# Path Loss Prediction based on machine learning methods in smart campus environment

Harsh Singh
*Department of Electronics and Communication*
*IIIT Naya Raipur*
Raipur, India
harshs17101@iiitnr.edu.in

***Abstract*—** Path loss models are important to predict signal accessibility, use limited network resources effectively and optimize performance of wireless communications.. Many researches have been carried out taking deployment of machine learning for modelling of path loss propagation in different scenarios such as railway, indoor and outdoor environment but contribution towards building path loss models for smart campus environment has not been processed yet. Our efforts in modelling the pass loss for campus environment through machine learning algorithms resulted in surprisingly highly accurate . In the presented work, measured data from Covenant University, Ota, Ogun State, Nigeria are used to train and evaluate the performance of different machine learning methods such as random forest, k-nearest neighbour, artificial neural networks (ANN) to prepare a suitable path loss model for smart campus environment. The statistical analysis shows that machine learning models outperform traditional propagation models such as log distance model but in this paper the research has been further extended to compare machine learning models with standard empirical models such as COST-231 Hata with regard to accuracy, complexity, and prediction time. The prediction performance of trained models is assessed on both the train set and test set according to the metric such as mean squared error.

***Keywords—5G Communication; machine learning; artificial neural networks; path loss prediction***

## I. Introduction

In the Information technology era, rapid increase of smart devices that have in-built sensors and embedded with Wireless Fidelity (Wi-Fi) module are revolutionizing the way things are done in campus environment [1] . Signal strength in wireless channel decreases as the distance between the mobile and base station increases. The complex propagation environment such as smart campus makes the prediction of the received signal strength very difficult. Path loss is used to describe the decrement in strength of an electromagnetic wave as it travels through space [2] . A highly efficient, easy, and general model for the path loss is essential for calculation of link budget, prediction of coverage made by signal, optimizing performance of system, and selection of base station (BS) locations. Contemporarily, researchers and engineers have made all possible efforts to find out logical models for the path loss prediction in different situations and at different frequencies.se models. The most awaited fifth-generation (5G) networks are planned to uphold increased efficiency, vast coverage, better connection strength, decreased radio latency. The network architecture of 5G mobile communication systems will overcome difficult challenges. Highly accurate models are required to predict path loss to avoid any discrepancies in such an efficient networks. With the development of the machine learning domain, it's applications deployment have been started in many domains such as economics, fintech [14] etc. It has been observed that these algorithms are well capable of solving many problems in wireless domain [15] . Many researchers have already implemented these algorithms in modelling of path loss considering the general scenario such as urban and suburban environment [13] or in other way indoor and outdoor scenario but the broad research has not been done in implementing this algorithms for specific cases like cabin environment . We consider a specific scenario of smart campus and modelled path loss accordingly in this through these algorithms and found these results highly accurate than the result obtained through the empirical models and practically implementable. Normally it has been found that researchers made comparison of the results obtained through these algorithms and simple model like log distance model [13] but here taking a step ahead we have made comparison of the results obtained through these algorithms and standard empirical model such as COST-231 Hata model. Further here section 2 deals with traditional methods including empirical and deterministic ones, after then there are short descriptions of the algorithms that we have implemented including tuned hyperparameters values specifically based on experiments on our model. Models comparisons, further opportunities and conclusion have been discussed later.

## II. Traditional methods

Conventionally, path loss is calculated through empirical models and deterministic models [3] . Both of them have different specific characteristics, advantages and disadvantages which are discussed below in the gist.

### *A. Empirical models*

These models are generally based on frequency range and a specific scenario. They describe statistical connection between path loss and scenario parameters such as distance between antennas, antenna heights, frequency, and so on. For example, COST-231 Hata model [2] uses mobile station antenna height correction, which is determined empirically for different scenarios such as suburban and urban environments. Mathematical formulation of COST-231 Hata model is described as

$$PL = A + Blog(d) + C \qquad (1)$$

Where A, B and C are factors that depend upon frequency and antenna height.

Other typical empirical models include the log-distance method, Okumra, Hata, and so on [2] [4] . These models are very simple because they require few parameters and equations of these models are concise. Parameters of these models are extracted from a specific scenario so when these models are applied to some general scenario then prediction highly deviated to measured result is obtained [5] . These models can only represent the statistics of path loss at a given distance , but they can not give the actual received power at a specific location. Further empirical models have their own limitations in frequency range such as Okumura model can work in frequency ranging from 1500-1920 Mhz, Okumura-Hata model works frequencies below 1500 Mhz and COST-231 model works for frequencies upto 2 Ghz.

### *B. Deterministic models*

These models use radio wave propagation mechanism and numerical analytics for modelling. Models which come under this category are ray-tracing [12] and finite-difference time-domain.In general, these models are highly accurate and can achieve desired accuracy but the problem with these models is that they are computationally full of complexity and take very much time in real environments. Scenario specific geometry and dielectric properties of materials involved are also required. Change in propagation environments causes to run the time consuming computation process once again.

## III. Path loss models with machine learning approach

Machine learning is a method based on measured data and generalized model frame to predict something. These methods can be classified as supervised and unsupervised learning .These methods are widely used in different applications such as data mining, computer vision and other fields. Path loss prediction is supervised learning problem because of labelled data.So this problem is solvable by machine learning algorithms such as random forest, ANN and support vector machine where output is turned into categorical variable. It has been observed that machine learning methods are more accurate than empirical methods and more computationally efficient than deterministic methods [3] [6] .

### *A. Dataset*

Measured data [7] contains columns for 6 input features and their corresponding path loss value. These input features are longitude, latitude, elevation, altitude, clutter height and distance. Measurements are performed from three base station antenna operating at 1800 Mhz frequency. There are 3617 measurements with three different routes. This data is divided with random seed value of 42 in 4:1 ratio into training and testing sets.

### *B. Feature Selection and Scaling*

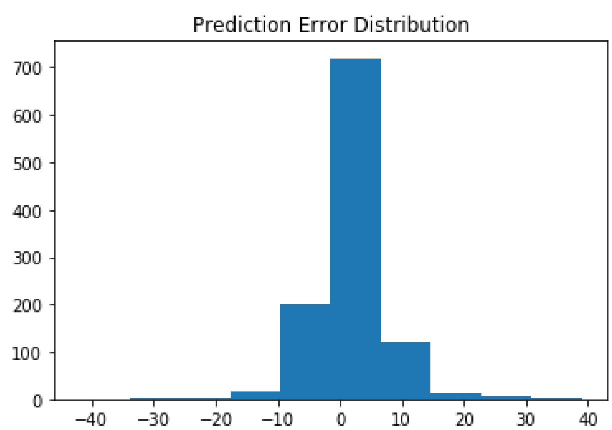
The goal of the feature selection [11] is to select those optimum features which contribute most to the accuracy of the model. In this scenario, feature selection is performed taking correlation of input features. While calculating correlation for each input features it has been found that longitude, elevation and altitude are highly correlated to each other. On the basis of feature importance ranking and correlation among features longitude and altitude dropped. Some machine learning algorithms such as artificial neural networks are sensitive to scale of the input space. So normalization process [8] should be finished before training of model.

### C. *Random Forest based model*

This is machine learning algorithms that combines decision tree [9] . A decision tree is normally understood as a tree containing a root node, some internal nodes, and some leaf nodes. These trees are often tends to be overfit. Due to random selection of features in tree training random forest [10] [11] is having high generalization ability. For predicting path loss, prediction of all individual decision trees are averaged and then highly voted predicted value is considered final predicted value.

$$y' = \frac{1}{T}\sum_{t=1}^{T}\widehat{h_t}(x) \qquad (2)$$

Where T is the number of decision tree models and $\widehat{h}_{t}$ (x) is the prediction of the $t$'th decision tree model. Hyperparameter selected for random forest such as number of trees and number of variable tried to split at each node in our case is 500 and 2 respectively. Prediction error distribution has been shown below in histogramatical form for 1085 observations.




