Automatic detection and prediction of signal strength degradation in urban areas using data-driven machine learning

Ibrahim El Moudden[1], Youssef Benmessaoud[2], Abdellah Chentouf[3], Loubna Cherrat[4], Ech-Charrat Mohammed Rida[5], Mostafa Ezziyyani[2]

[1]Mathematics and Applications Laboratory, Department of Computer Science, Faculty of Sciences and Technologies, Abdelmalek Essaadi University, Tangier, Morocco
[2]Department of Computer Science, Faculty of Sciences and Technologies, Abdelmalek Essaadi University, Tangier, Morocco
[3]Department of Physics, Faculty of Sciences and Technologies, Abdelmalek Essaadi University, Tangier, Morocco
[4]National School of Commerce and Management, Abdelmalek Essaadi University, Tangier, Morocco
[5]Department of Computer Science, National School of Applied Sciences of Tetouan, Abdelmalek Essaadi University, Tetouan, Morocco

Article Info

Article history:

Received Nov 14, 2023
Revised Mar 13, 2024
Accepted Apr 6, 2024

Keywords:

Coverage
Decision tree
Gradient-boosted
Machine learning
Signal strength

ABSTRACT

This study introduces an innovative approach to predicting signal strength degradation in urban areas by leveraging a data-driven machine learning methodology. Focusing on the issue of signal variation within diverse urban infrastructures, we introduce EzziSignal, a real-time signal data collection method through a mobile application. Traditional methods, such as manual drive tests, are labor-intensive and costly. In response, our research offers a novel and efficient alternative utilizing data mining and machine learning techniques. The study area was subdivided into multiple zones and sectors, independent of existing broadcast station locations. A custom mobile application was developed to systematically collect signal strength and location data across these zones, harnessing the collective power of over ten users' smartphones. This extensive dataset was then analyzed using the gradient-boosted algorithm, a sophisticated machine-learning technique. Its accuracy reached approximately 96.96%. The model exhibited promising results in predicting signal strength degradation, providing valuable insights into the dynamics of urban signal behavior. This research carries significant practical implications for telecommunication companies, offering an intelligent framework to optimize resource allocation and base station installation. By considering the deterioration of signal, strength determines the relationship between infrastructure and base station installation regarding the quality of networking. Ultimately, this optimization contributes to enhanced network coverage and improved service quality.

Corresponding Author:

Ibrahim El Moudden
Mathematics and Applications Laboratory, Department of Computer Science
Faculty of Sciences and Technologies, Abdelmalek Essaadi University
Old Airport Road, Km 10, Ziaten. BP: 416. Tangier, Morocco
Email: ibrahim.elmoudden@etu.uae.ac.ma

1. INTRODUCTION

Mobile network operators analyze signal quality and strength to optimize their services, this study investigated the effects of building altitude on coverage quality. We have studied the problem detection of possible signal strength deterioration in different areas around the city. This problem is seen as a challenge for the operator at all times, due to customer dissatisfaction with signal deterioration in different infrastructures.

In the literature review, most previous research has focused on improving the construction of network coverage from sparse received signal strength indicator (RSSI) data collected by test drivers in different locations as done in [1]-[4]. Other studies predicting path loss characteristics in areas of different sizes, heights of buildings, and surrounding environments (mountains) [5]-[7] have followed the approach. In Zhang and Shu [8] it was founded on the correlation between coverage quality and the quality of the transmission link. To accurately estimate the link quality, a gradient decision tree (GBDT) based link quality estimator is proposed. Link quality estimation is the main problem in ensuring the reliability of data transmission and the performance of the upper-layer network protocol. Instead of using the test driver, Wayessa [9] employed a universal mobile telecommunications system (UMTS) network coverage hole detection using a decision tree classifier. This approach addresses the concept of detecting network coverage holes along with the associated challenges posed by conventional data collection methods. It also provides the basic information to help acquire the root cause of such challenges so that the optimization team can trigger the optimization process on time. Minimization of drive test (MDT) data was used to build a model that classifies different coverage problem scenarios such as "poor coverage and poor quality", "poor coverage but good quality", and "good coverage but poor quality". Map construction involves collecting information about signal coverage in sparse locations, which can be done conventionally by measurement methods such as manual drive tests. The networked construction of a map of the indoor radio environment helps the operator define the problem area. Rufaida [10] applied gradient boosting algorithms (specifically XGBoost and light gradient boosting machine) to formulate radio environment map (REM) coverage maps. The performance of these algorithms was assessed through experimental evaluation, which involved the creation of heat maps depicting the coverage of base stations. They received reference signal power, quality, and signal-to-noise ratio, under different configuration parameters. The outcomes affirm the superior efficacy of both XGBoost and light gradient boosting machine compared to existing baseline methods k-nearest neighbor and support vector machine.

While earlier studies have explored the impact of path loss and the creation of a radio environment map, it is crucial to acknowledge their limitations, including the need for repetitive processes, organizational requirements, and considerable time and financial investments. Notably, they have not explicitly addressed the influence of signal strength deterioration as a predictive factor and its association with the radio environment. The approach proposed in this paper introduces a machine learning method, signal strength degradation prediction (SSDP), leveraging historical data gathered from diverse locations within the city.

The remainder of this paper is organized as follows. In section 2, we detail the method and materials used in our study, including the initial phases of data processing, starting with data collection, the meticulous process of data cleaning, and the construction of the data structure description. Subsequently, we presented the machine learning algorithms (decision tree algorithm and gradient-boosting trees) and proceeded to train the model to classify the area and predict signal intensity degradation. In section 3, we conduct a comprehensive analysis of results, comparing and analyzing two methods and interpreting the strength of signal degradation prediction (PSSD). Section 4 succinctly summarizes our conclusions, highlighting key results and their implications for our research goals.

2. MATERIALS AND METHODS

This section initiates with a discussion on the data processing phase, covering data collection and cleaning procedures. Subsequently, the data structure is described. In the subsequent subsection, we introduce the machine learning algorithms, encompassing both the decision tree and gradient-boosted tree (GBT) models. Finally, the model learning process is elaborated upon, providing a comprehensive overview of the training procedure.

2.1. Data processing

In the data processing phase, information was collected from the cell phone application, and the dataset was carefully cleaned to correct outliers, inconsistencies, and missing values. This rigorous process ensures the integrity and accuracy of our dataset, laying a solid foundation for subsequent analyses. The resulting high-quality data reinforce the reliability of our study results.

2.1.1. Data collection

The challenge lies in acquiring the test driver from the telecom operator for directly collecting data from base stations. RSSI is accessible on every cell phone, whether it is an iPhone, Android device, or any other cellular-connected device. The visual depiction of cellular signal strength is primarily conveyed through signal bars. Various phones utilize distinct decibel scales, often with negative values as shown in Figure 1. Those values were retrieved from [11] typically expressed in logarithmic units like decibels-milliwatts (dBm) or decibels (dB), RSSI measures the ratio of received power to a reference power level.

RSSI values usually range from -50 dBm to -120 dBm, where -50 dBm signifies an excellent signal and -120 dBm indicates a very weak signal strength.

We designed a mobile application called EzziSignal-a real-time cell phone application for collecting signal strength information (RSSI), as illustrated in Figure 2. This application encompasses four key parameters: the user's latitude, longitude, signal strength, and time. This application is deployed on diverse mobile phones placed within moving cars operating at a consistent speed.

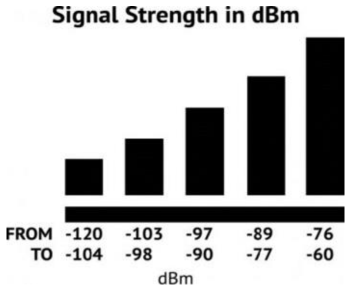
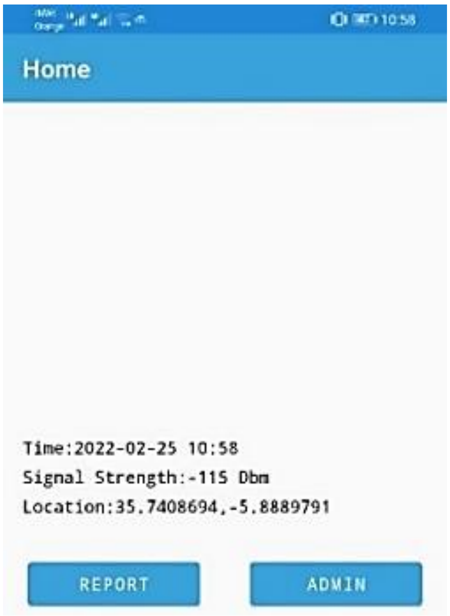


Figure 1. Signal strength bars in dB


Home

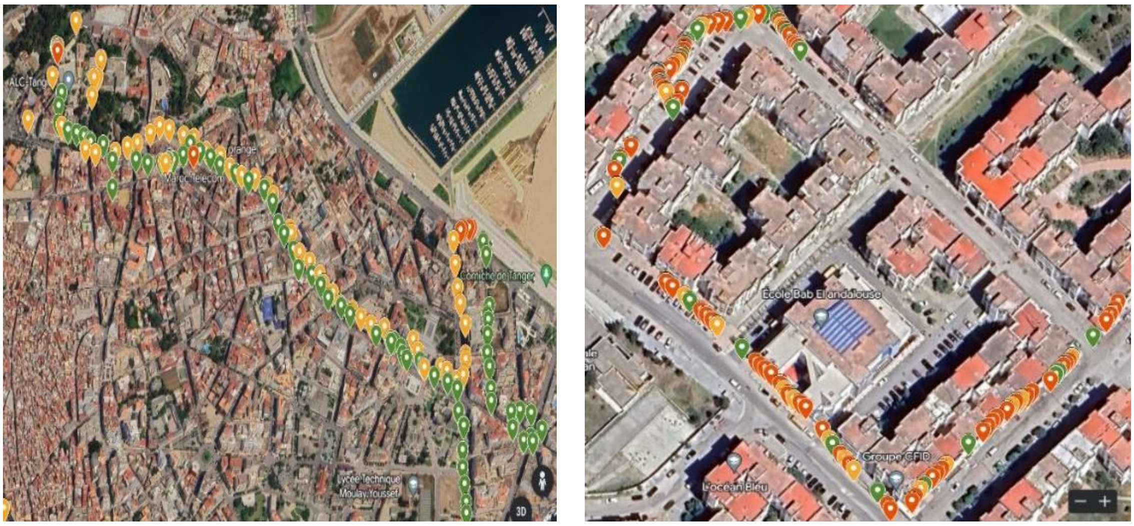
Time:2022-02-25 10:58
Signal Strength:-115 Dbm
Location:35.7408694,-5.8889791

Figure 2. EzziSignal-real time cell phone application for signal data collection

In Tangier, Morocco, our data collection application is strategically deployed across three distinct geographic zones, each with unique infrastructure characteristics. These zones include mid-rise buildings with a height of 13 meters, open spaces, and high-rise buildings with a height of 22 meters (as shown in Figure 3). Figure 3(a) highlights the high-rise buildings, providing a visual representation of the specific environment and the corresponding changes in signal strength. On the other hand, Figure 3(b) shows buildings at medium heights, highlighting the complexities of signal strength fluctuations within this specific area. We present an example of signal strength data collection categorized by colored nodes representing signal strength intervals in Table 1.

Table 1. Signal strength levels

Signal strength	Weak signals	Medium signals	High signals
interval	-91 to -120dBm	-76 to -90dBm	-50 to -75dBm
Node color	Red	Orange	Green



(a) (b)

Figure 3. Data collection areas, (a) high-altitude of buildings (22 meters) and (b) medium altitude of buildings (13 meters)

2.1.2. Data cleaning

Before moving on to the data learning and analysis phase, we devote this section to data pre-processing, a crucial step in the data analysis process. This step aims to prepare the raw data for analysis by eliminating errors, dealing with missing values and outliers, normalizing the data, and performing other necessary transformations. With this in mind, in the first phase, we used the median method to identify outliers and assess data dispersion. Before that, we had to keep a type for each value, classify them, and convert them into usable positive values in the next steps by applying the methods used by [12], [13].

- Step 1: Calculate the median: In this first step, we will calculate the median of the data set. We find the median equal -82 dBm.
- Step 2: In this step, we change the sign of the data numbers and calculate the absolute deviation from the median for each data point (Table 2).

Table 2. Absolute deviation

Signal strength	V1=58	V2=60	V3=62	V4=63	V5=64	V5=65	…..	V46=110
Absolute deviation	|Mx – V1|=21	|Mx – V2|=19	|Mx – V3|=17	|Mx – V4|=16	|Mx – V5|=15	|Mx – V6|=14	…….	|Mx – Vn|=31

- Step 3: Identify outliers after calculating the threshold to determine which data are considered outliers. The interquartile range (IQR) is calculated by subtracting the first quartile (Q1) and the third quartile (Q3). Q1 is the median of the data before the overall median and Q3 is the median of the data after the overall median. We calculate the position of Q1 and Q3 respectively using (1) and (2).

$$\text{Q1 position} = \frac{(n+1)}{4} \tag{1}$$

$$\text{Q3 position} = \frac{3(n+1)}{4} \tag{2}$$

Where n is the number of data points. Table 3 presents the position in data and the value of the median of Q1 and Q3. Therefore $\text{IQR} = \text{Q3} - \text{Q1} = 94 - 70 = 16$.

Table 3. Q1 and Q3 median

	Position in data	Value of median
Q1	11	70
Q3	35	94

– Step 4: Outlier replacement outliers are data that do not belong to the interval:

$$[\mathrm{Q1} - 1.5\mathrm{xIQR}, \mathrm{Q3} + 1.5\mathrm{xIQR}] = [70 - 1.5\mathrm{x}16, 94 + 1.5\mathrm{x}16] = [46, 118] = [-118\mathrm{Dbm}, -46\mathrm{Dbm}]$$

Furthermore, in Figure 4, we have replaced the outliers of this interval with the median value. Our data values range between -58 dBm and -110 dBm.

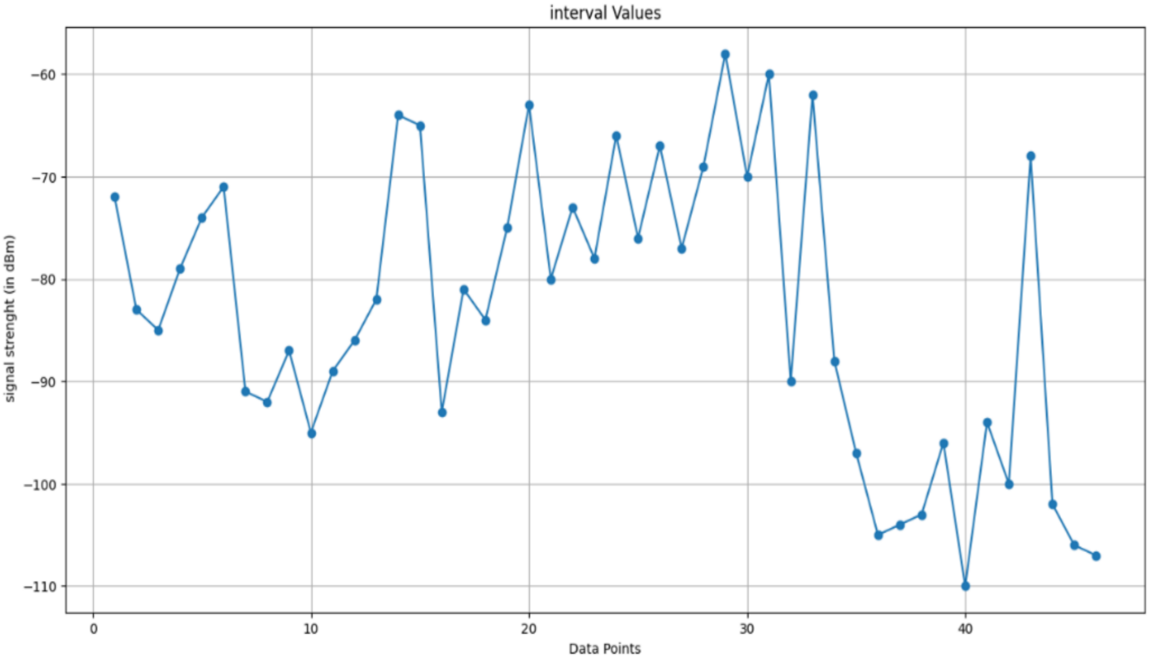


Figure 4. Data values limits

The second phase of the data cleaning process involves pinpointing duplicate values within the dataset that correspond to the same geolocalization. By isolating these duplicates, it becomes possible to streamline the dataset, ensuring accuracy and consistency in the information linked to specific geographic points.

2.1.3. Data structure description

Our dataset comprises 392 input lines, with each line containing twelve nodes featuring signal strength and location values. The signal strength is represented as a negative value in decibels, ranging between -50 decibels and -110 decibels. The data structure is organized based on location and the time of data collection. Additionally, each set of twelve lines includes a location category in the output data. The output from these lines falls into several categories, contingent upon the area from which the data was collected.

The data categories we have collected primarily revolve around three key areas: medium-altitude buildings, open space, and high-altitude buildings. Each distinct category is assigned a specific letter for classification purposes. The process of selecting these categories necessitated thorough on-site visits to meticulously document all building-related details within each area. Table 4 provides a representation of the sample data that we have gathered.

2.2. Machine learning algorithms

For the classification and prediction model. Two, machine learning-based models were trained using decision tree (DT) and gradient boosted tree (GBT). The choice of these machine learning classifiers was driven by the fact that the gradient boosted tree classifier is known to use ensemble Techniques, which have proven to increase the accuracy and performance of models.

