot be accurately computed especially in more general environments [10]. Due to the high cost of carrying out the measurement campaign and the need of developing evaluation metrics and tools for better judging the collected data [11], another candidate solution is to utilize deterministic methods [12] that are based on ray tracing or finite-difference time-domain (FDTD) methods. These deterministic methods are extremely accurate for predicting the spatial distribution of electromagnetic fields, where it is based on radio wave propagation mechanism and numerical analysis for computing the path loss values. The main disadvantage of these deterministic methods is that the computation procedure consumes appreciably more time and memory resources, and thus it is inappropriate to use these complex methods in real-time applications. For instance, ray-tracing method incurs sever computational burden when it is extended to model ultra-wideband (UWB) channel of several Giga-Hertz (GHz). Additionally, the exhaustive computation has to be run again once the propagation environment changes. So, there is a demand on having a trade-off between the estimation accuracy and complexity [2]. Relying on the free-space path loss model for LIS-assisted wireless communication is of no help [1]. The development of a path loss model for LIS-assisted wireless communication channel is of utmost importance. It is necessary, then to develop a physics-based model, to account for propagation environment characteristics and interference at the receiver [2] whilst maintaining low computational complexity.

Recently, machine learning (ML) algorithms have attracted huge interests for addressing the path loss prediction for LIS-assisted wireless communication in [11, 13]. ML techniques are assisted by training models with different channel characteristics as in [14, 15]. It can then instantly adapt to the changes in the propagation environment. In addition, updating the channel information can be done less frequently, which guarantees robust performance. It also aims at reducing the model complexity and increasing the prediction model accuracy. Path loss modeling based on ML methods can be regarded as a data mining task, and ML algorithms can be considered as a valid solution to predict the path loss. Path loss prediction model is classified as a supervised regression problem [11]. The prediction models based on ML methods are able to provide more accurate path loss prediction results than the empirical models [10], and they are more computationally efficient than the deterministic methods [15]. Though ML algorithms overcome challenging issues, namely complexity and time consuming, there is an inevitable trade-off between getting reliable predictions and the vital requirement of tremendous measurements to train the model.

In this paper, prediction models based on ML methods are built for LIS-assisted wireless communication. The dataset is acquired by developing a path loss model rather than relying on collecting data samples by measurement campaign. The resemblance between the LIS-to-UE and the reflector antennas allows to develop a path loss model based on the system gain of a reflector antenna system. For validation purpose, the developed path loss model is corroborated with the work in [16] for the far-field approximation. This simple path loss expression is applied to generate the required data samples, which fulfills the following algorithm functions, namely train, validate, and test the models. Two bagging ensemble algorithms, random forest and K-nearest neighbor (KNN), are implemented to build prediction models. The accuracy of ML-based models is assessed by using statistical error indicators, and the complexity is also evaluated. The importance of input features on the path loss prediction model based on random forest algorithm is analyzed and discussed. The remainder of this paper is organized as follows. The system model is presented in Section 2 followed by the channel model in Section 3. The presented path loss model and path loss model validation are described in Section 4. Section 5 presents the ML-based methods for path loss prediction details, including procedure, data wrangling, features selection, data division, model selection, model training and tuning, model evaluation, and complexity. In Section 6, the performance of path loss prediction models based on ML methods is evaluated by using Matlab and Anaconda software programs. At last, conclusions are drawn in Section 7.

2. SYSTEM MODEL

Consider a circular LIS located on the xy-plane with a diameter d, and K single users are in a three-dimensional space as shown in Fig. 1. The geometric center of the LIS is located at the origin $(x, y, z) = 0$ of the Cartesian coordinate system, while the locations of the users are only limited to the space where $z > 0$. LoS-dominated propagation channel is assumed for far-field scenario, and it can be found in many LIS applications, namely indoor or outdoor open spaces [13, 15], etc. The Euclidean distance between the kth user located at (x_k, y_k, z_k) and the LIS center is denoted as the effective distance d_k^c. The free-space path loss is expressed as a function of the distance between the LIS and the UE, and is given in [17] as

$$PL_k = \frac{1}{(2\kappa d_k^c)^2}, \tag{1}$$

where $\kappa = 2\pi/\lambda$ and λ is the wavelength. It is noteworthy that PL_k is valid wherever d_k^c is larger than Fraunhofer distance, i.e., $d_c^k > 2d^2/\lambda$.

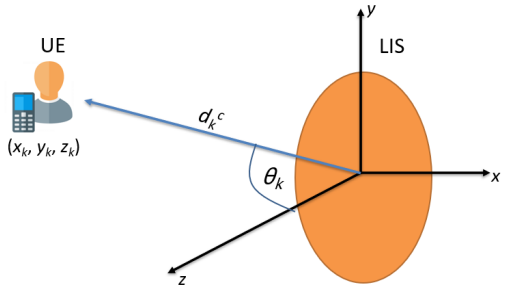


Figure 1. The circular LIS and UE in far-field scenario.

3. CHANNEL MODEL

For a far-field scenario, the channel propagation from the kth user to any point $(x, y, 0)$ located onto the LIS aperture is provided by [18] as

$$g_k(x, y) = \sqrt{PL_k} \cdot h_k(x, y), \tag{2}$$

where

$$\begin{aligned} h_k(x, y) &= e^{-j(\kappa d_k + \varphi_k)}, \\ d_k &= \sqrt{z_k^2 + (x - x_k^2) + (y - y_k^2)}, \end{aligned} \tag{3}$$

φ_k is a random phase following uniform distribution in the range of $[-\pi, \pi]$, and d_k is the Euclidean distance between the kth user and any point $(x, y, 0)$ located onto the LIS. The channel model analysis has to cover the path loss and other channel characteristics accordingly. The focus here is on the path loss estimation in the far-field region. The shape of the LIS can still be regarded as a continuous surface of fixed area for the far-field region [2], which will be further discussed for the evaluation of the path loss model and draws insights on the free-space path loss of the LIS-assisted wireless communications in the next section.

4. LIS-TO-UE PATH LOSS MODEL EVALUATION

The commonly used path loss in literature [4, 19] assums that the LIS and UE antennas are isotropic whilst deriving the path loss from the Friis transmission equation [20]. On the contrary, only a portion

of the power radiation pattern emitted by the LIS antenna is intercepted by the UE antenna, which represents the useful power in the LIS-assisted wireless communication. Therefore, it is desirable to separate various parameters associated with the propagation channel (including the LIS and UE), collectively referred to as the path loss from the antennas gain. To accomplish this, free-space path loss model has to be modified to account for these parameters i.e., the antenna efficiencies, Fresnel's correction factor, etc.

Owing to the likeness between the integrated reflector antennas and LIS-assisted wireless communication system, the LIS can be modeled as a uniform phase circular aperture and the UE as a square aperture with a linear phase error. This approximation of UE as a square aperture having a linear-phase error is plausible, because the LIS antenna radiation pattern is highly directive, and the attention is always with the radiation at angles confined within the projected area of the UE antenna normal to the incident radiation [21]. Also, the diffraction in the direction of the UE antenna is merely negligible for electrically large structures like LIS. The general integral equation for the radiation intensity in the far-field region for a circular aperture has already been presented by Silver in [22]. Though Silver derives a more intricate diffractivity expression pertaining to the general circular aperture problems in the far-field region, the qualitative aspects differ for the case of LIS-assisted wireless communication than the general case. Notably, there are many differences to consider: the nonuniform illumination of the circular aperture of the LIS and the UE aperture, to name a few.

The aim is to present a tractable and reliable physical and electromagnetic-based model for the LIS-to-UE, that does not include beamforming gain. In [21], the authors discuss the system gain and the radiation efficiency of a reflector antenna system, and a rigorous system gain formula is developed from the aperture field theory, including the radiation efficiency. This yields a gain expression for the reflector antenna system, that comprises a reflector antenna and a receiving antenna located in the far-field region. The system gain expression entails the nonuniform illumination of apertures, effective antenna aperture, near-field effects, etc. In the context of path loss evaluation, the interest is to propose a simple expression for the LIS-to-UE propagation channel, relying on the similarity with the reflector antenna system. So, the path loss model can adopt the following,

$$\begin{aligned}
PL''_k &= \eta_1\eta_3\eta_4, \\
\eta_1 &= a_0\frac{\pi a_k^2 d^2}{4\lambda^2 {d_k^c}^2}[1 - a_1\delta_1 + a_2\delta_1^2]\cos(\theta)^2, \\
\eta_3 &= [(\eta_0)(1 - \lambda/2a_k^2)], \\
\eta_4 &= \left[1 - a_3\left(\frac{\pi d^2}{4\lambda d_k^c}\right)^2\right]\left[\frac{C^2(\varpi) + S^2(\varpi)}{\varpi^2}\right]^2, \\
C(\varpi) &= \int_0^{\pi}\cos(\pi/2t^2)\,dt, \\
S(\varpi) &= \int_0^{\pi}\cos(\pi/2t^2)\,dt, \\
\varpi &= \frac{a_k}{\sqrt{2d_k^c\lambda}}\,.
\end{aligned} \tag{4}$$

The first term in the path loss PL''_k is the spillover efficiency between LIS and UE, designated as η_1. a_0 is the loss factor associated with LIS that causes a reduction in the peak radiation intensity. a_1 and a_2 are constant terms dependent on the power pattern intercepted by the UE, and δ_1 is the relative intensity at the edge of UE aperture in the principal plane. a_k^2 is the aperture area of the UE. The second term η_3 is the reflection efficiency factor. It is defined here as a multiplication of two terms. η_0 is the heat loss caused by currents on the antenna aperture of UE, and $(1 - \lambda/2a_k)^2$ is the edge efficiency loss due to the discontinuity in the field near the edge of the UE antenna aperture. Other efficiency factors are assumed to be negligible. The third term η_4 is Fresnel correction factor. a_3 is a constant term dependent on the illumination taper distribution, and $C(\varpi)$ and $C(\varpi)$ are cosine and sine Fresnel integrals. Clearly, the evaluation of path loss for the LIS-to-UE benefits from the similarity found with the reflector antenna system, where it allows to integrate different factors, e.g., radiation efficiency and

Fresnel correction. In subsequent section PL_k'' will be computed for LIS-to-UE in far-field scenario and compared with the plate-scattering based-model discussed in [16].

4.1. Path Loss Model Validation

Since the path loss model of LIS-assisted wireless communication follows the plate scattering paradigm [16] in the far-field region, this serves as a useful benchmark for the path loss in channels employing LIS. It is necessary to consider that this connection between the plate scattering and the LIS scattering is not universal. The equivalence demonstrated here for validation purpose is attributed to the assumption that LIS is a perfectly conducting flat plate and acts as a monostatic geometry. Therefore, the path loss of the LIS-assisted wireless communication for LIS-to-UE propagation channel can be modeled based on the far-field beamforming case [8], considering that the LIS is partitioned into continuous tiles separated by a distance $\lambda/2$ each as discussed in [2]. This yields the following