(a)

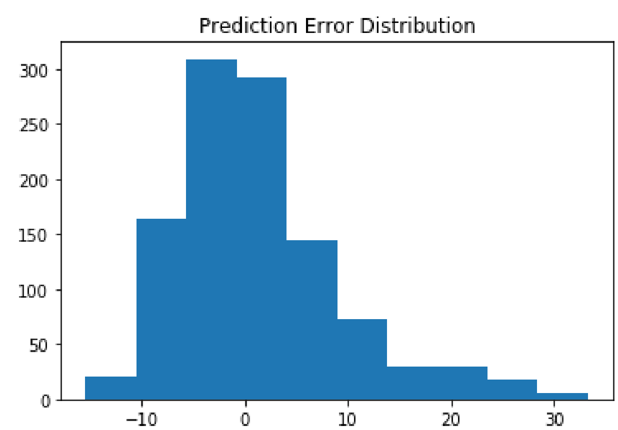
Fig. 1. (a) Prediction error distribution for random forest.

### D. *ANN based model*

ANN are networks formed by interconnecting different layers of neurons . These layers are input layer, one or more hidden layers and output layer for a feed forward multi layer perceptron based neural networks [3] [6] [10] . Neurons are connected to other neurons in other layer by different weights and there is no link up between neurons of the same layer and no cross layer connection among neurons. Number of hidden layers and number of neurons determine the size of network and affect the performance and accuracy of the model. Two hidden layers each with neurons of size 32 and activation function as relu are deployed in our neural network. Back propagation algorithm is used for training the neural network. One thousand epochs are selected to backpropagate the error which occurs due to the mismatch between the real output and predicted output because of untuned weights of connections between neurons. The error which is occuring at the output neuron is termed as mean squared error and mathematically formulated as

$$E = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2 \qquad (3)$$

Where $y_t$ is the target output, $\widehat{y}_t$ is predicted output and N is the number of training samples. We try to minimize this error using gradient descent approach [6] . Prediction error distribution has been shown below in histogramatical for 1085 observations.

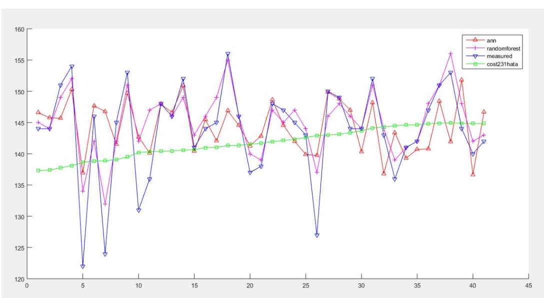



(a)

Fig. 2. (a) Prediction error distribution for ANN.

## IV. Comparison among different models

Neural networks are generally termed as black box in machine which means once inputs are provided to them, they map them with corresponding outputs or output after going through a complex training in the layerwise architecture. It is difficult to explain the learning process and predicted output in case of ANN. In addition, It is based on complicated non linear relationship. On the other side, the decision tree is easy to explain and understand. A random forest based algorithm provides ranking of features based on it's importance so this algorithms does the natural selection of features which is separately required based on correlation or any other feature selection methodology. So if one is opting to implement random forest algorithm on a particular dataset then one automatically gets feature selection in the provided dataset.




(a)

Fig. 2. (a) Plot representing comparison among ANN, random forest, COST-231 Hata and corresponding measured path loss values on 40 values selected from test set.

**Table 1.** Comparison of mean squared error and mean absolute error of random forest and ANN.

| **Model \ Metric** | **MSE** | **MAE** |
| --- | --- | --- |
| Random Forest | 1.927 | 1.62 |
| ANN | 65.89 | 5.88 |

After preparing machine learning models ANN and random forest from training set, these models are deployed for testing set. For better visualization perspective only 40 values are considered from testing set. As these values from measured data are corresponding to increase in distance so clearly path loss from empirical model COST-231 Hata is increasing in the graph. ANN and random forest both are tracing the curve formed by the measured value but random forest is tracing in a better way compared to ANN. Both the machine learning models are way beyond the COST-231 Hata model( an empirical model). Table 1. Is provided For Inner comparison between random forest and ANN considering metrics mean squared error and mean absolute error. It is obvious from the comparison random forest provide high accurate results in comparison of ANN.