Table 4. Sample of data structure

node 1	node 2	node 3	node 4	node 5	node 6	node 7	node 8	node 9	node 10	node 11	node 12	Class
-100 Dbm	-69 Dbm	-69 Dbm	-80 Dbm	-69 Dbm	-69 Dbm	-69 Dbm	-88 Dbm	-69 Dbm	-103 Dbm	-103 Dbm	-69 Dbm	Medium altitude of building (13m)
-88 Dbm	-103 Dbm	-69 Dbm	-88 Dbm	-106 Dbm	-103 Dbm	-106 Dbm	-106 Dbm	-69 Dbm	-69 Dbm	-80 Dbm	-69 Dbm	Medium altitude of building(13m)
-69 Dbm	-69 Dbm	-88 Dbm	-69 Dbm	-103 Dbm	-108 Dbm	-69 Dbm	-88 Dbm	-103 Dbm	-69 Dbm	-88 Dbm	-101 Dbm	Medium altitude of building(13m)
-87 Dbm	-91 Dbm	-85 Dbm	-85 Dbm	-85 Dbm	-81 Dbm	-81 Dbm	-81 Dbm	-81 Dbm	-81 Dbm	-81 Dbm	-81 Dbm	Open Space
-76 Dbm	-69 Dbm	-73 Dbm	-73 Dbm	-69 Dbm	-69 Dbm	-80 Dbm	-69 Dbm	-69 Dbm	-69 Dbm	-77 Dbm	-69 Dbm	Open Space
-69 Dbm	-78 Dbm	-69 Dbm	-78 Dbm	-78 Dbm	-69 Dbm	-69 Dbm	-80 Dbm	-69 Dbm	-69 Dbm	-69 Dbm	-78 Dbm	Open Space
-71 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	High-altitude of buildings(22m)
-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	High-altitude of buildings(22m)
-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-71 Dbm	-74 Dbm	-91 Dbm	-71 Dbm	-92 Dbm	High-altitude of buildings(22m)
-87 Dbm	-71 Dbm	-95 Dbm	-71 Dbm	-89 Dbm	-71 Dbm	-89 Dbm	-71 Dbm	-89 Dbm	-71 Dbm	-89 Dbm	-71 Dbm	High-altitude of buildings

2.2.1. Decision trees model

We present an adaptation of the classification and regression trees (CART) algorithm designed to address signal strength degradation in data. CART is capable of constructing both classification and regression trees through a process based on binary attribute division as in [14]. It is also based on Hunt's algorithm and can be implemented serially. In the context of a CART decision tree, the Gini index is a measure used to assess the impurity of a node in the tree. It is commonly used as a criterion for deciding on the division of nodes when building a decision tree. In decision trees, nodes represent subsets of data based on certain characteristics. The Gini index measures the impurity of a node, indicating how mixed the target labels are within that node. A node with low impurity means that it contains mainly instances of a particular class, while a node with high impurity contains a mixture of different classes. The Gini index is widely used and often preferred due to its computational efficiency and ability to work well in practice. We can formulate it as (3).

$$\mathrm{Gini} = 1 - \Sigma\,(\mathrm{Pi})^2 \tag{3}$$

2.2.2. Gradient boosted trees mode

We are dealing with structured data and looking for a high-performance algorithm, XGBoost is generally the preferred choice for such scenarios. We used XGBoost an optimized and enhanced version of GBT with additional features and regularizations, making it more robust, accurate, and efficient for practical applications [15]. In XGBoost, the learning rate (also known as the "eta" parameter) is a crucial hyperparameter that controls the step size at each iteration when fitting the gradient boosting model as shown [16]-[18]. This technique assesses the impact of each weak learner (tree) on the final ensemble model. A lower learning rate induces slower convergence, potentially yielding predictions that are more precise. Conversely, a higher learning rate may expedite convergence but heighten the risk of overfitting.

In Figure 5, we illustrate the gradient-boosted method. Initially, we use residuals from the initial prediction as target values to construct the first tree. In the subsequent step, we multiply the tree results by the learning rate and add them to the previous prediction. For the third step, residuals after combining the initial prediction with the first tree (scaled by the learning rate) serve as target values for the second tree. This process is iterated until the specified number of trees is reached or no further improvement is attainable. The ultimate prediction materializes as the sum of predictions from individual decision trees, each weighted by the learning rate (4).

$$\mathrm{Y\,(pred)} = \mathrm{y1} + (\mathrm{eta} * \mathrm{r1}) + (\mathrm{eta} * \mathrm{r2}) + \ldots + (\mathrm{eta} * \mathrm{rN}) \tag{4}$$

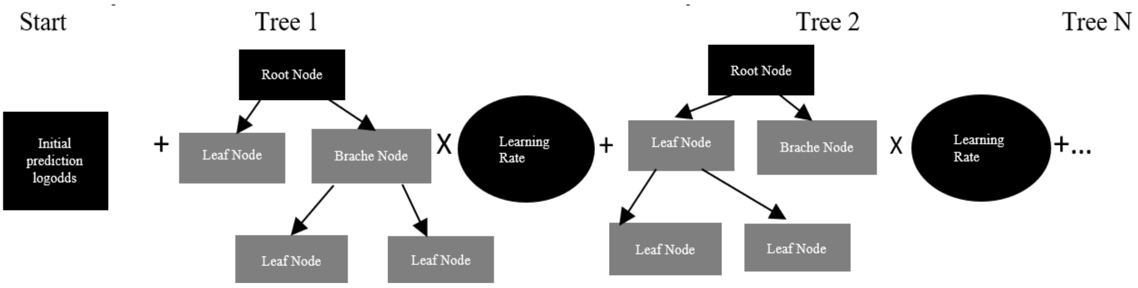


Figure 5. Gradient boosted method

2.3. Model training

We utilized KNIME software to simulate the system. Data in KNIME is arranged in a table format with a specified number of columns. The analysis begins by collecting and pre-processing the data. We drag and drop the "Column Filter" node from the "Node Repository" panel into the workflow editor panel. We connect the "Column Filter" node to the previous node (in our workflow, the "File Reader" "Adult Data Set" node). This node allows you to filter the columns in the input table and pass only the remaining columns to the output table. In the dialog box, columns can be moved between the Include and Exclude lists. The input table is split into two partitions (i.e., row-wise), the training and testing data. The two partitions are available at the two output ports. The partition has two options: “Absolute” and “Relative” The relative option permits specification of the percentage of rows from the input table in the first partition, ranging from zero to 100. In cases of the absolute value option, if the specified number of rows exceeds the available rows, all rows are placed in the first table, leaving the second table empty. In this study, we used the liner-sampling mode. This mode always includes the first and the last row and selects the remaining rows linearly over the whole table (e.g. every third row). This is useful to downsample a sorted column while maintaining minimum and maximum values.