$$PL_r \leq \frac{A_c \eta_t}{(2\pi d_k^c)^2}, \tag{5}$$

where η_t is the total radiation efficiency. The total radiation efficiency term is assumed to be equal to the difference between PL_k'' and $\frac{A_c}{(2\pi d_k^c)^2}$. In the evaluation of PL_r, the upper bound is considered for comparison with other path loss models, because the work in [16] assumes a planar square LIS with uniform antenna elements patterns and $\lambda/2$ spacing between antenna elements. It is now intriguing to visualize the three path loss models PL_k, PL_k'', and PL_r for demonstration purpose. Consider that the LIS has a diameter; $d = 1.22$ (m), and UE has a side length; $a_k = 0.05$ (m). The UE can move to a distance up to $d_k^c < 40$ (m). The operating frequency is $f = 2.6\,\text{GHz}$. The values of parameters, including a_0, a_1, a_2, a_3, and η_0, are used for a typical reflector antenna discussed in [21] and are given below in Table 1. The curves for the three path loss models PL_k, PL_k'', and PL_r are plotted by using a Matlab program [23] and are shown in Fig. 2. The curves show that the PL_k'' and PL_r are equivalent in the far-field range, and consequently, the results of Equations (4) and (5) are same for the far-field approximation. PL_k'' results are compared to PL_k having the same path length, and it is shown that the path loss of the LIS-to-UE is lower than the case of free-space path loss. Referring to the assumption that the LIS is partitioned into continuous tiles, thus LIS can be viewed as an array of diffuse scatters [24] that phase-align their signals at the UE. Hence, PL_r is merely a multitude of PL_k associated with diffuse scatters as shown in Fig. 2. It should be emphasized that PL_k'' can also show the Fresnel region, where the minimum path loss value occurs at near-radiating distance. Moreover, the area of the LIS antenna aperture plays an important role in controlling the path loss for the LIS-assisted wireless communication. Similarly, this conclusion has been drawn for the power gain between LIS and small/medium intelligent surfaces (SIS/MIS) in [3]. This discussion without resorting to electromagnetic

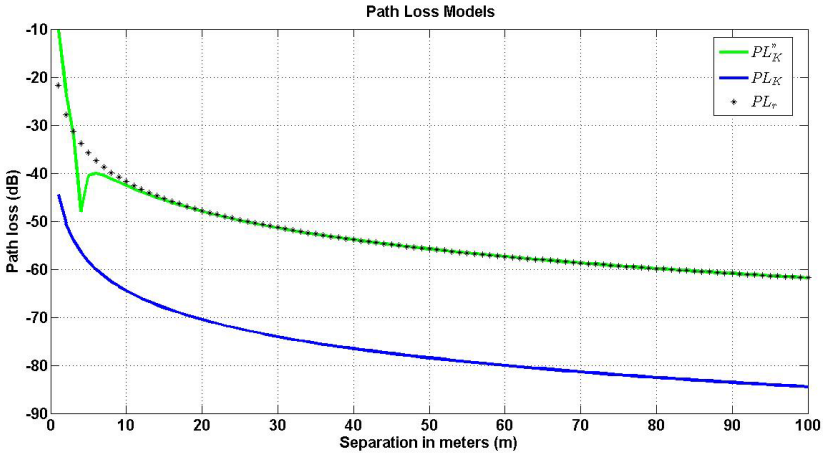


Figure 2. Path loss models.

Table 1. Path loss model PL_k'' parameters.

Parameter	Value
a_0	0.65
a_1	1.3
a_2	0.62
a_3	0.0684
η_0	0.99

extensive simulations yields a useful guideline for identifying the required LIS physical area to achieve a path loss less than the path length free space channel. In contrast, the physical area is of great concern due to the potentially high cost of the LIS, and the cost is expected to be proportional to physical area and number of antenna elements [19].

5. MODELING PATH LOSS USING MACHINE LEARNING METHODS

Machine learning-based model is data driven modelling. Therefore, the number of data samples is very useful to the improvement of prediction accuracy. Ideally, the data samples are collected from the measurements to get accurate knowledge of the channel characteristics. In some cases, the measurement campaign is even not possible to carry out. Also, the number of measured data sometimes cannot meet the requirements of ML algorithm [11]. Alternatively, a path loss prediction scheme relies on combining the ML-based model and a classical path loss model to generate the required data samples likewise expansion method discussed in [10]. Generally, the goal of the ML method is to find the optimal function, which best describes the relationship between the input features and the output path loss value. In the next subsections, the procedure of ML-based method for path loss prediction will be explored in more details.

5.1. Procedure of Machine Learning-based Methods for Path Loss Prediction Model

The procedure for ML-based method for path loss prediction is described by a block diagram in Fig. 3, which is reproduced from [11] and adapted to the context. The data samples are gathered for a far-field

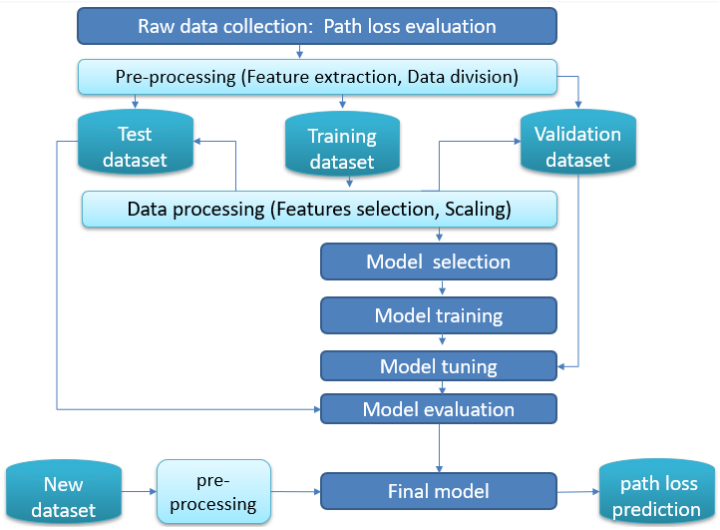


Figure 3. Procedure of machine learning-based method for path loss prediction, reproduced from [11].

scenario based on the developed path loss model discussed earlier. Each data sample is composed of path loss value and input features, e.g., d_k^c. Next, the data samples are segregated into three datasets, namely training dataset, validation dataset, and test dataset. This is followed by feature selection and feature normalization. Based on the training dataset and selected ML algorithm, the ML-based model is trained, and its hyperparameters are subsequently optimized by using the validation dataset. After the optimal path loss prediction model has been built, the path loss can be predicted by using the test dataset. The statistical error indicators are applied to assess the prediction accuracy of the trained model. The subsections broadly discuss the various workflow steps of the ML-based method for path loss prediction.

5.2. Data Wrangling

The data samples are computed by randomly displacing the UEs at different locations and calculating the corresponding path loss values. The input features for the free-space case are limited to the system-dependent parameters [11]. The data wrangling implies different functions, e.g., inspect and diagnose data samples, and dealing with outliers [25]. Then, data samples are shuffled by using a pseudorandom number generator to assert that the datasets contain data samples from all classes. This guarantees the same prediction results if one runs the ML algorithm several times.

5.3. Features Selection and Extraction

Feature selection aims to select the optimal subset with the least number of input features, which mostly contribute to learning accuracy [15]. The independent variables accompanied with the propagation channel, including LIS and UE, are collectively referred to as model input features. Retaining the relevant features and discarding irrelevant features directly lead to having good quality of the predictor. The essence of a feature selection method is to make a trade-off between adding the most significant features and minimizing the model complexity; otherwise, it could increase the occurrence of overfitting problem [25]. Although the ML algorithm essentially avoids creating a set of expert-designed rules, it does not mean that prior knowledge of the electromagnetics should be discarded. On contrary, domain expertise with electromagnetic problems can help in identifying the useful model input features, which are much more informative than the initial representation of the data. Five input features have been selected by using correlation method discussed in [26]. These input features of the prediction model are listed below, and their data types and range values are tabulated in Table 2.

Table 2. Features and range values.

Feature	Data Type	Range Values
d_k^c	Float	$2.5 < d_k^c \leq 40$ (m)
$h_{y,k}$	Float	$0 < h_{y,k} \leq 70$ (m)
a_k	Float	[0.01, 0.02, 0.03, 0.04, 0.05] (m)
θ_k	Float	$0 < \theta_k < \pi/2$
$\tilde{\chi}$	Boolean	[0, 1]

1) Propagation distance (d_k^c, in meter). Regression with KNN and random forest algorithms are known to have discrete prediction values, which introduce discontinuities problem. In addition to the presented far-field model consider the Fresnel correction factor into the path loss evaluation as depicted in Fig. 2. To have smoother predicted values, the chosen range for the propagation distance has been chosen to extend beyond the near-radiating region.
2) UE height ($h_{y,k}$, in meter): the height of the UE measured from the LIS center in y-coordinate.
3) UE aperture area (a_k^2, in meter^2).
4) Elevation angle (θ_k): the angle between the LoS path and xy plane.

5) $\tilde{\chi}$: The one-hot encoding [25] is used to replace spatial resolution values with one or more new features, which can have the values "0" and "1". So, any number of categories can be represented by a new feature per category. For instance, the class label "far-apart" corresponds to the case where UEs are far from each other by a distance $\geq \tilde{\chi}$, where it is encoded by "True". On the other hand, the class of closely spaced UEs is encoded by "False" for a distance $< \tilde{\chi}$.

The one-hot encoding consequently causes an imbalanced dataset, which necessitates a restoration of an even balance between majority class and minor class. The under-sampling technique [27] is then used to balance the dataset by reducing the size of the major class (True) in proportion to the minor class (False). This method is usually used when the number of data samples is sufficiently large. The method keeps all data samples in the minor class, and it randomly selects an equal number or percentage of data samples out of the major class. A balanced new dataset can be retrieved for modelling. The resulting number of data samples after preprocessing is 1,183,962 in total. The input features are further scaled by the normalization process [25]. Feature normalization is very imperative to ensure that all input features are set to the same scale. This guarantees faster convergence on model learning and satisfies uniform influence on the weights of input features [28].

5.4. Data Division

Data samples are partitioned into three datasets: training dataset, validation dataset, and test dataset. The training dataset is used to train the model, the validation dataset examines the performance of the potential model (i.e., model with different hyperparameters), while the test dataset is used to evaluate the final selected model. Deciding how much dataset is partitioned as the training dataset, validation dataset, and test dataset is somewhat arbitrary. A test dataset containing almost 20% of the dataset is a good rule of thumb [15]. The performance of the path loss prediction model is strongly dependent on the number and quality of training dataset. More training dataset leads to having more accurate reflections on the inherent laws and interrelationships among the labels and input features [11]. In this paper, the proportions of the training dataset, validation dataset, and test dataset of the whole dataset are divided into 65%, 15%, and 20%, respectively.

5.5. Model Selection

Model selection in ML is defined as the process of choosing one model among many candidate models. Notably, there may be many competing concerns whilst performing model selection goes beyond model performance, including complexity, resources availability, etc. Additionally, the purpose of ML method is to improve the performance on a specific task based on extensive data samples and a flexible model architecture [25]. The path loss prediction problem is classified as a typical supervised learning regression task [10], because the predicted path loss values are continuous and have labels. Therefore, it can be solved by using different regression algorithms, like neural network regression [29], random forest [30], support vector machine (SVM) [12], etc. These ML algorithms establish the mapping relationship between the inputs and outputs based on the input features and the corresponding path loss values by using the training dataset. Furthermore, the path loss values under new conditions are predicted by the trained model. KNN and random forest algorithms are selected from ensemble methods to build the path loss prediction models. It has been reported that these algorithms have good performance in predicting path loss values in [10, 15]. In addition, other many advantages will be revealed shortly in more details. The major principles of these algorithms are introduced as follows.

1) KKN ensemble learning algorithm belongs to the class of bagging methods [31]. This algorithm for KNN regression builds several instances of a KNN estimator on random subsets of the original training dataset and then aggregates their individual predictions to form a final prediction. It serves as a mean to reduce the variance of a single KNN estimator by introducing randomization into its construction procedure [31] and then makes an ensemble out of it. The mechanism of KNN algorithm itself is to find the k_{nn} samples closest to the query sample to be predicted based on a distance metric [32]. Then it performs the prediction based on averaging the information of these k_{nn} nearest neighbors. Here, the used distance metric is the Euclidean distance. The

KNN algorithm is characterized by no explicit training process, and its implementation is very straightforward [28].

2) Random forest predictor is also a bagging method. Multiple individual ensembles are used to solve this regression problem, where it achieves an improvement in controlling overfitting [28]. It employs a decision tree algorithm as an ensemble member. Then it implements the bootstrap aggregating ensembles [30] in order to select the size of decision tree member from randomly-sampled out of the training dataset. Decision tree members are trained based on these samples, and then the result is obtained by averaging the results of all the decision tree members [28]. In addition, the input features are always randomly permuted at each split of the decision tree. Therefore, the best-found split may vary, even with the same training dataset. Thus, the greater the diversity of decision tree members is, the better the prediction performance and predictive accuracy improvement are [30]. By introducing both sample perturbations and feature perturbations, the diversity of decision tree members in random forest is increased and directly leads to fully grown and unpruned trees. Random forest algorithm is insensitive to input data samples and simple to implement, and it can handle thousands of input features [11].

5.6. Model Training and Tuning

The process of training a prediction path loss model based on ML methods involves providing the KNN and random forest algorithms with training dataset to learn from. The path loss prediction model is learned, but the question arises if the prediction model is generalizing well for making path loss predictions on new dataset. So, it is necessary to maximize the performance of trained model without having an overfitting problem or causing too high a variance. The best way to think about model tuning and hyperparameters is just like the settings of an algorithm, which can be finely adjusted to optimize the performance. Concisely, model tuning aims at acquiring the optimal hyperparameters values [11]. Model hyperparameters are those parameters defining the model architecture. Random forest and KNN algorithms are characterized by employing a few hyperparameters compared to other ML algorithms in model tuning [28].

Determination of the optimal hyperparameters values enables achieving high performance of the path loss prediction model. For random forest algorithm, the model accuracy is affected by these hyperparameters, including maximum tree depth (d_{tree}), and the number of ensemble members [30]. Generally, a small ensemble with deep decision trees has a greater tendency toward overfitting than a shallow ensemble of many decision trees [15]. For KNN algorithm, k_{nn} is very crucial for the prediction performance. If k_{nn} is too small, the model becomes more complicated and may overlearn when the neighboring points are noises. Contrarily, large k_{nn} makes the model structure simple, but the neighboring samples with large differences will affect the prediction accuracy.

The GridsearchCV module of SciKit-Learn [33] is employed to find the optimal combination of hyperparameters by exhaustively searching over possible range values of the hyperparameters as tabulated in Table 3. After finding the optimal values, the hyperparameters values are then adjusted, and the optimal prediction model is built. Path loss can now be predicted by using the test dataset.

Table 3. Hyperparameters for KNN and random forest algorithms.

Algorithm	Hyperparameter	Range	Optimum Value
Random Forest	Number of Ensemble	10–200	100
	d_{tree}	5–50	40
KNN	k_{nn}	1–20	3

5.7. Model Evaluation and Regression Metrics

The performance of the path loss prediction models is measured by samples in the testing dataset. The evaluation metrics include complexity and prediction accuracy. These metrics are applied to assess the

prediction performance of the trained model. The statistical error indicators, namely mean absolute error (MAE), root mean square error (RMSE), and R^2 score [25], are chosen as accuracy metrics. These accuracy metrics give a more congenital perspective for the evaluation of regression models. R^2 score measures how much variability in dependent variables can be explained by the trained model. A bigger R^2 score value indicates that a better fit between the prediction and the path loss value. MAE is an absolute measure of the goodness for the fitted model. RMSE gives an absolute number on how much the predicted results deviate from the actual ones [15]. MAE and RMSE are better used to compare performance between different regression models [12]. The accuracy metrics can be calculated by comparing the predictions with path loss values in the test dataset as in [11].

$$\begin{aligned} MAE &= \frac{1}{I}\sum_{i}^{I}|PL''_{k,i} - PL'_{k,i}|, \\ RMSE &= \sqrt{\frac{1}{I}\sum_{i}^{I}(PL''_{k,i} - PL'_{k,i})^2}, \\ R^2 &= 1 - \frac{\sum_{i}(PL''_{k,i} - PL'_{k,i})^2}{\sum_{i}(PL''_{k,i} - \overline{PL})^2}, \end{aligned} \tag{6}$$

where I is the total number of samples of test dataset, $PL''_{k,i}$ the path loss value of the ith sample in the test dataset, $PL'_{k,i}$ the predicted value, and $\overline{PL}$ the mean value $\frac{1}{I}\sum_i PL''_{k,i}$. It is worth mentioning that R^2 score is the only metric that manifests the overfitting problem, where one of the symptoms of occurrence in an overfitting problem is a high R^2 score [25].