## V. Opportunities for further research

It has been seen that enough training data collected from different scenarios plays an important role in generalization ability and enough accuracy of the machine learning model. So if models are built on this bigger and better diversified data then such accuracy and generalization ability can be achieved. Selecting hyperparameters are difficult tasks in machine learning. Approaches like grid search can be used for this. As for ANN-based methods, the number of hidden layers and number of neurons are important to provide a good model. Some techniques should be find to optimize these hyperparameter for better formation of machine learning based models. As the machine learning is evolving area, new algorithms are coming with the passage of time. Models should be formed based on these algorithms and their performance should be checked. Models should be built such that Once they are trained then with the new training examples they can be trained afterwards without retraining them from initial level. This type of learning is called incremental learning. So models should adopt incremental learning as their one of the advantageous characteristics.

## VI. Conclusion

As 5G communication in it's development phase, systems to be deployed and networks to be used must be highly accurate, simple and more diversified. Being a supervised problem path loss can be modelled using machine learning methods. Provided dataset machine learning methods can provide the relationship among path loss and input features. In our analysis, It has been observed that machine learning methods ANN and random forest show good agreement with the measured value. These models are better than empirical models in terms of accuracy and reliability and they do not need geometrical parameters and site specific information to predict path loss as in the case of deterministic models. At the end, we summarize the problem and further research perspective in this area..

## VII. References


[1] V.O. Matthews, Q. Osuoyah, S.I. Popoola, E. Adetiba, A.A. Atayero, C-BRIG: a network architecture for real-time information exchange in smart and connected campuses, in: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2017, London, U.K., vol. 398–401, 5–7 July, 2017.

[2] Phillips, C.; Sicker, D.; Grunwald, D. A survey of wireless path loss prediction and coverage mapping methods. IEEE Commun. Surv. Tutor. 2013, 15, 255–270.

[3] Östlin, E.; Zepernick, H.J.; Suzuki, H. Macrocell path-loss prediction using artificial neural networks. IEEE Trans. Veh. Technol. 2010, 59, 2735–2747.

[4] Erceg, V.; Greenstein, L.J.; Tjandra, S.Y.; Parkoff, S.R.; Gupta, A.; Kulic, B.; Julius, A.A.; Bianchi, R. An empirically based path loss model for wireless channels in suburban environments. IEEE J. Sel. Areas Commun. 1999, 17, 1205–1211.

[5] Ayadi, M.; Zineb, A.B.; Tabbane, S. A UHF path loss model using learning machine for heterogeneous networks. IEEE Trans. Antennas Propag. 2017, 65, 3675–3683.

[6] Isabona, J.; Srivastava, V.M. Hybrid neural network approach for predicting signal propagation loss in urban microcells. In Proceedings of the 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Agra, India, 21–23 December 2016; pp. 1–5.

[7] Path loss dataset for modeling radio wave propagation in smart campus environment Popoola S.I., Atayero A.A., Arausi O.D., Matthews V.O. (2018) *Data in Brief*, 17 , pp. 1062-1073.

[8] Wu, D.; Zhu, G.; Ai, B. Application of artificial neural networks for path loss prediction in railway environments. In Proceedings of the 2010 5th International ICST Conference on Communications and Networking, Beijing, China, 25–27 August 2010; pp. 1–5.

[9] Oroza, C.A[redacted]; Zhang, Z.; Watteyne, T.; Glaser, S.D. A machine-learning based connectivity model for complex terrain large-scale low-power wireless deployments. IEEE Trans. Cogn. Commun. Netw. 2017, 3, 576–584.

[10] Zineb, A.B.; Ayadi, M. A multi-wall and multi-frequency indoor path loss prediction model using artificial neural networks. Arabian J. Sci. Eng. 2016, 41, 987–996.

[11] Liu H. (2011) Feature Selection. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

[12] Piacentini, M. & Rinaldi, F. Comput Manag Sci (2011) 8: 371. https://doi.org/10.1007/s10287-010-0121-8.

[13] Zhang, Y.; Wen, J.; Yang, G.; He, Z.; Wang, J. Path Loss Prediction Based on Machine Learning: Principle, Method, and Data Expansion. Appl. Sci. 2019, 9(9), 1908; https://doi.org/10.3390/app9091908.

[14] N. Golowich, H. Narasimhan, and D. C. Parkes. Deep learning for multi-facility location mechanism design. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 261–267, 2018.

[15] Sun, Yaohua & Peng, Mugen & Zhou, Yangcheng & Huang, Yuzhe & Mao, Shiwen. (2018). Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues.