Random Forest and XGBoost Based Fingerprinting Using MMSE: An Approach to Data-Centric AI to Enhance Indoor Wi-Fi Localization Systems.

Mariame Niang [1], Philippe Canalda[2], François Spies[2], