Development of Path Loss Prediction Model using Feature Selection-Machine Learning Approach

Improvement of Path Loss Prediction Accuracy in Mixed Land-water Case

Bengawan Alfaresi[1], Zainuddin Nawawi[2]*, Bhakti Yudho Suprapto[3]
Doctoral Students of Electrical Engineering, Universitas Sriwijaya, Palembang, Indonesia[1]
Electrical Engineering, Universitas Muhammadiyah Palembang, Palembang, Indonesia[1]
Electrical Engineering, Universitas Sriwijaya, Palembang, Indonesia[2, 3]

***Abstract*—Wireless network planning requires accurate coverage predictions to get good quality. The path loss accurate model requires a flexible model for each area including land and water. The purpose of this research is to develop a Cost-Hatta model that can be applied to the mixed land-water area. The approach used of this research is the three methods of feature selection of machine learning. The first stage of the research was the collection of field data. The measurement data included system, weather, and geographical parameters. The next stage was feature selection to obtain the best composition of features for the development of the model. The feature selection methods used were Univariate FS, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). After obtaining the best features from each method, the next stage was to form a model using four machine learning algorithms, namely Random Forest Regression (RF), Deep Neural Network (DNN), K-Nearest Neighbor Regression (KNN), and Support Vector Regression (SVR). The results of the improvements to the path loss prediction model were tested using the evaluation parameters of Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Percentage Error (MAPE). The results of the testing showed that the improved Cost-Hatta model using the proposed Univariate-RF combination produced a very small RMSE value of 1.52. This indicates that the proposed model framework is highly suitable to be used in a mixed land-water area.**

Keywords*—Path loss; feature selection; machine learning; mixed land-water; Cost-Hatta

I. Introduction

Path loss predication is of great importance in the planning and optimization of coverage in wireless networks [1]. Path loss is used to predict the strength of the signal received by the user. The accuracy of path loss prediction plays an important role in determining the quality of a network design [2].The complexity of environmental characteristics influences the level of complexity in the prediction of received signal strength.In the propagation process, electromagnetic waves undergo a number of treatments caused by various environmental and weather factors that are present around the propagation media. Some of the nearby objects can affect the treatment of electromagnetic wave propagation [3]. The electromagnetic wave treatments that occur include diffraction, refraction, and reflection [4], [5]. These treatments cause fluctuations in the signal power of the receiver due to a weakening in the power of the electromagnetic signal. This signal attenuation is the result of power loss that arises during the electromagnetic wave propagation process in wireless networks.Electromagnetic wave propagation is extremely important in wireless communication systems [6].

Indonesia is an archipelago and has at least 5,590 main rivers and 65,017 tributaries spread across several islands in Indonesia, where the main rivers, watersheds (DAS) in Indonesia reach 1,512,466 square kilometers. People who live in watersheds use water transportation in carrying out the economic activities. The current problem is that there is no path loss prediction model that can be used for water areas, thus, current modeling is not accurate if it is used to plan networks in water areas, especially for areas that are passed by water transportation. Therefore it needs predictive modeling of path loss that can be flexibly used in land and water areas.

Research on the modeling of path loss prediction continues to be carried out to obtain high accuracy predictions in various area conditions. A number of researchers have studied path loss prediction in various kinds of conditions using different variables. Future challenges include the development of high speed wireless telecommunication technology with low latency. Accurate path loss prediction has a strong impact on good quality, low latency, and high throughput.

Conventional predication models developed in the past include empirical and deterministic modeling. Empirical modeling is based on measurements and direct observation in the field. Empirical models provide a statistical picture of the connection between the dependent variables of path loss and a number of measured parameters, specifically frequency, transmitter height, receiver height, and distance between transmitter and receiver [7]. Empirical models are quick and easy to be implemented but have a low level of accuracy, which presents a challenge in empirical model development. Empirical models include Okumura-Hatta, Cost231-Hatta, the ECC model and the Ericsson model [8], [9]. Empirical modeling is the most frequently used type of modeling in planning and optimization systems of wireless networks of telecommunication vendors.

Machine learning is a method of learning about a data set which is used to create a model that can perform a particular task[7]. Machine learning carries out a study of data by learning with the use of algorithms and statistics. Machine learning can be divided into three types, namely supervised learning, unsupervised learning, and reinforcement learning.

*Corresponding Author.

The algorithm of supervised learning can be further divided into two types, namely regression and classification. Based on the research data that exists, the modeling used in path loss prediction falls into the category of supervised learning regression. Regression is characterized by input and output data types in the form of numeric data. Examples of regression types include Support Vector Regression (SVR), Random Forest (RF), Artificial Neural Network (ANN), and K-Nearest Neighbor (KNN). The advantage of machine learning is its high level of accuracy compared with empirical methods [10].

II. Related Work

The focus of the following literature review is the various types of research field, feature variables, and feature selection methods used in the development of path loss prediction models. A number of researchers have developed path loss prediction models using machine learning in various kinds of area condition. These include a study by [11] on indoor building types using an Artificial Neural Network (ANN) model. The research of [12] and [13] also studies path loss prediction with several machine learning models in suburban areas, while [14], [6], and [15] investigate different area types, namely rural, suburban, and urban. Various other research has been developed in different places with a special measuring field, such as the research of [16] which focuses on a vegetation area, and [17], which focuses on the study of path loss prediction in the indoor area of an aircraft cabin. Other studies, such as those by [18], [19], and [20], use an unmanned aerial vehicle (UAV), while the research of [21] focuses on a mixed city-river area. Only a small number of studies have been carried out on path loss prediction in a mixed city-river area. This indicates that there is still room for development of research on path loss prediction in a mixed city-river area.

The types and numbers of input features used in the development of path loss prediction models are highly varied. Research in [22], [23] uses a single input feature, namely distance between the transmitter (TX) and receiver (RX). In addition to using the TX-RX distance feature, the research of [1] includes the feature of frequency as an additional input feature, while [7] also uses two features, with the addition of onboard GPS sensors. The study by [24] uses the feature of TX-RX distance and adds the features of PCC downlink throughput and PDCP downlink throughput as parameters of the input feature. User position based on longitude and latitude is also used as an input parameter, amongst others in the research of [17], to study the indoor area of an aircraft cabin. Longitude and latitude are also used in the research of [11] and [25] to study an outdoor location. In addition to using system parameters, some studies also use environmental parameters as input features. The research of [26], [15] uses the environmental parameters of humidity, temperature, and dew point as input feature parameters. In order to obtain results with a maximum degree of accuracy, some studies use a more complex combination of parameters in accordance with the focus of the characteristics of the research field. The research of [27], [28] uses six input parameters such as longitude, latitude, elevation, altitude, clutter height, and TX-RX distance. This shows that the types and numbers of features can still be developed to match the specific object of the research field.

From the point of view of feature selection process, it is evident that feature selection is still rarely used in most studies. This is because the number of features used in the modeling is relatively small so there is no need to use a feature selection method. In the research of [12], PCA is used to reduce the number of data features and to simplify appropriate modeling. In addition, some research recommends Opportunities for Further Research that are related to types and development of feature selection methods. The research of [7] recommends the use of a feature selection method for further research on path loss prediction, while [16] also recommends further research on the development of feature selection methods. The purpose of including feature selection is to minimize the possibility of eliminating features that are important and relevant to the prediction model.

Main contribution of this research are proposes development of a path loss prediction model for a land-river area by varying the input parameters derived from system parameters and environmental parameters. The second contribution of this research are this model research combine an empirical model with a machine learning model by using three method features selection approach, namely Univariate, GA and PSO combined with the use of four Machine learning models namely Random Forest (RF), Support Vector Regression (SVR), K-Nearest Neighbor (KNN), and Deep Neural Network (DNN) where from the literature study has never been done before.

III. Material and Method

This section is divided into a number of stages: measurement, data processing, delta path loss empirical model calculation, feature selection, and modeling.

Fig. 1 shows the stages carried out in this study starting from the measurement stage then the data processing stage, followed by the best feature selection stage using three approaches methods namely Univariate, GA and PSO. The data with the best features are processed and modeled using four types of machine learning models, namely Random Forest (RF), Support Vector Regression (SVR), KNN Regressor and Deep Neural Network (DNN). The four models produced will be evaluated and obtained as the best model based on the level accuracy of RMSE.

A. Measurement Location

The collection of the research data was carried out in the city of Palembang, Indonesia, which is located at 2°59′27.99″S 104°45′24.24″E. The city of Palembang covers an area of 400.61 km², with an average altitude of 8 meters above sea level. Measurements were taken at Ultra High Frequency (UHF) on a 4G LTE 1800 MHz and 2100 MHz network. The data consisted of a number of input features which were divided into two groups, namely system parameters and environmental parameters.

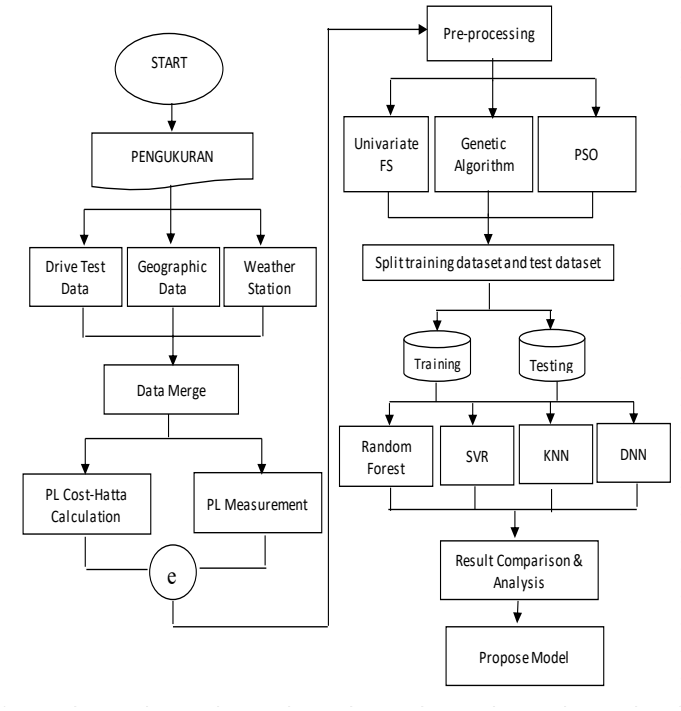

Fig. 1. Research Flow Chart Process.

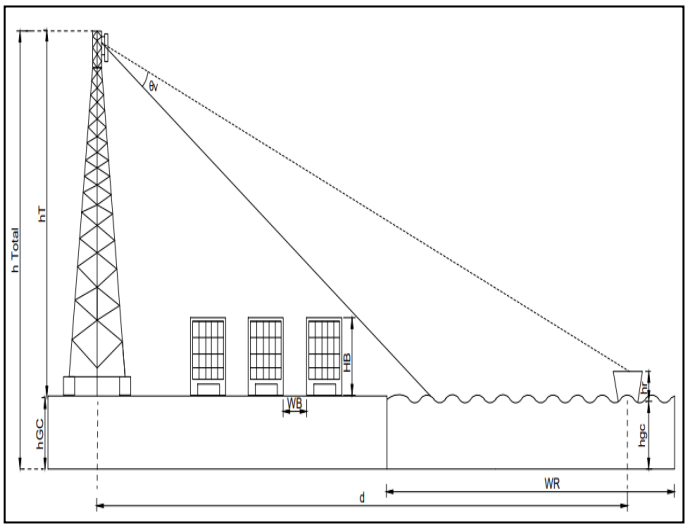

Fig. 2. Measurement Methodology.

Fig. 2 shows the data collection methods for several system parameters such as TX-RX distance, frequency, transmitter height, receiver height, river width, distance between buildings, building height, difference between TX-RX height, distance between transmitter and river border, distance between ship users and river border. The environmental parameters consisted of two segments, namely geographical parameters and weather parameters. The geographical parameters used in this research were slope contour and building density, while the four weather parameters were barometric pressure, temperature, humidity, and dew point.

The data collection was conducted using three measurement methods to obtain the various input and output parameters. A war driving measurement method was used to obtain values of path loss and system parameters. Fig. 3 shows the route taken at the measurement of path loss along the river with a distance of 13 km. The tools used for the war driving measurements were GPS and Dongle which were attached to a laptop. The war driving measurements used two handsets Samsung S5 as receiver, both of which were connected to the laptop. The handsets were locked at the two 4G LTE frequencies. The war driving methodology was used in dedicated conditions or conditions where the handsets were in active (download) mode. At the same time, the weather station tools located at Sriwijaya University took measurements of the weather parameters throughout the data collection. The geographical parameters were obtained based on the geographical maps which were processed using QGIS.

Fig. 3. Selected Route and Building Map on Google Earth for Data Collection.

B. Data Processing

The data processing stage began with the preparation of data from the results of the measurements collected in each of the measurement stages to obtain a number of variables that could be used to create a framework for a model of path loss prediction. The data preparation stage started by processing the data. The results of the war driving measurements in the form of logfiles were treated with time based binning (in seconds) and exported in the form of excel files. The parameter used as path loss value was PUCCH Path Loss. The data of the users' location with the longitude and latitude positions of the data collection were also obtained from this data processing stage. The vertical angle, horizontal angle, and TX-RX distance parameters were calculated based on the angle and position of the BTS transmitter in relation to the user. The distances between the transmitter and river border, and user and river border were calculated based on the straight line intersection of the signal transmission and the river border, which was processed using QGIS.