In our dataset partitioning strategy, we allocated 78% of the data for training and reserved the remaining 22% for testing, facilitating a robust evaluation of DT and GBT (Figure 6). The decision tree learner, used for DT classification, exclusively trained on the 78% subset, employing the Gini index for split calculation (Figure 6(a)). For GBT, the gradient boosted trees learner utilized the same 78% partition, offering advanced options like XGBOOST and surrogate (Figure 6(b)). This consistent partitioning approach ensures a fair comparison between DT and GBT models, allowing for a meaningful assessment of their classification performance. Both learners connected with the column filter node and finally, we have a node of scorers who have accuracy statistics content accuracy and Cohen’s kappa.

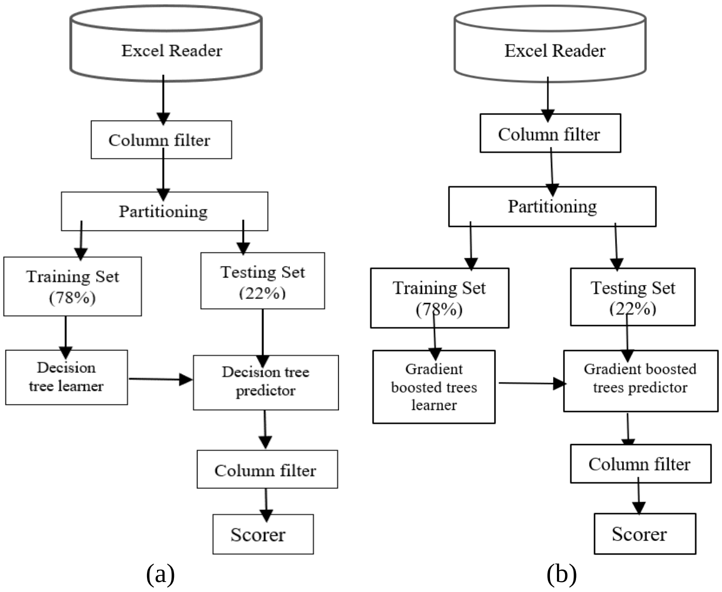


Figure 6. Model of (a) DT and (b) GBT

3. RESULTS AND DISCUSSION

In this section, we will elucidate the outcomes of our experiment. We have tested two different models on the testing set which consists of 22% of our dataset, those models are the advanced XGBoost option within the GBT framework and the CART decision tree employing the Gini index the outcomes were compared using a wide range of metrics which are stated in the next section.

3.1. Parameter metric

According to Ohsaki [19], Patro and Patra [20], a confusion matrix can provide the required information to determine how a classification model performs correctly. However, by summarizing this information in a single figure, it is more appropriate to compare the relative performance of different models. we use evaluation metrics to measure the effectiveness of the classifier, including the confusion matrix, accuracy stated in [9], [21], [22], precision in [23], recall in [24], and the f-measure in [25]-[27], which have been calculated as:

- Accuracy: this represents the ratio between all correctly classified instances and the total number of instances and is given in (5).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{5}$$

- Precision (positive predictive value): This represents the ratio between the number of correctly classified positive instances and the number of all correctly and incorrectly classified positive instances. It is also known as the positive predictive value and can be calculated using (6).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{6}$$

- Recall (positive sensitivity value): It represents the ratio of the number of correctly classified positives to the number of all the positive instances. It is also called a positive sensitivity value, which can be calculated by (7).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{7}$$

- F-measure: It is a model metric that can be used when seeking a balance between precision and recall, as shown in (8).

$$\text{F} - \text{Measure} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \tag{8}$$

In classification models, recall, precision, sensitivity, specificity, and F-measure measures are crucial for performance evaluation. Recall measures how well the model captures true positives, precision evaluates accurate positive predictions, sensitivity focuses on correct identification, specificity evaluates correct negative identification, and the precision and recall of the F-measure balance. These measures provide nuanced information that enables the model to be fine-tuned to the specific needs of the application.

Key terms are crucial in assessing the classification model's performance with two-class labeled data (positive and negative). True positives (TP) denote correctly labeled positive instances, while true negatives (TN) represent accurately labeled negative instances. False positives (FP) occur when negative instances are wrongly labeled as positive, and false negatives (FN) arise when the classifier erroneously labels positive instances as negative.

3.2. Results of the decision tree learning model

Concerning the decision tree algorithm, we have Table 5 illustrating the confusion matrix. The true positive class values of medium-altitude buildings (13 meters), open space, and high-altitude buildings (22 meters) are 62, 5, and 13, respectively, which represent the diagonal in Table 6. For DT, accuracy was 95.23%, and Cohen's kappa was 89.7%. Table 7 represents the parameters for each category.

Table 5. DT confusing matrix results

Row ID	True positives	False positives	True negatives	False negatives
Medium altitude of building (13 meters)	62	2	22	2
Open space	5	2	80	1
High-altitude of buildings (22 meters)	13	0	74	1

Table 6. DT prediction classes

Classe\prediction (Classe)	Medium altitude of building	Open Space	High-altitude of buildings
Medium altitude of building (13 meters)	62	2	0
Open space	1	5	0
High-altitude of buildings (22 meters)	1	0	13

Table 7. DT Parameters result

Name of category	Recall %	Precision %	Sensitivity %	Specificity %	F-measure %
Medium altitude of building (13 meters)	96.9	96.9	96.9	91.7	96.9
Open space	83.3	71.4	83.3	97.6	76.9
High-altitude of buildings (22 meters)	92.9	100	92.9	100	96.3

3.3. Results of the boosted gradient tree learning model

The ensuing Table 7 showcases the confusion matrix for the gradient-boosted tree predictor, GBT. We have the highest number of true positives in areas that refer to positive instances that have been correctly labeled as positive by the classifier compared to a few of the FP the negative instances that were incorrectly labeled as positive by the classifier. On the other side, we have TN these refer to the negative instances that were correctly labeled as negatives by the classifier their number is very large compared to the false negatives the positive instances that were mislabeled as negative by the classifier. Based on the GBT confusing matrix results in Table 8, we find that out of 84 test instances, the algorithm made only three misclassifications. As is observed from the table, there are three classes of zone categories, such as medium altitude buildings (13 meters), open space, and high-altitude buildings (22 meters).