5.8. Computational Complexity

Another aspect in the assessment of the performance of the path loss prediction models based on ML methods is the complexity. It is necessary to evaluate the path loss value in a very short time, so that the spatial distribution of electromagnetic fields can be quickly updated in response to the propagation environment changes. The complexity associated with the ML-based algorithms is tabulated in Table 4 as presented in [32]. The ensemble methods generally multiply the complexity of the original model by the number of ensembles within the model and replace the training size by the size of each ensemble. For KNN algorithm, it is necessary to compare the distance between the query point and every point in the test dataset. This amounts to N operations for N samples in the dataset. For Euclidean distance, this operation is performed in $\mathcal{O}(p)$ operations, where p is the number of features. Meanwhile, in training a number of decision trees (n_{trees}) within the random forest algorithm, a split has to be found until a d_{tree} has been reached [34]. It is obvious from the table that the KNN algorithm has a lower complexity than the random forest algorithm. Moreover, KNN algorithm is defined as a memory-based non-parametric approach [32], hence it can immediately adapt to the new changes in the smart radio environment.

Table 4. Path loss prediction model complexity.

Algorithm	Training	Testing
KNN	-	$\mathcal{O}(Np)$
Random Forest	$\mathcal{O}(N^2pn_{trees})$	$\mathcal{O}(pn_{trees})$

Bagging ensemble based KNN algorithm frequently requires distance computation of k_{nn} nearest neighbors for ensembles [28]. With the growth in dataset size, the KNN algorithm becomes computationally intensive. There are many accelerated variants of KNN algorithm, and they achieve better saving in the complexity. These algorithms are called KD-trees and locality-sensitive hashing

[35]. Although these algorithms can significantly reduce the complexity, there is a trade-off between the prediction accuracy and reducing a model complexity burden. Alternatively, there are newly developed application-specific integrated circuit (ASIC) customized so-called tensor processing unit (TPS) for ML workloads instead of traditional central processing units (CPU) [36]. TPS can best offer faster processing and more memory capacity for ML workloads.

6. MODELS VALIDATION AND RESULTS

The performance is evaluated for the path loss prediction models and compared with the path loss values in test dataset by using Tables 3 and 1. The path loss calculations are implemented in Anaconda environment using different packages [37] and Matlab software [23]. To visually demonstrate the prediction performance of path loss models, the first 100-samples in the test dataset is used to predict the path loss. The predicted path loss results are shown in Fig. 4. It is shown that the ML-based models can accurately predict the path loss values. As the ML-based models give good prediction results, they well capture the link between input features and path loss value. In support of the obtained results, the residual plots and error distribution histograms are chosen to further compare between the KNN algorithm and random forest algorithm.

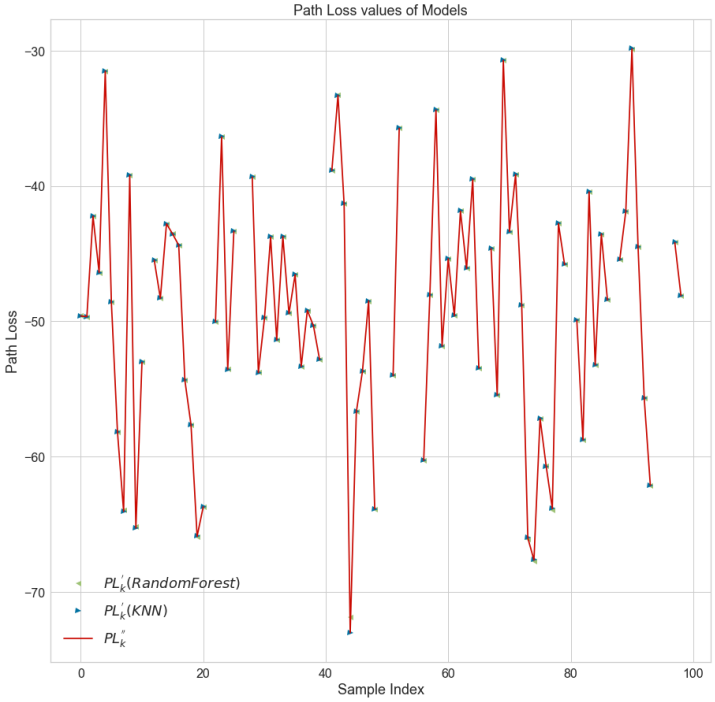
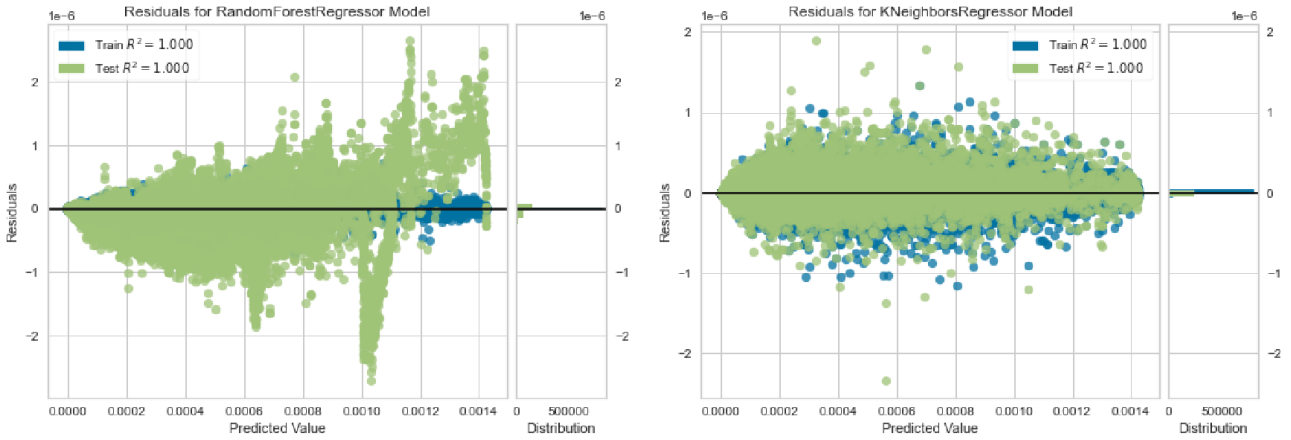


Figure 4. Comparison between evaluated and predicted path loss values.

