Machine Learning-Based Methods for Path Loss Prediction in Urban Environment for LTE Networks

Nektarios Moraitis
School of Electrical and Computer Engineering
National Technical Univ. of Athens
Athens, Greece
morai@mobile.ntua.gr

Lefteris Tsipi, Demosthenes Vouyioukas
Dept. of Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
dvouyiou@aegean.gr

***Abstract*—This paper presents prediction path loss models in an urban environment for cellular networks with the help of machine learning methods. For this goal, Support Vector Regression (SVR), Random Forest (RF) and K-Nearest Neighbor (KNN) algorithms are exploited and assessed. The training and testing procedure is carried out with the help of a path loss dataset generated by simulated results considering a Long Term Evolution (LTE) network utilizing a digital terrain model. The simulation takes into account an urban environment for both line-of-sight (LOS) and non-LOS (NLOS) propagation condition. The results reveal that all the evaluated algorithms forecast path loss with a remarkable accuracy, providing root-mean-square errors on the order of 2.1-2.2 dB for LOS and 3.4-4.1 dB for NLOS locations, respectively. Among the examined algorithms, KNN shows the best performance, thus being an appealing option to predict path loss in urban areas. For comparison purposes, the COST231 Walfisch-Ikegami empirical model was applied, which presents the worst performance, providing the highest errors under-predicting path loss, especially in NLOS locations.**

Keywords—3D simulation, Long Term Evolution (LTE), machine learning, path loss prediction, urban environment.

I. INTRODUCTION

Machine learning has been considered to play a key role in modern wireless communications [1]. Machine-learning-based methods are increasingly adopted by the research community and vendors, as a promising alternative for the prediction of path loss, substituting the traditional empirical and deterministic models. Path loss models are very important not only for a comprehensive network planning, but also for evaluating interference [2]. Therefore, it is crucial to develop precise prediction models, since a possible under or overestimation of the path loss could compromise the planning procedure [3]. In essence, path loss prediction can be regarded as a supervised regression problem which can be dealt with various machine learning algorithms, such as support vector machine (SVM), random forest (RF), or K-Nearest Neighbor (KNN) [4]. Traditionally, path loss was predicted by using empirical or deterministic models [5]. However, both methods have their benefits and shortcomings in terms of computational efficiency, applicability, and accuracy [6]. Current research attempts have shown that machine-learning-based path loss models provide more precise prediction results than the empirical ones, and are more computationally efficient than the deterministic models [7], [8].

In this context, machine learning algorithms are endorsed as an attractive option to predict path loss, encompassing both exceptional accuracy and efficiency. To date, path loss prediction based on support vector regression (SVR) and RF methods have been proposed in [6] for an urban microcell scenario at 2 GHz, as well as in [4], and [9] for in-cabin scenarios at 2.4, and 3.52 GHz. Additionally, SVR prediction results have been reported in [10], taking into account an urban environment around 850 MHz. Finally, path loss prediction using KNN method, regarding an unmanned aerial vehicles (UAV) scenario for an urban environment has been examined in [11], exploiting simulated data at 2.4 GHz. Based on the existing literature, there are few research studies that exploit and evaluate different machine learning algorithms, oriented at predicting path loss in urban environments. This renders such studies a challenging endeavor.

In contrast to the existing research efforts, this work aims to assess the performance of various machine learning methods in urban environments for a Long Term Evolution (LTE) network at 2.1 GHz. The results incorporate both line-of-sight (LOS) and non-LOS (NLOS) propagation conditions, with more than 5000 samples produced by ray tracing simulations. To the best of our knowledge it is the first time that different machine learning algorithms are assessed in urban environment for LTE networks. Moreover, it is the first time that the KNN method is tested for an urban LTE cellular scenario. Finally, the predicted path loss is also compared with the COST231 Walfisch-Ikegami model, a widely used empirical model applied in urban environments [12].

The rest of the paper is organized as follows. Section II describes the employed machine learning methods, whereas Section III outlines the simulation procedure. The learning and validation method is described in Section IV. The performance of each algorithm is evaluated and discussed in Section V, followed by a summary in Section VI.

II. MACHINE LEARNING METHODS

A. SVR method

SVR is an extension of SVM and has been successfully exploited to deal with non-linear regression problems for path loss prediction. The main idea is to depict input data from the low-dimensional space to a high-dimensional one through non-

978-1-7281-9722-7/20/$31.00 ©2020 IEEE

linear functions and search for an ideal hyperplane in the high-dimensional feature space, so as the samples fall on this hyperplane as far as possible [4], [13]. The specific hyperplane is given by the following linear expression [6]

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b \tag{1}$$

where **w** is the normal vector that regulates the direction of the hyperplane, **x** is the input vector, $\varphi(\cdot)$ is the non-linear mapping function, and b stands for the bias. Then, using the Lagrange multipliers, and solving the problem according to [10], the prediction path loss function can be determined by

$$PL^{pred} = f(\mathbf{x}) = \sum_{i}^{l} (a_i - a_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \tag{2}$$

where PL^{pred} is the predicted path loss, a_i and a_i^* are the Lagrange multipliers and $K(\cdot,\cdot)$ is a kernel function that realizes the nonlinear mapping from the low- to the high-dimensional space [6]. Finally, l denotes the total number of samples. A very significant feature in SVR is the appropriate selection of the kernel function that achieves the mapping process. In the following, the radial basis function (RBF) kernel is tested for its performance. The related expression is given by

$$K_{\mathrm{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{3}$$

where $\|\cdot\|$ denotes the norm, and γ is the adjustable parameter that is fitted to the data and controls the performance of the kernel ($\gamma = 1/2\sigma^2$). The complexity of the model does not depend on the dimension of the input parameters, averting any dimensionality problems.

B. RF method

RF is a machine learning technique that incorporates both decision tree and bagging [6], whereas it is easy to implement, and parallel computing can be used. Initially, each tree in RF learns from a random sample of the training inputs. The technique of bootstrapping is used where the samples are drawn with replacement, which means that some samples will be used multiple times in a single tree. The next process is known as the feature selection. The node division is performed by randomly selecting k features out of all features. Then the optimal feature is selected out of the k features for node division and the division rule is to minimize the estimation error. Repeating the aforementioned steps, multiple decision trees can be created. The overall prediction can be defined as the average response from all the independently trained trees [4]. This model can forecast the path loss by averaging the predictions according to

$$PL^{pred} = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t(x) \tag{4}$$

where $\hat{p}_t(x)$ stands for the predicted path loss value of the t-th decision tree model, x is the feature set, and T represents the number of the decision tree models. Because of the random selection of features at each node split, RF is not susceptible to sample and feature disturbances, thus providing a better generalization performance.

Fig. 1. Simulated path loss results for BS1 in urban environment.

C. KNN method

KNN is a supervised machine learning algorithm often used for classification problems [11]. The main idea is to find the K training samples closest to the sample under prediction using distance metrics, and then predict the results based on these K neighbours. For regression problems, the final predicted value is the outcome from averaging the K-nearest neighbours. The selected distance metric is the Euclidean formula, and the predicted path loss is provided by

$$PL^{pred} = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{5}$$

where K is the total number of the K-nearest neighbors, n is the number of predictions, and q_i, p_i are the training and the prediction samples, respectively.

III. Simulated Dataset

To train properly the above-mentioned machine learning models, simulated path loss results are exploited. The simulation was carried out using WinProp software (Altair Hyperworks™) [14]. More specifically, ProMan suite was utilized, which includes wave propagation models for different scenarios and network planning simulators for various air interfaces. The simulation takes into account an urban environment, incorporating a three-dimensional digital terrain map of Frankfurt with a resolution of 5 m, including the building layer. Fig. 1 illustrates the simulated environment where three different base stations (BS) locations comprising the LTE network were considered. The produced path loss, indicatively for BS1, is also provided. Each BS has different height (above ground level) and operating frequency. The mobile station (MS)

is fixed at 1.5 m above the ground level. The spatial resolution for the MS locations (i.e. the sampling resolution) was 5 m. The rest of the selected parameters regarding the LTE network are listed in Table I.

The WinProp software, for pure path loss predictions in urban areas, applies the Dominant Path Model (DPM), which uses simplified ray tracing techniques, combining high accuracy with short computation time [15]. To evaluate the feasibility of the proposed machine learning methods, simulations were carried out using the DPM model, so as to generate path loss data for training and testing process. Furthermore, the empirical COST231 Walfisch-Ikegami model was also tested for comparison purposes. On the whole, 5150 path loss samples were collected, out of which 2644 are for LOS and 2506 for NLOS conditions, respectively.

TABLE I. TECHNICAL PARAMETERS OF THE SIMULATED NETWORK.

Frequency	2120 / 2140 / 2160 MHz
BS EIRP	40 dBm
BS Height (a.g.l.)	75 / 95 / 115 m
Bandwidth	20 MHz
BS Antenna type	Omnidirectional
MS Height	1.5 m
MS Antenna type	Omnidirectional
MS Antenna Gain	0 dBi
Transmitted signal	OFDMA/QPSK

IV. TRAINING PROCESS AND VALIDATION

It is worth remarking that machine learning models are susceptible to the range of the training inputs. Furthermore, normalization processing is suitable, so as to accelerate the convergence and improve the algorithm's performance. Therefore, normalization was applied before training the models, where all inputs and outputs are limited within the range of 0 to 1 according to [7]

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

where x is the value that requires normalization, x_{max} and x_{min} represent the maximum and minimum value of the data range, respectively, and x_{norm} is the value after normalization.

In the following, seven parameters that affect path loss, act as the input variables to the machine learning algorithms for the training procedure. The distance between BS and MS, the BS and MS altitudes, the carrier frequency, the coordinates (X,Y) of the receiver location, and finally, the path visibility that indicates the existence of LOS or NLOS conditions between the terminals. To avert overfitting, the k-fold cross-validation method was adopted, which incorporates one training and one validation subprocess. The performance of each model is evaluated during the testing phase. The input dataset is segmented into k-subsets of equal size. A single subset out of k is retained as the test dataset and the remaining k-1 subsets are used as training data. The cross-validation process is then repeated k times, and each of the k subsets are used exactly one time as the test data. The results from all the iterations are then averaged to produce a single estimation.

The training process aims at acquiring the parameters for each machine learning algorithm and optimize the performance and effectiveness of the path loss prediction. The learning and validation procedure for the three selected algorithms was carried out using RapidMiner™ [16], a software platform that provides an integrated environment for data preparation, machine learning and results visualization. For the k-fold method, k was selected to be equal to 10 (515 samples) and the learning procedure was applied for the entire dataset. For SVR the training accuracy depends on the RBF kernel provided in (3), more specifically on γ. After many trials, the best results were obtained for $\gamma = 1$ (i.e. $\sigma = 0.707$). In respect, the accuracy of the Random Forest algorithm is affected by the maximum tree depth and the number of the ensemble members. The best outcome is obtained for a tree depth and an ensemble members number of 10 and 110, respectively. Finally, KNN depends directly on K, which denotes the number of K-nearest neighbors. During the learning process, the best results are yielded for $K = 20$.

To validate the performance of each algorithm, it is important to assess the statistical error between the simulated and the predicted path loss values. The mean error (ME), in decibels, is determined by

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\left(PL_i^{sim} - PL_i^{pred}\right) \tag{7}$$

where PL_i^{sim} and PL_i^{pred} are the simulated and predicted path loss values, respectively, i is the index of the measured sample, and N the total number of samples. A positive value of ME indicates that one machine learning algorithm in general under-predicts the simulated path loss, and over-estimation stands for a negative value. The mean absolute percentage error (MAPE), is provided by

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{PL_i^{sim} - PL_i^{pred}}{PL_i^{sim}}\right| \times 100 \tag{8}$$

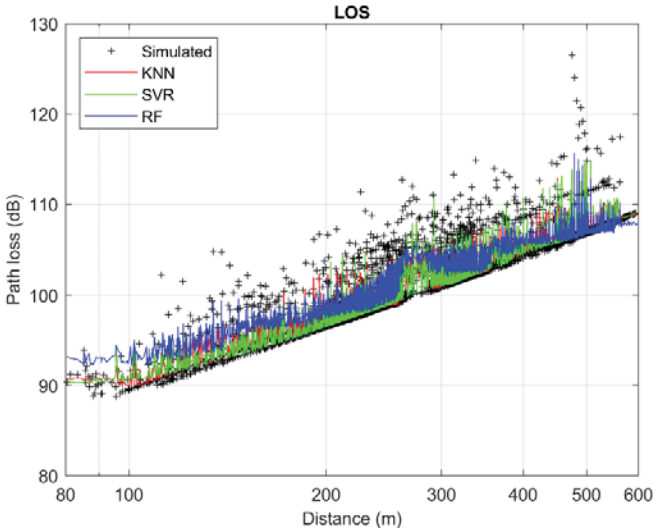
Accordingly, the root-mean-square error (RMSE), is a common metric to assess the performance of the predictive algorithms. It is given, in decibels, by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(PL_i^{sim} - PL_i^{pred}\right)^2} \tag{9}$$

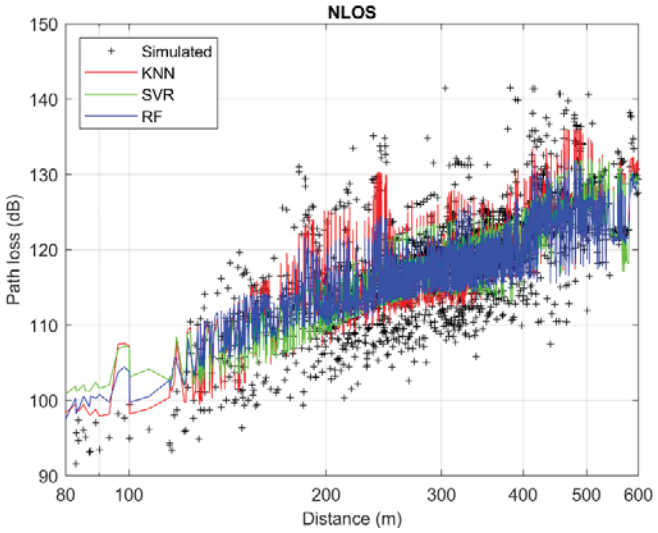
In urban areas, an acceptable RMSE is on the order of 6-7 dB [17], [18].

TABLE II. STATISTICAL ERROR METRICS OF THE EXAMINED MACHINE LEARNING ALGORITHMS AND THE SIMULATED RESULTS IN URBAN ENVIRONMENT.

Examined Model	**KNN**		**SVR RBF**		**Random Forest**		**COST231 WI**	
Propagation Condition	***LOS***	***NLOS***	***LOS***	***NLOS***	***LOS***	***NLOS***	***LOS***	***NLOS***
ME [dB]	0.2	-0.2	0.4	0.3	-0.4	0.4	7.9	18.3
MAPE [%]	1.2	2.2	1.2	2.6	1.6	2.3	7.9	15.5
RMSE [dB]	2.1	3.4	2.2	4.1	2.1	3.6	8.4	19.1
ρ	0.93	0.89	0.92	0.83	0.93	0.89	0.68	0.61
T_{train} [s]	2		304		36		-	

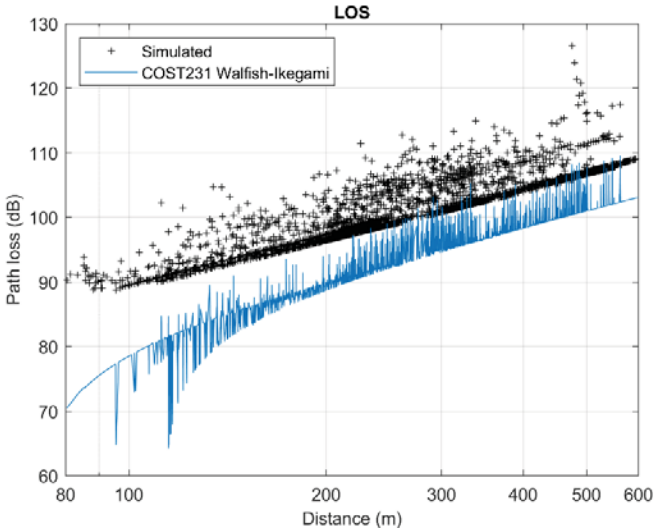


(a)

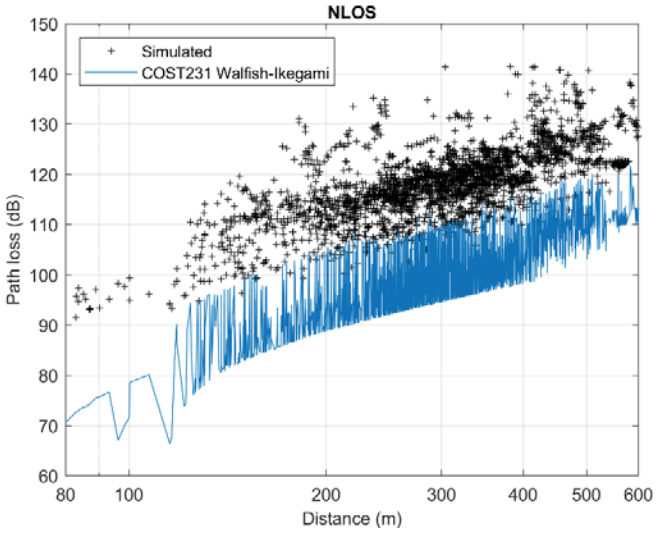


(b)

Fig. 2. Comparison between the simulated path loss and the prediction by machine learning algorithms. (a) LOS condition. (b) NLOS condition.



(a)



(b)

Fig. 3. Comparison between the simulated path loss and the prediction by COST231 Walfisch-Ikegami empirical model. (a) LOS condition. (b) NLOS condition.

Finally, the cross-correlation coefficient, defined as the Pearson product moment correlation [19], is given by

$$\rho = \frac{\sum_{i=1}^{N}\left(PL_i^{sim} - \overline{PL_i}^{sim}\right)\left(PL_i^{pred} - \overline{PL_i}^{pred}\right)}{\sqrt{\sum_{i=1}^{N}\left(PL_i^{sim} - \overline{PL_i}^{sim}\right)^2 \sum_{i=1}^{N}\left(PL_i^{pred} - \overline{PL_i}^{pred}\right)^2}} \quad (10)$$

The cross correlation is a nonparametric measure of the statistical dependence between the simulated, PL_i^{sim}, and the predicted path loss, PL_i^{pred}, respectively. Furthermore, $\overline{PL_i}^{sim}$, $\overline{PL_i}^{pred}$, are the mean values, and N the total number of samples. A correlation greater than or equal to 0.60-0.70 is considered acceptable, indicating the appropriateness of an assessed algorithm.

In the following, all the statistical error metrics are computed using (7)-(10), for each examined machine learning algorithm and the numerical results are summarized in Table II, separately for LOS and NLOS cases. Finally, the training time, T_{train}, for all the examined algorithms was recorded in seconds and presented in Table II. The simulation procedure was carried out on a desktop computer, having an Intel Core i3™ processor at 2.3 GHz and 6-GB of RAM.

V. Results and Discussion

In this section, the three aforementioned machine learning models are evaluated and depicted in terms of path loss in dB versus distance and the produced simulated data for an urban scenario for LTE network. The simulated path loss versus distance for LOS and NLOS conditions is shown in Fig. 2, superimposing the machine learning predictions applying KNN, SVR and RF algorithms. In respect, Fig. 3 compares the path loss between the simulated outcome and the empirical COST231 Walfisch-Ikegami model. The results are provided separately for presentation purposes. All the machine learning models can reproduce in detail all the path loss fluctuations, as it is obvious in Fig. 2. According to the error parameters, shown in Table II, all the machine learning algorithms outperform the empirical model prediction. In essence, the applied COST231 Walfisch-Ikegami model under-predicts path loss, as one can observe in Fig. 3. The mean error reaches 18.3 dB, and the percentage deviation approximates 15.5%, especially in NLOS cases. The low cross-correlation values also validate this inconsistency.

On the other hand, the statistical errors regarding the examined machine learning algorithms remain remarkably low, especially for LOS conditions, as one can observe in Table II. The RMSE and MAPE are on the order of 2.1-2.2 dB, and 1.2-1.6%, respectively. Furthermore, in NLOS cases the applied algorithms present comparable performance, although the RMSE is increased about 1.5 dB compared with LOS. Lower cross-correlation values are also observed. Finally, MAPE increases up to 2.2-2.6%, about 1.0% higher than in LOS locations.

The best performance, according to the cross-correlation metric, is observed by the KNN and RF algorithms for both LOS (0.93) and NLOS cases (0.89). In LOS cases, both exhibit an RMSE value of 2.1 dB, which is the lowest. Additionally, KNN presents the best performance in NLOS cases providing an RMSE of 3.1 dB, whereas RF and SVR algorithms yield an RMSE of 3.6, and 4.1 dB, respectively. Furthermore, for the seven training input parameters, SVR algorithm shows the highest training time (304 s) among the examined models, followed by RF (36 s) and KNN (2 s). The computational complexity of SVR model accounts for the highest training time, which is the main weakness of the specific method. On the other hand, the lowest observed time for KNN is attributed to its simplicity, since it calculates only the Euclidean distances. It is regarded as a "Lazy Algorithm", since, in essence, there is no training phase, thus reducing the entire calculation time. Finally, based on the ME parameter, all the examined algorithms present very good adaptability to the simulated path loss, without a severe incidence of over- or under-prediction. Only SVR and RF models present a minor under- and over-estimate of path loss (between -0.4 and 0.4 dB).

The results from the examined SVR-RBF algorithm are comparable with those reported in [10], where measurements in a mixed urban environment (incorporating LOS and NLOS cases) were taken into account at frequencies around 850 MHz. The yielded RMSE was on the order of 4.6 dB for a Gaussian kernel, yet the best performance is achieved for a Laplacian SVR. Similar RMSE values were also obtained in [6], testing an RBF-SVR (4.2 dB) and an RF (3.9 dB) method, considering a measured dataset from a mixed urban area around 2 GHz. Furthermore, higher RMSE values were found, applying various algorithms for digital terrestrial television based on point-to-point measurements in urban areas [8]. Values between 5.9 and 6.3 dB were obtained for SVR, RF and KNN models, where RF presented the best performance. Moreover, SVR and RF methods exhibit very good performance providing lower RMSE when applied to measured datasets at 3.52 GHz [4], [9]. However, this comparison is not straightforward since the specific results refer to in-cabin scenarios.

Finally, simulated datasets in [11] were exploited to investigate the performance of RF and KNN algorithms. An urban propagation environment was taken into account for a UAV scenario at 2.4 GHz. The best results are achieved for an RF model, which provides an RMSE of 3.1 dB, a value in between those presented in Table II. On the other hand, KNN demonstrated higher RMSE that reached up to 8.9 dB, far above from those yielded in this work. However, the air-to-ground propagation scenario accounts for any of the observed differences in error metrics. Based on the above results and discussion, among the machine learning methods applied in this study, KNN shows a remarkable adaptability to the simulated

dataset with low training time, thus being an attractive option for accurate path loss predictions in urban locations.

VI. Conclusion

This paper presented machine learning methods that aimed to predict path loss in urban environment for LOS and NLOS conditions. For this purpose, SVR, RF and KNN algorithms were evaluated. The training and testing procedures were carried out with a path loss dataset derived by simulated results considering an LTE network at 2.1 GHz. The results showed that all machine learning algorithms outperformed other, widely used, empirical models such as COST231 Walfisch-Ikegami, providing much lower statistical errors. Comparing the performance between the assessed machine learning algorithms, KNN demonstrated the best performance, thus can be considered as an option for precise path loss predictions in urban locations.

References

[1] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60-69, Sep. 2019.

[2] S. I. Popoola, S. Misra, and A. A. Atayero, "Outdoor path loss predictions based on extreme learning machine," *Wireless Pers. Commun.*, vol. 99, no. 1, pp. 441-460, Springer, Mar. 2018.

[3] N. Moraitis, D. Vouyioukas, A. Gkioni, and S. Louvros, "Measurements and path loss models for a TD-LTE network at 3.7 GHz in rural areas," *Wireless Netw.*, vol. 26, no. 4, pp. 2891-2904, Springer, Jan. 2020.

[4] J. Wen, 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, pp. 159251-159261, Oct. 2019.

[5] E. Östlin, H. J. Zepernick, and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 2735-2747, Aug. 1980.

[6] Y. Zhang, J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: principle, method, and data expansion," *Appl. Sci.*, vol. 9, no. 9, pp. 1-18, May 2019.

[7] M. Ayadi, A. Ben Zineb, and S. Tabbane, "A UHF path loss model using learning machine for heterogeneous networks," *IEEE Trans. Antennas Propag.*, vol. 65, no. 7, pp. 3675-3683, July 2017.

[8] C. E. G. Moreta, M. R. C. Acosta, and I. Koo, "Prediction of digital terrestrial television coverage using machine learning regression," *IEEE Trans. Broadcast.*, vol. 65, no. 4, pp. 702-712, Dec. 2019.

[9] X. Zhao, C. Hou, and Q. Wang, "A new SVM-based modeling method of cabin path loss prediction," *Int. J. Antennas Propag.*, vol. 2013, Apr. 2013, Art. no. 279070.

[10] R. D. A. Timoteo, D. Cunha, and G. D. C. Cavalcanti, "A proposal for path loss prediction in urban environments using support vector regression," *in Proc. The Tenth Advanced Int. Conf. on Telecommunications (AICT'14)*, pp. 119-124, Paris, France, Jul. 2014.

[11] Y. Zhang, J. Wen, G. Yang, Z. He, and X. Luo, "Air-to-air path loss prediction based on machine learning methods in urban environments," *Wireless Commun. Mobile Computing*, vol. 2018, Article ID 8489326, Jun. 2018.

[12] COST Action 231 (1999). Digital mobile radio towards future generation systems, final report. Tech. Rep. European Communities, 18957.

[13] M. Awad and R. Khanna, "Support vector regression," *in Proc. Neural Inf. Process. Lett. Rev.*, vol. 11, no. 10, pp. 203-224, Oct. 2007.

[14] Accessed on May 14, 2020. [Online]. Available: https://altairhyperworks.com/product/feko/winprop-propagation-modeling.html.

[15] R. Wahl, G. Wölfle, P. Wertz, P. Wildbolz, and F. Landstorfer, "Dominant path prediction model for urban scenarios," *in Proc. 14th IST Mobile and Wireless Communications Summit*, Dresden, Germany, Jun. 2005.

[16] Acessed on May 15, 2020. [Online]. Available: https://rapidminer.com.

[17] W. C. Y. Lee, *Mobile Communications Design Fundamentals*. Wiley Series in Telecommunications and Signal Processing. John Wiley & Sons, Ltd, Chichester, UK, 1993.

[18] J. D. Parsons, *The Mobile Radio Propagation Channel*. 2nd ed., New York: Wiley, 2000.

[19] X. Zhou *et al.*, "Experimental characterization and correlation analysis of indoor channels at 15 GHz," *Int. J. of Antennas and Propag.*, vol. 2015, Article ID 601835, Hindawi, 2015.