Table 8. GBT confusing matrix results

Row ID	True positives	False positives	True negatives	False negatives
Medium altitude of building (13 meters)	62	1	23	2
Open Space	5	2	80	1
High-altitude of buildings (22 meters)	14	0	74	0

The true positive class values of medium-altitude buildings (13 meters), open space, and high-high-altitude buildings (22 meters) are 62, 5, and 14, respectively, which represent the diagonal in Table 9. Regarding GBT, the achieved accuracy was a satisfactory 96.96%. Additionally, Cohen's kappa statistic yielded a value of 92.4%. Table 10 presents the parameters associated with each category.

Table 9. GBT prediction classes

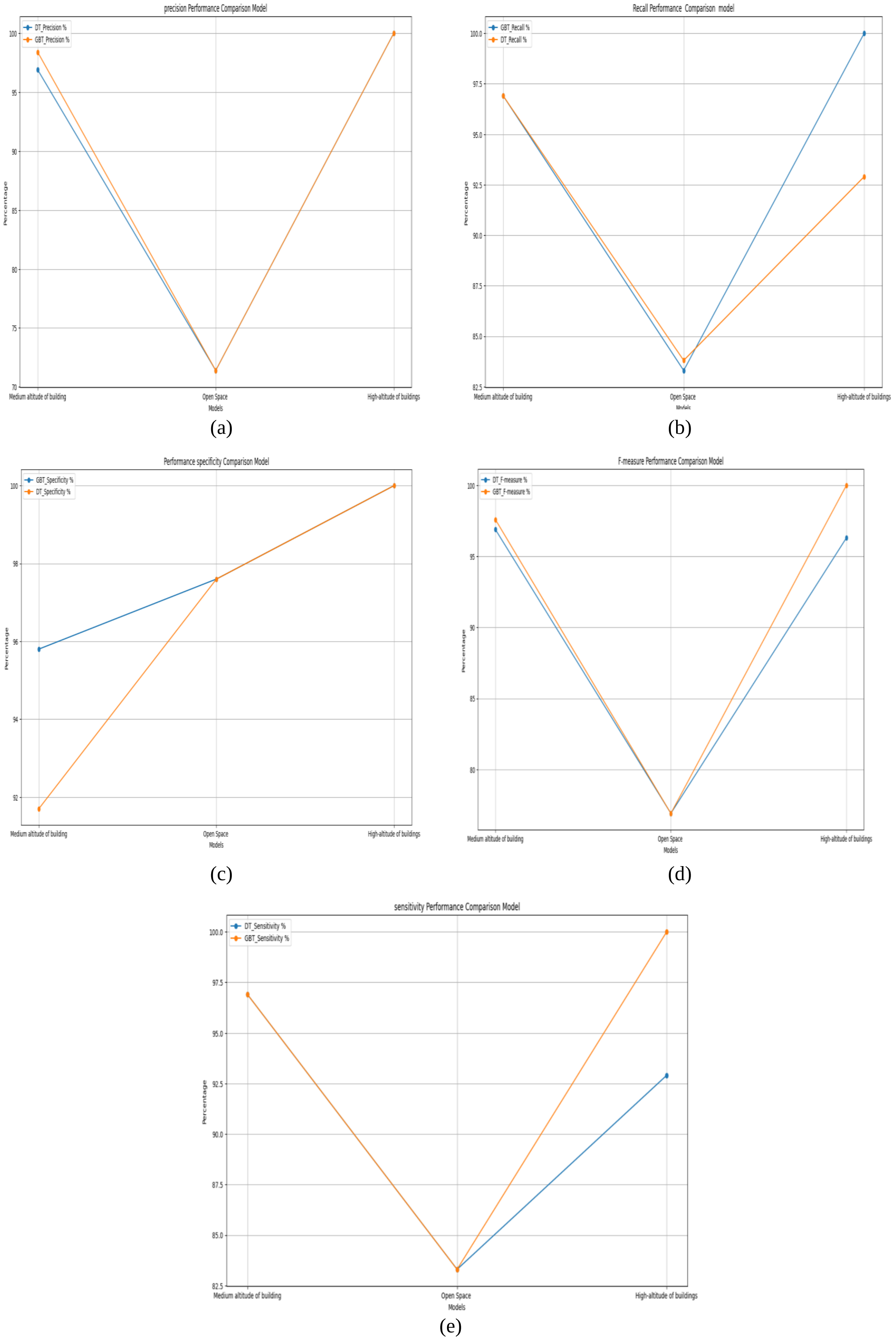
Classes/prediction (Classes)	Medium altitude of building	Open space	High-altitude of buildings
Medium altitude of building (13 meters)	62	2	0
Open Space	1	5	0
High-altitude of buildings (22 meters)	0	0	14

Table 10. GBT Parameters result

Name of category	Recall %	Precision %	Sensitivity %	Specificity %	F- measure %
Medium altitude of building (13 meters)	96.9	98.4	96.9	95.8	97.6
Open space	83.3	71.4	83.3	97.6	76.9
High-altitude of buildings (22 meters)	100	100	100	100	100

3.4. Experiments/analysis

After a comprehensive analysis of performance measures in Figure 7, including precision (Figure 7(a)), recall (Figure 7(b)), specificity (Figure 7(c)), f-measure (Figure 7(d)), and sensitivity (Figure 7(e)), we found the GBT method outperforms the DT method on several criteria. The performance of the GBT method is consistently superior, with greater precision for mid-rise buildings, a better recall parameter for high-rise building structures, greater specificity, and significantly better F-measurements for all infrastructure categories. In addition, GBT displays superior sensitivity measurements, outperforming DT in particular in the medium and high building categories. Thus, based on these assessments, GBT emerges as the superior method, offering more robust and accurate predictive capabilities.