The residual plots for the regression models are shown in Fig. 5(a) and Fig. 5(b) (left subplots). The plots show the residuals of each path loss prediction model on the vertical axis computed for the training dataset and test dataset, and the predicted path loss values on the horizontal axis. This allows to finely observe the regions within the target values, which may be error prone. The residuals for KNN algorithm are more randomly dispersed around the horizontal axis than the random forest algorithm. Mostly, the residuals for KNN and random forest algorithm are at an order $10e-6$, which indicates that the path loss prediction models are performing well. Recall that the low residuals are attributed to the model tuning, where the optimal hyperparameters significantly influence the optimization of the model accuracy performance.

The prediction error distribution metric representing the proportion of occurrence of the prediction error situated in each given interval is employed to compare the model accuracy. For an accurate



(a) Random Forest (b) KNN

Figure 5. Residuals and prediction error distribution.

model, prediction errors should concentrate on both sides of zero, and the maximum error must be close to zero [10]. The histograms (right subplots) for the path loss prediction models show that the error is normally distributed around zero for KNN algorithm, which also indicates a well fitted model. Meanwhile, the prediction error distribution of random forest algorithm shows uneven error distribution around zero. Later, to further investigate other accuracy metrics, the resulting MAE, RMSE, and R^2 score of the prediction results are tabulated in Table 5. It confirms that the KNN algorithm outperforms random forest algorithm for the three-accuracy metrics.

Table 5. Accuracy metrics.

Algorithm	MAE	RMSE	R^2 test data
KNN	1.336E-08	4.742E-08	0.9999998
Random Forest	8.070E-08	3.1445E-07	0.9999950

For random forest algorithm, the significance of different input parameters is analyzed by using the permutation-based importance method described in [33]. The importance of a feature basically shows how much the input feature is frequently used in a decision tree within the random forest model. This method will randomly shuffle each feature and compute the change in the performance of the model. From Table 6, the input features, including propagation distance, and UE aperture area record the higher scores with respect to other input features in building random forest model. This can be as well inferred from the developed path loss model, and also the path loss model is mainly a function of the propagation distance as discussed in [17].

Table 6. Normalized importance of input features for random forest algorithm.

Features	Importance %
Propagation Distance	44.61
UE Aperture Area	36.09
Spatial Resolution	13.22
UE Altitude	3.95
Elevation Angle	2.12

7. CONCLUSION

The paper presents a machine learning framework for the path loss prediction problem of the LIS-to-UE in far-field region. The framework is based on machine learning and electromagnetics in predicting path loss model. By considering the qualitative aspects of the LIS-assisted wireless communication rather than relying on commonly used free-space path loss, the developed path loss model reveals an insightful aspect in controlling the path loss. It shows that the area of the LIS antenna aperture plays an important role in controlling the path loss for the LIS-assisted wireless communication. Therefore, it can serve as a guidance for determining the LIS physical area to achieve a certain path loss value less than the path length free space channel. It has also been demonstrated that machine learning framework provides a flexible modeling approach based on the training dataset for LIS-to-UE architecture. The path loss evaluation problem for the LIS-assisted wireless communication in the far-field region can be better solved by the supervised methods in ML based. Furthermore, the KNN algorithm has better prediction performance in terms of accuracy metrics like MAE, RMSE, R^2 score, and complexity than random forest algorithm. In addition, the importance of the input features has been analyzed for the random forest algorithm. Importance results show that the propagation distance is the dominant input feature. The LIS-assisted wireless communication is a newly emerging scenario, and the channel modeling and path loss prediction are still very interesting research topics. Future work will incorporate introduction of more machine learning-based models and different scenarios.

REFERENCES

1. Yuan, J., H. Q. Ngo, and M. Matthaiou, "Towards large intelligent surface (LIS)-based communications," *IEEE Transactions on Communications*, Vol. 68, No. 10, 6568–6582, 2020.
2. Najafi, M., V. Jamali, R. Schober, and H. V. Poor, "Physics-based modeling and scalable optimization of large intelligent reflecting surfaces," *IEEE Transactions on Communications*, Vol. 69, No. 4, 2673–2691, 2021.
3. Dardari, D., "Communicating with large intelligent surfaces: Fundamental limits and models," *IEEE Journal on Selected Areas in Communications*, Vol. 38, No. 11, 2526–2537, 2020.
4. Han, Y., W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 8, 8238–8242, 2019.
5. Kundu, N. K. and M. R. Mckay, "Large intelligent surfaces with channel estimation overhead: Achievable rate and optimal configuration," *IEEE Wireless Communications Letters*, Vol. 10, No. 5, 986–990, 2021.
6. Taha, A., M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," *2019 IEEE Global Communications Conference (GLOBECOM)*, 1–6, 2019.
7. Di Renzo, M., F. Habibi Danufane, X. Xi, J. de Rosny, and S. Tretyakov, "Analytical modeling of the path-loss for reconfigurable intelligent surfaces — Anomalous mirror or scatterer?," *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 1–5, 2020.
8. Tang, W., M. Z. Chen, X. Chen, J. Y. Dai, Y. Han, M. Di Renzo, Y. Zeng, S. Jin, Q. Cheng, and T. J. Cui, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Transactions on Wireless Communications*, Vol. 20, No. 1, 421–439, 2021.
9. Yildirim, I., A. Uyrus, and E. Basar, "Modeling and analysis of reconfigurable intelligent surfaces for indoor and outdoor applications in future wireless networks," *IEEE Transactions on Communications*, Vol. 69, No. 2, 1290–1301, 2021.
10. Wen, J., Y. Zhang, G. Yang, Z. He, and W. Zhang, "Path loss prediction based on machine learning methods for aircraft cabin environments," *IEEE Access*, Vol. 7, 159251–159261, 2019.