The geographical parameters of slope contour, building density and distance between buildings were processed using Arcgis. The slope contour was obtained by determining the

difference between the elevation height at the transmitter point and the elevation height at four other points, specifically, 0.25%, 0.5%, and 0.75% of the TX-RX distance, and the difference between the elevation height of the transmitter point (TX) and the elevation height of the receiver point (RX). The average of these values was calculated to find the slope contour value. The building density was found by calculating the number of buildings that were crossed on a straight line between TX-RX, using the intersection on Arcgis. Fig. 4 show a building map was obtained from google earth and converted with Arcgis to find the building points. The distances between buildings were found by calculating the distance between TX-river border (on land) divided by the area of buildings crossed using the intersection on QGIS.

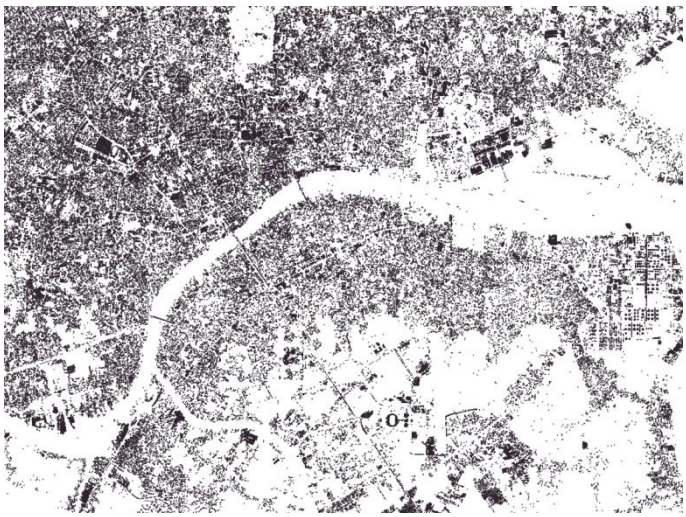

Fig. 4. Building Distribution Map using Arcgis.

Measurements were taken at the weather station to collect weather data about barometric pressure, temperature, humidity, and dew point, which were used as input parameters. The data from the results of the measurements in these three stages were combined to obtain a number of input and output variables.

C. Model Cost-Hatta

The next stage was to find the delta value of the difference between path loss from the result of the measurements and path loss from the result of empirical calculations. The empirical model used in this research was the Cost231-Hatta model. This Cost-Hatta model is a combination of the Cost-231 model and the Hatta model. It can be used to calculate a number of factors, including TX-RX distance, frequency, transmitter height, and receiver height. The Cost-Hatta model is suitable for use in urban areas with a frequency range between 500 MHz – 2000 MHz [29][30]. The formula of the

Cost-Hatta model is shown below:

$$Lu\ (dB) = 46.3 + 33.9 \times \log(f) - 13.82 \times \log(hte) - a(hre) + (44.9 - 6.55 \times \log(ht)) \times \log(d) + CM \quad (1)$$

For urban area:

$$a(hre) = 3.2 \times ((\log(11.75 \times hre)^2) - 4.97 \quad (2)$$

For Sub Urban dan Rural:

$$a(hre) = (1.1 \times \log(f) - 0.7) \times hre - (1.56 \times \log(f) - 0.8) \quad (3)$$

CM: 0 dB for medium size towns and suburban areas

CM: 3 dB for downtown areas

Where *f* is frequency (MHz); *hte* is height of BTS transmitter antenna (m); *hre* is height of receiver antenna (m); *d* is distance between transmitter-receiver (m). The result of the delta calculation of the the difference in path loss was used as an output variable in the modeling.

D. Feature Selection Dan Modelling

The next stage was the process of selecting the features that would be used in the process of developing the model. Feature selection is an important stage in machine learning modelling [31]. This research used three models of feature selection, namely Univariate Feature Selection, Genetic Algorithm, and Particle Swarm Optimization (PSO). The features selected from the three methods were then tested and compared using machine learning.

The Univariate method is a filter method. This kind of method makes an evaluation of every feature in relation to the output variables, then ranks the input features to determine the best features. The Univariate method uses the application of statistical calculations to assign the ranking of each feature. The main criteria used in the Univariate method for the selection of variables are statistical ranking technique and ranking order. After obtaining the ranking results, the next step in this research was to evaluate the number of best features based on the highest ranking, using the "MLPRegressor" model. The python script used was SelectKBest.

The next feature selection method used was Genetic Algorithm. The way this method works is to look for the most suitable composition of features, with the aim of achieving the best prediction accuracy. The Genetic Algorithm method is a search technique based on principles that arise as a result of the inspiration of genetic and evolutionary mechanisms found in a natural system and population of living organisms [32]. In a Genetic Algorithm, every individual in the population represents a candidate solution to the designated problem. The Genetic Algorithm changes a population of individuals by using several genetic functions such as selection, crossover, and mutation [33][34]. Genetic Algorithm is a wrapper method which evaluates every composition of parameter features using machine learning performance as the criteria of evaluation.The genetic algorithm approach is acceptable for various types of solving solutions such as optjmization and calls for scheduling[35].

PSO is based on the idea of the social and cooperative behavior of various species to fulfil their food needs, in this case existing in a multidimensional search space [36][37]. The PSO algorithm consists of a number of main parameters that are used by particles to determine the direction and steps that are then used to determine subsequent movement, in *Pbest* and *Gbest*[36][38]. The position of every particle represents a solution that has a particular fitness value. Particles have their own memory in which they store their best position, referred

to as personal best or *Pbest*[39]. These particles are evaluated in terms of a particular optimization function to identify their compatibility value and ability to hold the best solution. Every particle determines its next position in the search space based on the function of velocity, which calculates the best position of a particle and the best particle position in a population (*Gbest*). These particles will move at each iteration to a different position until they reach an optimal position [40].

After obtaining the best feature composition, the next step of the research was the modeling phase with machine learning. This research used four machine learning algorithms in the modeling process, namely Multi-Layer Perceptron (MLP), Random Forest, Support Vector Regression (SVR), and K-Nearest Neighbor. In the modeling stage, the prediction model was developed by studying how closely the data of the selected input features correlated with the results of the measurement output data.

E. Evaluation and Deployment

The final step was to evaluate the results of the machine learning modeling using an evaluation matrix. This evaluation parameter was used to observe the best accuracy level of the various machine learning models that had been developed. The path loss prediction model is a regression model in which the performance level of the model is calculated by comparing the prediction value with the actual value. The evaluation matrix used in this research included 3 parameters, namely RMSE, MSE, and MAPERMSE is the root of the Mean Square Error which is the evaluation parameter of the regression case.

$$RMSE=\sqrt{\frac{1}{N}\sum_{j=1}^{n}(yj - \hat{y}j)^2} \quad (4)$$

MAE is the average of absolute value of the difference between the actual value and the predicted value. MAE measures the average error between predictions and actual values.

$$MAE=\frac{1}{N}\sum_{j=1}^{N}|yj-\hat{y}j| \quad (5)$$

MAPE (Mean absolute Percentage Error) is the average value of the percentage error error between the actual value and the predicted value.

$$MAPE=\frac{1}{N}\sum_{j=1}^{n}\left(\frac{yj - \hat{y}j}{yj}\right) \text{ x } 100\% \quad (6)$$

yj is the measured path loss, ŷj is the predicted path loss, and N is the number of samples.

IV. Result and Discussion

The preliminary data in this research included a total of 18 candidate variables that consisted of system parameters and environmental parameters. The environmental parameters were divided into two segments, namely geographical parameters and weather parameters. In The first step was to make calculations using the Cost-Hatta model, which is an empirical model. Only four parameter variables were used in this model, namely distance, frequency, height of TX and height of RX. These data were used to determine the value of calculated path loss and to find the delta value of the difference between the measured path loss and calculated path loss.

The next step was to analyze the selection of the best variables from the 18 candidate variables using the stage of feature selection, as shown in Table I.

TABLE I. Candidate Variable

Variable Name	Variable Description	Level
Distance	Distance between transmitter (TX) and receiver (RX)	meters
Frequency	Frequency used in signal transmission	MHz
Height TX	Transmitter antenna height + altitude location	meters
Height RX	Receiver antenna height + altitude location	meters
Vertical angle	The angle difference between the vertical direction of the antenna and the vertical direction of the receiver	degree
Horizontal angle	The angle difference between the horizontal azimuth of the antenna and the horizontal direction of the receiver	degree
Width of River	River Width	meters
Height of Building	Surrounding building height	meters
Distance between Building	Distance between surrounding buildings	meters
Distance_TX to Border (Land)	Distance between transmitter (TX) and river border / distance on land	meters
Distance Border to User (Water)	The distance between the river border and the user / distance on the waters	meters
Delta Height of TX-RX	The difference between the height of the transmitting antenna and the receiving antenna	meters
Slope Contour	The angle between the horizontal plane and the direction of the contour of the ground	degree
Building Density	The building density between tranceiver and receiver	-
Barometric Pressure	Barometric pressure at the time of measurement	hPa
Temperature	Air temperature at the time of measurement	°C
Humidity	Air humidity at the time of measurement	%
Dew Point	Dew Point Value at the time of measurement	°C

A. Result of Empirical Model by using Cost-Hatta

In the preliminary stage, the path loss value was calculated using the Cost-Hatta empirical model. In the existing models, the Okumura-Hatta models are divided according to type of area, whether urban, suburban, or rural. The results of the Cost-Hatta model calculations were compared with the path loss value from the results of the measurements collected in order to obtain the evaluation parameter value.

TABLE II. Evaluation Performance of Cost-Hatta Model

Model	Area	Evaluation		
		RMSE	MAE	MAPE
Cost-Hatta	Urban	31.643	27.114	18.554
	Sub-Urban	25.832	21.319	15.375
	Rural	25.832	21.319	15.375

Table II shows that the Cost-Hatta calculation of urban area had an RMSE value of 27.114 while the suburban/rural

area calculation had an RMSE value of 21.319. This result shows that the calculation model with Cost-Hatta in a suburban area had a higher level of accuracy than the calculation value in an urban area. This indicates that the measurement value in a mixed land-water area was more compatible with the calculation in a suburban area. However, the RMSE value still showed a inadequate level of accuracy because it is exceeded the limit of an RMSE value, which should be less than 10. Therefore, there is a need to improve the model of the Cost-Hatta formula by modeling path loss error using machine learning. Path loss error is the difference between the path loss result from the Cost-Hatta calculation and the path loss value from the measurement result. The path loss error value obtained was used as an output variable in the process of prediction improvement.

B. Results of Feature Selection Process

The first step in the process of developing the path loss prediction model was the feature selection process. The feature selection process used three methods to obtain the best composition of features. The composition of features produced by these three methods is as follows:

1) Univariate feature selection: The Univariate method was used to select the features with the highest level of correlation. The type of score function used was Mutual Information Regression. The next step was to provide alternative feature combinations based on the different degrees of correlation. The first feature selection method is by using Univariate FS. The score function used is mutual info regression. Univariate feature selection works by selecting the best features based on univariate statistical tests.

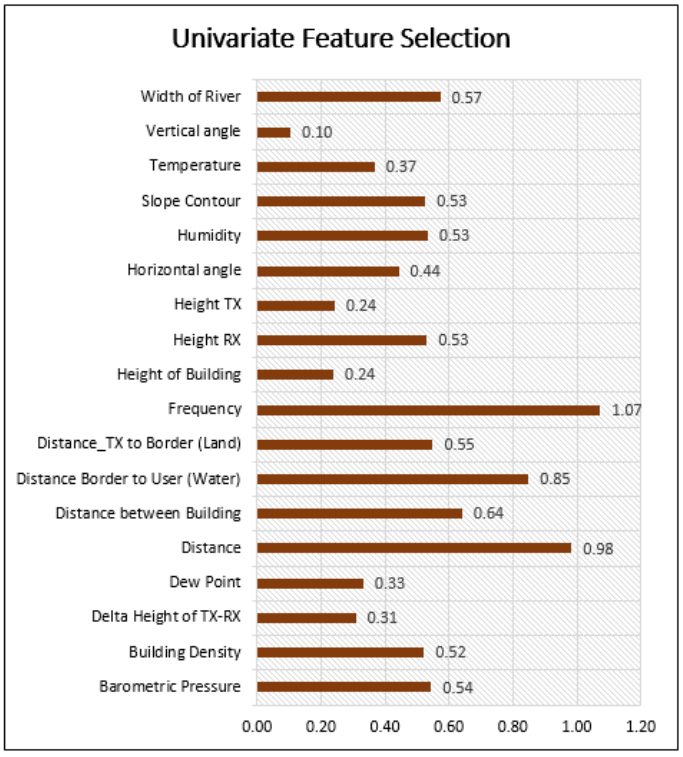

Fig. 5. Value of Mutual Information Univariate for Path Loss Prediction.

Fig. 5 shows that based on metode Univariate FS, frequency has the highest ranking. This indicates that the frequency value had the strongest correlation with the path loss variable, followed by TX-RX distance, border to user distance (water), and distance between buildings (m). This research shows that frequency and distance variables, including TX-RX distance, border to user distance (water), and distance between buildings, play an extremely important role in path loss prediction. The features with the weakest correlation were the parameters of RX vertical angle from TX main beam and building height. The RX vertical angle from main beam variable did not have a significant effect because the measurements were carried out in the NLOS area, so there were many measurement factors that influenced this parameter, such as blockage from buildings and other nearby objects. The building height parameter also had no significant influence because the collection of building data only took into account the height of buildings in the area of the user location point but did not take into account all the buildings between the BTS transmitter location and the receiver location.

Some of these candidate features were modeled simply and evaluated using a machine learning classifier in the form of Random Forest Regression. Table III shows that the best feature combination was achieved by combining the best 17 features, with an RMSE value of 3.07. The combination of these 17 features eliminated the variable with the lowest correlation level to output, which was the RX vertical angle from main beam variable.

TABLE III. CANDIDATE VARIABLE OF UNIVARIATE FEATURE SELECTION

Number of Variables	MAE	MSE	RMSE
1	10.59	171.99	13.11
2	9.13	138.71	11.78
3	9.11	137.15	11.71
4	9.11	140.95	11.87
5	6.68	94.48	9.72
6	6.87	97.13	9.86
7	6.71	93.52	9.67
8	6.74	95.75	9.79
9	6.68	91.89	9.59
10	6.87	98.30	9.91
11	6.53	89.96	9.48
12	6.34	85.27	9.23
13	6.58	88.15	9.39
14	6.62	91.12	9.55
15	6.66	92.65	9.63
16	2.30	10.89	3.30
17	2.18	9.41	3.07
18	2.25	10.38	3.22

2) Genetic algorithm feature selection: The Genetic Algorithm feature searches for the best feature composition by performing an evaluation of every feature combination using a classifier with machine learning. The classifier used in this

research was Random Forest Regression. This research searched for the best composition by altering the values of the parameter settings on the Genetic Algorithm. The population values were changed between 20, 50, and 80. The crossover % values were changed between 0.5, 0.7, and 0.9, and the mutation % values were varied between 0.3, 0.5, and 0.7.

TABLE IV. CANDIDATE VARIABLE OF GENETIC ALGORITHM

Max Iteration	Population	%Crossover	%Mutation	Selected Variable	RMSE
20	20	0.5	0.3	[,3,6,8,13,15,16,18]	9.870
20	20	0.7	0.3	[,2,3,4,8,9,10,11,17]	7.990
20	20	0.9	0.3	[3,4,5,7,10,13,4,16,18]	11.345
20	20	0.5	0.5	[1,2,3,6,7,8,9,15,16]	8.234
20	20	0.7	0.5	[2,3,4,5,8,10,11,4,17,18]	7.651
20	20	0.9	0.5	[2,3,4,5,7,8,9,10,11,13,15,17,18]	8.382
20	20	0.5	0.7	[2,3,4,6,7,8,9,4,15,16,17]	7.936
20	20	0.7	0.7	[2,3,4,5,7,8,10,12,4,16,17]	7.643
20	20	0.9	0.7	[1,3,4,5,6,9,13,4,15,16,18]	9.037
20	50	0.5	0.3	[3,4,5,8,9,11]	14.477
20	50	0.7	0.3	[2,3,4,5,6,11,12,13,15,16,17]	8.857
20	50	0.9	0.3	[2,3,4,5,9,10,11,15,18]	7.603
20	50	0.5	0.5	[3,5,9,12,13,4,17,18]	10.948
20	50	0.7	0.5	[3,4,5,6,8,9,10,18]	9.988
20	50	0.9	0.5	[1,2,3,6,8,9,10,11,13,15,16,18]	8.183
20	50	0.5	0.7	[2,5,8,9,11,4,16,17]	11.978
20	50	0.7	0.7	[2,3,5,8,9,10,12,15]	8.050
20	50	0.9	0.7	[1,2,3,6,10,13,4,15,16,17,18]	8.298
20	80	0.5	0.3	[3,5,6,8,9,10,12,13,16,17,18]	10.012
20	80	0.7	0.3	[1,2,3,5,6,7,10,11,12,4,15,17,18]	8.058
20	80	0.9	0.3	[3,5,8,10,11,12,13,18]	8.549
20	80	0.5	0.5	[2,3,4,10,11,12,13,4,15,16]	8.167
20	80	0.7	0.5	[,2,3,4,6,8,12,13,15,18]	8.399
20	80	0.9	0.5	[1,2,4,5,6,10,11,12,13,4,15\]	7.966
20	80	0.5	0.7	[1,2,3,9,10,11,13,17,18]	8.190
20	80	0.7	0.7	[1,3,6,8,9,11,15,16]	8.316
20	80	0.9	0.7	[1,2,3,4,7,12,4,15,16]	7.864

Table IV shows that the best composition of variables achieved was using the variable numbers [2, 3, 4, 5, 9, 10, 11, 15, 18]. These variables are frequency, TX height, RX height, RX vertical angle from TX main beam, distance between buildings, barometric pressure, temperature, slope contour, and border to user distance (water). The parameters selected in the Genetic Algorithm showed quite a marked difference with the Univariate FS. The RX vertical angle from TX main beam and distance between buildings, which had a low correlation with output, were included in the selected parameter composition, as was the TX-RX distance parameter. This was because the GA method did not take into consideration the correlation level between the input and output variables but performed a combination search with the mutation and crossover between the variables.

3) Particle Swarm Optimization (PSO) feature selection: PSO is also a wrapper method, which searches for the best composition by evaluating every possibility for each candidate combination using machine learning, searching for the best accuracy based on the results of the evaluation. This research carried out a number of trials by altering the values of the PSO parameter settings. Particle number, weighting, and C1/C2 values were changed to obtain the best accuracy value from the selected variables. The number of particles was varied with the values of 40, 70, and 100, while the W and C1/C2 values were varied with the values of 0.2, 0.5, and 0.8.

TABLE V. CANDIDATE VARIABLE OF PARTICLE SWARM OPTIMIZATION

Max Iteration	Particle	W	C1/C2	Selected Variable	Score Error
100	40	0.2	0.2	[1,2,3,4,5,6,8,9,10,11,12,13,4,16,17]	2.433
100	40	0.5	0.2	[2,3,5,6,7,8,11,13,4,15,16,17,18]	2.423
100	40	0.8	0.2	[1,2,3,4,6,7,8,9,12,13,4,15,16,17,18]	2.415
100	40	0.2	0.5	[2,3,4,5,6,7,8,12,13,4,16,17,18]	2.421
100	40	0.5	0.5	[1,2,3,5,6,7,9,10,4,15,16,17,18]	2.419
100	40	0.8	0.5	[1,2,3,4,5,6,7,8,9,10,12,13,16,17,18]	2.416
100	40	0.2	0.8	[2,3,4,5,6,9,10,12,13,4,15,16,17,18]	2.417
100	40	0.5	0.8	[2,3,5,6,7,8,9,10,11,12,13,15,16,17, 18]	2.415
100	40	0.8	0.8	[2,3,4,6,7,9,10,11,12,13,4,15,16,17, 18]	2.411
100	70	0.2	0.2	[2,3,6,7,8,9,10,11,13,4,15,16,17,18]	2.425
100	70	0.5	0.2	[1,2,3,4,6,7,9,11,12,13,15,16,17,18]	2.419
100	70	0.8	0.2	[2,3,4,6,7,9,11,12,13,4,15,16,17,18]	2.414
100	70	0.2	0.5	[1,2,3,4,5,6,7,8,11,12,13,4,15,16,17 ,18]	2.419
100	70	0.5	0.5	[2,3,4,5,6,8,9,10,12,13,16,17,18]	2.419
100	70	0.8	0.5	[2,3,4,5,6,7,8,9,10,11,12,13,4,15,16 ,17,18]	2.411
100	70	0.2	0.8	[1,2,3,4,6,7,8,9,12,13,4,15,16,17,18]	2.415
100	70	0.5	0.8	[2,3,4,5,6,7,8,9,10,11,12,13,4,16,17 ,18]	2.411
100	70	0.8	0.8	[2,3,4,6,7,9,10,11,12,13,4,15,16,17, 18]	2.411
100	100	0.2	0.2	[1,2,3,6,7,9,11,12,4,15,16,17,18]	2.428
100	100	0.5	0.2	[2,3,4,5,6,9,12,13,4,15,16,17,18]	2.420
100	100	0.8	0.2	[1,2,3,5,6,7,8,9,10,12,13,4,15,16,17 ,18]	2.413
100	100	0.2	0.5	[1,2,3,4,6,7,9,4,15,16,17,18]	2.423
100	100	0.5	0.5	[1,2,3,4,5,6,7,8,9,11,12,13,4,16,17, 18]	2.414
100	100	0.8	0.5	[1,2,3,4,6,7,8,9,10,11,13,4,15,16,17 ,18]	2.417
100	100	0.2	0.8	[2,3,4,5,6,9,10,12,13,4,15,16,17,18]	2.417
100	100	0.5	0.8	[1,2,3,4,5,6,7,8,9,10,11,12,13,4,15, 16,17,18]	2.411
100	100	0.8	0.8	[2,3,4,5,6,7,8,9,11,12,13,4,15,16,17 ,18]	2.413

Table V shows that the best composition of variables was the composition of variables [2, 3, 4, 6, 7, 9, 10, 11, 12, 13,

14, 15, 16, 17, 18]. The variables eliminated from the selected variables were TX-RX distance, RX vertical angle to mainbeam, and building height. This was the same as the Univariate results, where the parameters of RX vertical angle to mainbeam and building height, which had a low correlation, were not included in the selected variables. However, what was significantly different was the parameter of TX-RX distance, which had a sufficiently high level of correlation but was not included in the selected variables in PSO. As in the case of GA, PSO did not take into account the level of correlation between the input and output variables, but carried out a combination search with a particular method. In PSO, the search uses a swarm technique, which is a search based on the history of the best values, whether Pbest or Gbest.

C. Machine Learning Evaluation

In this research, four machine learning models were used to improve path loss prediction. The results of the evaluation parameters using four machine learning model are presented.

Table VI shows the evaluation of combination parameters to the feature selection methods with machine learning models. These results indicate that the best feature selection in DNN modeling is the Univariate method with an RMSE value of 4.49. These results are also shown in the KNN Regressor and Random Forest modeling where the smallest RMSE value uses the Univariate feature selection method where the RMSE values are 3.75 and 1.52 respectively, while the feature selection method using the GA method has the lowest accuracy rate for the use of the three types of algorithms. the. The SVR modeling shows different results, namely the best feature selection using GA, while the selection feature with the largest RMSE uses Univariate.

Table VII shows the best combination of feature selection methodology and machine learning. In the improvement of the Cost-Hatta model using Univariate-Random Forest, it has the smallest level of accuracy, namely the RMSE value of 1.52, MAE of 1.09 and MAPE of 14.08. The second accuracy value is model improvement using Univariate-KNN with a RMSE value of 3.75, an MAE value of 2.76 and a MAPE value of 35.75. On the third device, using univariate-DNN with an RMSE value of 4.49, an MAE value of 3.31 and a MAPE value of 308.8. While the worst value of accuracy improvement is by using GA-SVR. The value of the level of accuracy in the model is RMSE of 17.09, MAE value of 13.9 and MAPE value of 592.29.

The results of the Cost-Hatta model improvement using a machine learning approach can be seen in Fig. 6. The graph shows that the increasing accuracy using univariate-RF was the highest in RMSE accuracy, which is around 94.12%, followed by Univariate-KNN Regression and Univariate-DNN where the increase values are 85.48% and 82.67%, respectively. The lowest RMSE accuracy increase value is in the GA-SVR combination, with an accuracy increase of 33.84% from the Cost-Hatta RMSE value of 25,832 to 1.52.

TABLE VI. EVALUATION PERFORMANCE OF FEATURE SELECTION – MACHINE LEARNING MODEL

Model	FS	RMSE	MAE	MAPE
DNN	GA	5.58	4.15	720.57
	PSO	5.18	3.86	705.96
	Univariate	4.49	3.31	308.80
KNN Regressor	GA	4.61	3.24	148.09
	PSO	3.88	2.80	89.56
	Univariate	3.75	2.76	35.75
Random Forest Regressor	GA	1.76	1.22	105.79
	PSO	1.58	1.12	36.54
	Univariate	1.52	1.09	14.08
SV Regressor	GA	17.09	13.90	592.29
	PSO	18.08	14.65	634.28
	Univariate	18.21	14.74	252.81

TABLE VII. COMPARISON OF COST-HATTA MODEL WITH MACHINE LEARNING MODIFICATION MODEL

Model	**Evaluation**		
	RMSE	**MAE**	**MAPE**
Cost-Hatta	25.832	21.319	15.375
Cost-Hatta - Univ-DNN	4.49	3.31	308.8
Cost-Hatta-Univ-KNN	3.75	2.76	35.75
Cost-Hatta-Univ-RF	1.52	1.09	14.08
Cost-Hatta-GA-SVR	17.09	13.9	592.29

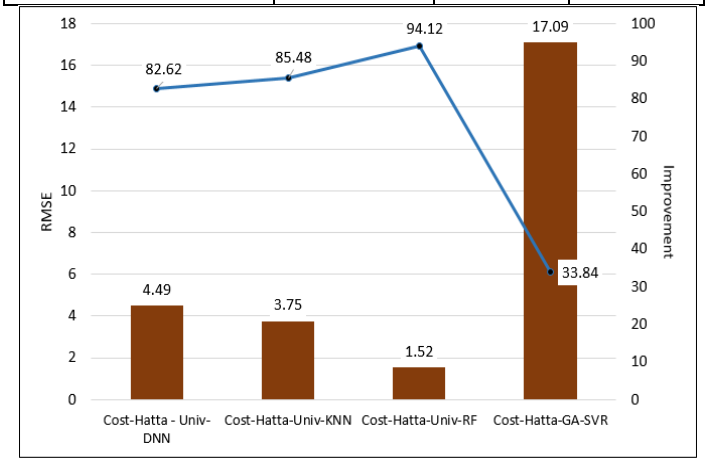

Fig. 6. Improvement Percentage of Machine Learning Model Approachment.

V. CONCLUSIONS

The determination of the features in the modeling will determine the accuracy of the path loss prediction. The condition of different signal propagation area causes the complexity of the various parameters needed in predicting path loss modeling, especially in mixed land-water areas. System parameters and environmental parameters have an influence on the path loss value. The feature selection method approach is needed to choose the best combination of

parameters in the construction of the prediction model. Improvements to the Cost-Hatta model with the feature-selection and machine learning approach resulted in a significant improvement in accuracy. The combination of the Univariate-RF model is the best combination with an increase in accuracy of 94.12% from the previous RMSE Cost-Hatta value. This indicates that the proposed model framework is highly suitable to be used in a mixed land-water area.

VI. Future Work

In the next research, some suggestion to increase path loss prediction accuracy especially in mixed land water area:

1) Hyper-parameter optimization of machine learning models can be carried out. Metaheuristic methods such as Genetic Algorithm and Particle Swarm Optimization can be used in determining the composition of hyper-parameters to get the best accuracy value.

2) Expand measurement data to get a more varied sample value, especially for weather parameters.

Acknowledgment

The first author is a doctoral student at the Faculty of Engineering Science, Universitas Sriwijaya. The authors would like to thank Universitas Sriwijaya for their support in carrying out this research.

References

[1] Park, D. K. Tettey, and H.-S. Jo, "Artificial Neural Network Modeling for Path Loss Prediction in Urban Environments," J. LATEX Cl. FILES, vol. 14, no. 8, pp. 9–13, 2019, [Online]. Available: http://arxiv.org/abs/1904.02383.

[2] B. Alfaresi, T. Barlian, F. Ardianto, and M. Hurairah, "Path Loss Propagation Evaluation and Modelling based ECC-Model in Lowland Area on 1800 MHz Frequency," J. Robot. Control, vol. 1, no. 5, pp. 167–172, 2020, doi: 10.18196/jrc.1534.

[3] I. Oluwafemi and O. Femi-Jemilohun, "Suburban area path loss propagation prediction and optimization at 900 and 1800 MHz," J. Eng. Appl. Sci., vol. 13, no. 9, pp. 2521–2529, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85049520402&origin=inward.

[4] A. O. A, T. O. A, M. O. S, and A. J. A, "Experimental Study of Variation of Path Loss with Respect to Heights at GSM Frequency Band," Int. J. Sci. Res. Sci. Eng. Technol., vol. 3, no. 3, pp. 347–351, 2016.

[5] Alor MO, "Efficient Pathloss Model for determining Mobile Radio Link Design," Int. J. Sci. Res. Sci. Eng. Technol., vol. 3, no. 3, pp. 270–276, 2015.

[6] 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, 2017, doi: 10.1109/TAP.2017.2705112.

[7] 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, 2019, doi: 10.3390/app9091908.

[8] D. A. V. Sreevardhan Cheerla, K. Sindhuja, Ch. Indra Kiran, "Analysis of different path loss models in urban suburban and rural environment," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 7, pp. 2972–2976, 2020, doi: 10.30534/ijeter/2020/14872020.

[9] O. Shoewu, L. A. Akinyemi, and L. Oborkhale, "Modelling Path Loss in Mobile Communication 4G Network System for Dryland and Wetland Terrains," in Southern Africa Telecommunication Networks and Applications Conference (SATNAC), 2019, pp. 44–49.

[10] J. Isabona and V. M. Srivastava, "Hybrid neural network approach for predicting signal propagation loss in urban microcells," IEEE Reg. 10 Humanit. Technol. Conf. 2016, R10-HTC 2016 - Proc., 2017, doi: 10.1109/R10-HTC.2016.7906853.

[11] K. Saito, Y. Jin, C. Kang, J. Takada, and J.-S. Leu, "Two-step path loss prediction by artificial neural network for wireless service area planning," IEICE Commun. Express, vol. 8, no. 12, pp. 611–616, 2019, doi: 10.1587/comex.2019gcl0038.

[12] H. S. Jo, C. Park, E. Lee, H. K. Choi, and J. Park, "Path loss prediction based on machine learning techniques: Principal component analysis, artificial neural network and gaussian process," Sensors (Switzerland), vol. 20, no. 7, pp. 1–23, 2020, doi: 10.3390/s20071927.

[13] B. J. Cavalcanti, G. A. Cavalcante, L. M. De Mendonça, G. M. Cantanhede, M. M. M. De Oliveira, and A. G. D'Assunção, "A hybrid path loss prediction model based on artificial neural networks using empirical models for LTE and LTE-A at 800 MHz and 2600 MHz," J. Microwaves, Optoelectron. Electromagn. Appl., vol. 16, no. 3, pp. 708–722, 2017, doi: 10.1590/2179-10742017v16i3925.

[14] L. Wu et al., "Artificial Neural Network Based Path Loss Prediction for Wireless Communication Network," IEEE Access, vol. 8, pp. 199523–199538, 2020, doi: 10.1109/access.2020.3035209.

[15] J. E. Ofure, O. D. Oyedum, M. O. Ajewole, and A. M. Aibinu, "Comparative analysis of basic models and artificial neural network based model for path loss prediction," Prog. Electromagn. Res. M, vol. 61, no. October, pp. 133–146, 2017, doi: 10.2528/PIERM17060601.

[16] C. Oroza et al., "A Machine-Learning Based Connectivity Model for Deployments To cite this version : A Machine-Learning Based Connectivity Model for Complex Terrain Large-Scale Low-Power Wireless Deployments," IEEE Trans. Cogn. Commun. Netw., vol. 3, no. 4, pp. 576–584, 2017, doi: 10.1109/TCCN.2017.2741468.

[17] 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, 2019, doi: 10.1109/ACCESS.2019.2950634.

[18] S. Duangsuwan, "Comparison of path loss prediction models for UAV and IoT air-to-ground communication system in rural precision farming environment," J. Commun., vol. 16, no. 2, pp. 60–66, 2021, doi: 10.12720/jcm.16.2.60-66.

[19] G. Yang, Y. Zhang, Z. He, J. Wen, Z. Ji, and Y. Li, "Machine-learning-based prediction methods for path loss and delay spread in air-to-ground millimetre-wave channels," IET Microwaves, Antennas Propag., vol. 13, no. 8, pp. 1113–1121, 2019, doi: 10.1049/iet-map.2018.6187.

[20] 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," Wirel. Commun. Mob. Comput., vol. 2018, 2018, doi: 10.1155/2018/8489326.

[21] ALLAN DOS S. BRAGA et al., "Radio Propagation Models Based on Machine Learning Using Geometric Parameters for a Mixed City-River Path," IEEE Access, vol. 8, pp. 146395–146407, 2020, doi: 10.1109/ACCESS.2020.3012661.

[22] T. Zhang, S. Liu, W. Xiang, L. Xu, K. Qin, and X. Yan, "A real-time channel prediction model based on neural networks for dedicated short-range communications," Sensors (Switzerland), vol. 19, no. 16, 2019, doi: 10.3390/s19163541.

[23] S. I. Popoola, S. Misra, and A. A. Atayero, "Outdoor Path Loss Predictions Based on Extreme Learning Machine," Wirel. Pers. Commun., vol. 99, no. 1, pp. 441–460, 2017, doi: 10.1007/s11277-017-5119-x.

[24] S. Ojo, "Radial basis function neural network path loss prediction model for LTE networks in multitransmitter signal propagation environments," Int. J. Commun. Syst., vol. 34, no. 3, 2021, doi: 10.1002/dac.4680.

[25] A. Tahat, T. Edwan, H. Al-Sawwaf, J. Al-Baw, and M. Amayreh, "Simplistic Machine Learning-Based Air-to-Ground Path Loss Modeling in an Urban Environment," 2020 5th Int. Conf. Fog Mob. Edge Comput. FMEC 2020, pp. 158–163, 2020, doi: 10.1109/FMEC49853.2020.9144965.

[26] E. J. Ofure, O. D. Oyedum, M. O. Ajewole, and A. M. Aibinu, "Artificial Neural Network model for the determination of GSM Rxlevel from atmospheric parameters," Eng. Sci. Technol. an Int. J., vol. 20, no. 2, pp. 795–804, 2016, doi: 10.1016/j.jestch.2016.11.002.

[27] S. I. Popoola, E. Adetiba, A. A. Atayero, N. Faruk, and C. T. Calafate, "Optimal model for path loss predictions using feed-forward neural networks," Cogent Eng., vol. 5, no. 1, 2018, doi: 10.1080/23311916.2018.1444345.

[28] H. Singh, S. Gupta, C. Dhawan, and A. Mishra, "Path Loss Prediction in Smart Campus Environment: Machine Learning-based Approaches," IEEE Veh. Technol. Conf., vol. 2020-May, 2020, doi: 10.1109/VTC2020-Spring48590.2020.9129444.

[29] C. Emeruwa and P. . Iwuji, "Determination Of A Pathloss Model For Long Term Evolution (Lte) In Yenagoa," Int. J. Eng. Sci., vol. 7, no. 10, pp. 38–44, 2018, doi: 10.9790/1813-0710033844.

[30] E. R. Abboud, "Propagation Model For the 900 MHz Almadar Aljadid Mobile Network at Tripoli Area Using Linear Regression Method," in The Proceedings of Second International Conference on Electrical and Electronics Engineering, Clean Energy and Green Computing, 2015, pp. 5–11.

[31] A. Sanmorino, Ermatita, Samsuryadi, and D. P. Rini, "Building Research Productivity Framework in Higher Education Institution," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, pp. 184–191, 2021, doi: 10.14569/IJACSA.2021.0120620.

[32] A. Bhuvaneshwari, "Path loss model optimization using stochastic hybrid genetic algorithm," Int. J. Eng. Technol., vol. 7, no. 4, pp. 464–469, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85082355283&origin=inward.

[33] F. Moslehi and A. Haeri, "A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection," J. Ambient Intell. Humaniz. Comput., vol. 11, no. 3, pp. 1105–1127, 2020, doi: 10.1007/s12652-019-01364-5.

[34] Y. C. Hsieh, P. J. Lee, and P. S. You, "Immune-based evolutionary algorithm for determining the optimal sequence of multiple disinfection operations," Sci. Iran., vol. 26, no. 2 C, pp. 959–974, 2019, doi: 10.24200/sci.2018.20324.

[35] A. Brezulianu, L. Fira, and M. Fira, "A genetic algorithm approach for scheduling of resources in well-services companies," Int. J. Adv. Res. Artif. Intell., vol. 1, no. 5, pp. 1–6, 2012, doi: 10.14569/ijarai.2012.010501.

[36] S. Rukhaiyar, M. N. Alam, and N. K. Samadhiya, "A PSO-ANN hybrid model for predicting factor of safety of slope," Int. J. Geotech. Eng., vol. 12, no. 6, pp. 556–566, 2018, doi: 10.1080/19386362.2017.1305652.

[37] B. A. A. Yousef, H. Rezk, M. A. Abdelkareem, A. G. Olabi, and A. M. Nassef, "Fuzzy modeling and particle swarm optimization for determining the optimal operating parameters to enhance the bio-methanol production from sugar cane bagasse," Int. J. Energy Res., vol. 44, no. 11, pp. 8964–8973, 2020, doi: 10.1002/er.5605.

[38] S. Karkheiran, A. Kabiri-Samani, M. Zekri, and H. M. Azamathulla, "Scour at bridge piers in uniform and armored beds under steady and unsteady flow conditions using ANN-APSO and ANN-GA algorithms," ISH J. Hydraul. Eng., vol. 00, no. 00, pp. 1–9, 2019, doi: 10.1080/09715010.2019.1617796.

[39] T. Si and R. Dutta, Partial Opposition-Based Particle Swarm Optimizer in Artificial Neural Network Training for Medical Data Classification, vol. 18, no. 5. 2019.

[40] M. Khari, D. Jahed Armaghani, and A. Dehghanbanadaki, "Prediction of Lateral Deflection of Small-Scale Piles Using Hybrid PSO–ANN Model," Arab. J. Sci. Eng., vol. 45, no. 5, pp. 3499–3509, 2020, doi: 10.1007/s13369-019-04134-9.