(a) (b)

(c) (d)

(e)

Figure 7. Compare performance measurements for GBT and DT to (a) precision, (b) recall, (c) specificity, (d) F-measure, and (e) sensitivity

3.5. Interpretation

Our study suggests that the deterioration in signal strength is different depending on the zone environment. Our experimental results on PSSD are very encouraging. We found that the prediction results correlate with the signal strength variation in the positions. The proposed method in this study tended to have an inordinately higher proportion as the accuracy of GBT is 96.42% and the accuracy of DT is 95.23%. Our study suggests that higher deterioration of signal strength is not associated with poor performance in the network. Rather, it is related to the environment of the area. This study explained that PSSD could characterize the type of infrastructure in terms of signal strength degradation. The proposed method may benefit from PSSD; PSSD can provide a lot of time and potential to optimize the resource allocation for the telecom company, to detect the problem of signal strength deterioration without adversely influencing and using a test drive tool several times without arriving at a complete solution.

This study explored comprehensive network coverage; using the data, collection in real-time but the difference between the environments stays the impact to determine the homogeneous area. However, further in-depth studies are required to investigate the signal strength deterioration in other infrastructural areas, considering varying altitudes of buildings and curved spaces. Future studies may explore signal strength degradation and the relationship between base station location and building altitude, examining feasible methods for optimizing electrical energy consumption and evaluating strategies for achieving efficient coverage in networks. This research is essential to enable the telecom company to intelligently distribute the installation of new base stations. The distribution strategy should account for the specific characteristics of each infrastructure type and its associated signal strength deterioration, ensuring an effective deployment aligned with the unique features of each infrastructure category.

4. CONCLUSION

Automatic detection and prediction of signal strength degradation in urban areas using data-driven machine learning shows that the model can predict infrastructure as a function of RSSI, our results provide conclusive evidence that this phenomenon is associated with building height characteristics, this parameter being considered important for the prediction of signal strength deterioration. This shows that there is a clear difference in network coverage. These results are very important for future studies and for companies' decision-making to achieve a balance between all regions and make them subject to quality standards.in addition to modifying the way base stations is installed in such a way that it suits the characteristics of the region. Our future work will involve generating a model that predicts the rate of consumption of elective energies based on the area's infrastructure.

REFERENCES

[1] N. Basiran *et al*., "Analysis of signal strength variations for an urban public university campus in Bangladesh," *International Journal of Engineering Research and Technology*, vol. 9, no. 7, Jul. 2020, doi: 10.17577/IJERTV9IS070301.

[2] H. Mohsin, K. Abdulameer, and Z. N. Khudhair, "Study and performance analysis of received signal strength indicator (RSSI) in wireless communication systems," *International Journal of Engineering and Technology*, vol. 6, pp. 195-200, Jan. 2017, doi: 10.14419/ijet.v6i4.29558.

[3] I. Ahmed, R. Mohamed, and G. Abdalla, "Android-based drive test platform for cellular system," in *2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE)*, Sep. 2015, pp. 330-335, doi: 10.1109/ICCNEEE.2015.7381386.

[4] L. Sa'adu, "Assessment of mobile network signal strength for GSM networks in Gusau, Zamfara State," *International Journal of Innovative Science, Engineering & Technology*, vol. 6, no. 4, 2019.

[5] E. Ostlin, H.-J. Zepernick, and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 2735-2747, Jul. 2010, doi: 10.1109/TVT.2010.2050502.

[6] L. C. Fernandes and A. J. M. Soares, "Path loss prediction in microcellular environments at 900 MHz," *AEU - International Journal of Electronics and Communications*, vol. 68, no. 10, pp. 983-989, Oct. 2014, doi: 10.1016/j.aeue.2014.04.020.

[7] D. Green, Z. Yun, and M. F. Iskander, "Path loss characteristics in urban environments using ray-tracing methods," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 3063-3066, 2017, doi: 10.1109/LAWP.2017.2761299

[8] Y. Zhang and J. Shu, "Link quality estimation method based on gradient boosting decision tree," *International Journal of Sensor Networks*, vol. 36, no. 3, pp. 159-166, Jan. 2021, doi: 10.1504/IJSNET.2021.117232.

[9] G. A. Wayessa, "UMTS network coverage hole detection using decision tree classifier machine learning approach," PhD Thesis, Addis Ababa University, 2020.

[10] S. I. Rufaida *et al*., "Construction of an indoor radio environment map using gradient boosting decision tree," *Wireless Networks*, vol. 26, no. 8, pp. 6215-6236, Nov. 2020, doi: 10.1007/s11276-020-02428-7.

[11] E. Alepis and A. Kontogianni, "Smartphone Crowdsourcing and Data Sharing Towards Advancing User Experience and Mobile Services," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 14, p. 38, 2020, doi: 10.3991/ijim.v14i03.11815.

[12] J. T. Nearing *et al*., "Microbiome differential abundance methods produce different results across 38 datasets," *Nature Communications*, vol. 13, no. 1, p. 342, Jan. 2022, doi: 10.1038/s41467-022-28034-z.

[13] P. Bruce, A. Bruce, and P. Gedeck, *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

[14] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105400, Aug. 2020, doi: 10.1016/j.cmpb.2020.105400.

[15] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA: ACM, Aug. 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[16] S. K. Mohapatra, R. Khilar, A. Das, and M. N. Mohanty, "Design of gradient boosting ensemble classifier with variation of learning rate for automated cardiac data classification," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Aug. 2021, pp. 11-14, doi: 10.1109/SPIN52536.2021.9566084.

[17] C. Ma, X. Qiu, D. Beutel, and N. Lane, "Gradient-less federated gradient boosting tree with learnable learning rates," in *Proceedings of the 3rd Workshop on Machine Learning and Systems*, Rome, Italy: ACM, May 2023, pp. 56-63, doi: 10.1145/3578356.3592579.

[18] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of XGBoost," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937-1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[19] M. Ohsaki *et al*., "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1806-1819, 2017.

[20] V. M. Patro and M. R. Patra, "Augmenting weighted average with confusion matrix to enhance classification accuracy," *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, no. 4, pp. 77-91, 2014, doi: 10.14738/tmlai.24.328.

[21] A. E. Maxwell, T. A. Warner, and L. A. Guillén, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review," *Remote Sensing*, vol. 13, no. 13, p. 2450, 2021, doi: 10.3390/rs13132450.

[22] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Computer Science & Information Technology*, vol. 1, 2020, doi: 10.5121/csit.2020.100801.

[23] A. Husejinović, "Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers," *Periodicals of Engineering and Natural Sciences*, vol. 8, no. 1, pp. 1-5, 2020.

[24] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545-1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.

[25] D. Hand and P. Christen, "A note on using the F-measure for evaluating record linkage algorithms," *Statistical Computing*, vol. 28, no. 3, pp. 539-547, May 2018, doi: 10.1007/s11222-017-9746-6.

[26] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv*, October 10, 2020.

[27] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Machine Learning*, vol. 110, no. 3, pp. 451-456, Mar. 2021, doi: 10.1007/s10994-021-05964-1.

BIOGRAPHIES OF AUTHORS

Ibrahim El Moudden graduated in 2015 with a bachelor's degree in electronic physics at the Faculty of Science, Tetouan, Morocco, and in 2018 was awarded a fundamental master's degree in electronics and telecommunications at the Faculty of Science, Abdelmalek Essaadi University, Tetouan, Morocco. Now a Ph.D. candidate specializing in Artificial Intelligence and Machine Learning at the Faculty of Science and Technology, Université Abdelmalek Essaadi de Tanger, Tangier, Morocco, in the Department of Computer Science. He can be contacted at email: ibrahim.elmoudden@etu.uae.ac.ma.

Youssef Benmessaoud A Computer Science Doctor and Engineer, who graduated as an engineer from the International University of Rabat and later attained a doctoral degree from the University of Sciences and Technologies, specializing in Artificial Intelligence and Machine Learning. With over five years of professional experience in web development, database administration, and a profound understanding of complex data analysis. My career has taken me to diverse environments, including working for Onnix in Valencia, Spain, and RLM in Morocco as an IT engineer. Currently deepening my expertise through advanced research, I impart knowledge as a Part-time Tutor and have a history of impactful roles in IT engineering. My scholarly work is distinguished by publications in prestigious forums and journals including Computers a Q2 journal, and my role as a keynote speaker at the last AI2SD conference in 2023 reflects my commitment to advancing research in AI and ML. He can be contacted at email: ybenmessaoud@uae.ac.ma.

Prof. Abdellah Chentouf Professor at Fst Tangier Abdelmalek Essaadi University. Ph.D. in engineering science, energy, process engineering 1999–2004. Thesis: "contribution to the electrical, magnetic, and thermal modeling of a high frequency inductive plasma applicator", 5 articles published at IEEE Trans on. Magn. Use of numerical methods for the resolution of a coupled Electromagnetic and Thermal problem leading to systems of nonlinear equations. Development of software allowing global modeling of an HF induction plasma installation. He can be contacted at email: achentouf@uae.ac.ma.

Loubna Cherrat Professor-researcher at the National School of Commerce and Management in Tanger (ENCGT), holds a Ph.D. in Computer Science specializing in Distributed Database Engineering. With extensive teaching experience since 2018, encompassing modules from computing to advanced data management across various academic levels, she also contributes to educational programs at the Faculty of Sciences in El Jadida since 2016. Actively involved in supervising doctoral theses, Cherrat has guided research on predictive analysis for agricultural adaptation, intelligent systems for cancer monitoring. Additionally, she is a research internship at the Laboratory of Computer Science in Paris. Engaged in diverse scientific activities, she moderates workshops, co-chairs and participates in international conferences on advanced intelligent systems, serves on scientific committees, and registers as chair in related conferences. Holding patents for innovative devices, Cherrat contributes her expertise to multiple research projects, addressing topics such as heterogeneous information system mediation, satellite image integration for flood risk management and machine learning tools for effective customer segmentation in digital marketing. She can be contacted at email: l.cherrat@uae.ac.ma.

Ech-Charrat Mohammed Rida attained his Ph.D. in applied mathematics and computing, specializing in reverse logistics, from the National School of Applied Sciences of Tangier (ENSAT) at Abdelmalek Essaadi University in Morocco. Presently, he serves as an Associate Professor at Abdelmalek Essaadi University, affiliated with the National School of Applied Sciences of Tetouan (ENSATe). His academic pursuits center around optimization, supply chain management, data science, and artificial intelligence. He can be contacted at email: charrat.mohammed@uae.ac.ma.

Prof. Mostafa Ezziyyani IEEE and ASTF Member received the "Licence en Informatique" degree, the "Diplôme de Cycle Supérieur en Informatique" degree and the PhD "Doctorat (1)" degree in Information System Engineering, respectively in 1994, 1996 and 1999 from Mahmmeed V University in Rabat, Morocco. Also, he received the second PhD degree "Doctorat (2)" in 2006, From Abdelmalek Essaadi University" in Distributed Systems and Web Technologies. In 2008 he receives a Researcher Professor Ability Grade. In 2015, he receives a PES Grade the highest degree at Morocco University. Now he is a professor of Computer Engineering and Information System in Faculty of Science and Technologies of Abdelmalek Essaadi University since 1996. His research activities focus on the modelling databases and integration of heterogeneous systems (with the various developments to the knowledge base, the object BD, active BD, Multi- System Agents, distributed systems and mediation). This research is at the crossroads of databases, artificial intelligence, Software Engineering and Programming. Professor at computer Science Department, Member of La.S.I.T laboratory, and responsible of the research direction Information Systems and Technologies, he formed a research team that works around this theme and more particularly in the area of integration of heterogeneous information systems and muti-agents systems using WSN as technology for communication. AI2SD (Advanced Intelligent Systems for Sustainable Development) Global Summit Symposium Series General Chair. He can be contacted at email: mezziyyani@uae.ac.ma.