11. Zhang, Y., J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: Principle, method, and data expansion," *Applied Sciences*, Vol. 9, No. 9, 2019, [Online]. Available: https://www.mdpi.com/2076-3417/9/9/1908.
12. Duangsuwan, S., P. Juengkittikul, and M. Myint Maw, "Path loss characterization using machine learning models for GS-to-UAV-enabled communication in smart farming scenarios," *International Journal of Antennas and Propagation*, Vol. 2021, 5524709, Aug. 2021, [Online]. Available: https://doi.org/10.1155/2021/5524709.
13. Aldossari, S. and K.-C. Chen, "Predicting the path loss of wireless channel models using machine learning techniques in mmWave urban communications," *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 1–6, 2019.
14. Juang, R.-T., "Explainable deep-learning-based path loss prediction from path profiles in urban environments," *Applied Sciences*, Vol. 11, No. 15, 2021, [Online]. Available: https://www.mdpi.com/2076-3417/11/15/6690.
15. Zhang, Y., J. Wen, G. Yang, Z. He, and X. Luo, "Air-toair path loss prediction based on machine learning methods in urban environments," *Wireless Communications and Mobile Computing*, Vol. 2018, 8489326, Jun. 2018, [Online]. Available: https://doi.org/10.1155/2018/8489326.
16. Ellingson, S. W., "Path loss in reconfigurable intelligent surface-enabled channels," *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 829–835, IEEE, 2021.
17. Molisch, A. F., K. Balakrishnan, D. Cassioli, C.-C. Chong, S. Emami, A. Fort, Johan, Karedal, J. Kunisch, H. G. Schantz, U. G. Schuster, and K. Siwiak, "IEEE 802.15.4a channel model-final report," 2004.
18. Hu, S., F. Rusek, and O. Edfors, "Beyond massive mimo: The potential of data transmission with large intelligent surfaces," *IEEE Transactions on Signal Processing*, Vol. 66, No. 10, 2746–2758, 2018.
19. Jide, Y., H.-Q. Ngo, and M. Matthaiou, "Large intelligent surface (LIS)-based communications: New features and system layouts," *IEEE International Conference on Communications*, IEEE, arXiv:2002.12183, Feb. 2020.
20. Balanis, C. A., *Antenna Theory: Analysis and Design*, Wiley-Interscience, 2005.
21. Greenquist, R. E. and A. J. Orlando, "An analysis of passive reflector antenna systems," *Proceedings of the IRE*, Vol. 42, No. 7, 1173–1178, 1954.
22. Silver, S., *Microwave Antenna Theory and Design*, S. Silver, et al., Ed., [Massachusetts Institute of Technology. Radiation Laboratory Series. No. 12], McGraw-Hill Book Company, 1949, [Online]. Available: https://books.google.com.eg/books?id=Fi42MwEACAAJ.
23. MATLAB, "Version 7.10.0 (R2010a)," Natick, The MathWorks Inc., Massachusetts, 2010.
24. Ozdogan, O., E. Bjornson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Communications Letters*, Vol. 9, No. 5, 581–585, 2020.
25. Muller, A. and S. Guido, "Introduction to machine learning with python: A guide for data scientists," O'Reilly Media, 2016, [Online]. Available: https://books.google.com.eg/books?id=vbQlDQAAQBAJ.
26. Huang, J., N. Huang, L. Zhang, and H. Xu, "A method for feature selection based on the correlation analysis," *Proceedings of 2012 International Conference on Measurement, Information and Control*, Vol. 1, 529–532, 2012.
27. Mohammed, R., J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248, 2020.
28. Ray, S., "A quick review of machine learning algorithms," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39, 2019.
29. Popoola, S. I., A. Jefia, A. A. Atayero, O. Kingsley, N. Faruk, O. F. Oseni, and R. O. Abolade, "Determination of neural network parameters for path loss prediction in very high frequency wireless channel," *IEEE Access*, Vol. 7, 150 462–150 483, 2019.

30. Breiman, L., "Random forests," *Machine Learning*, Vol. 45, No. 1, 5–32, Oct. 2001, [Online]. Available: https://doi.org/10.1023/A:1010933404324.
31. Breiman, L., "Bagging predictors," *Machine Learning*, Vol. 24, No. 2, 123–140, Aug. 1996, [Online]. Available: https://doi.org/10.1007/BF00058655.
32. Rahim, A., Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, Vol. 9, 106 575–106 588, 2021.
33. Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: Experiences from the scikit-learn project," *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122, 2013.
34. Lin, W., Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *IEEE Access*, Vol. 5, 16 568–16 575, 2017.
35. Matsushita, Y. and T. Wada, "Principal component hashing: An accelerated approximate nearest neighbor search," *Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology*, Ser. PSIVT'09, 374–385, Springer-Verlag, Heidelberg, Berlin, 2009, [Online]. Available: https://doi.org/10.1007/978-3-540-92957-4_33.
36. Jouppi, N. P., C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-L. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-datacenter performance analysis of a tensor processing unit," *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, 1–12, 2017.
37. "Anaconda software distribution," 2020, [Online]. Available: https://docs.anaconda.com/.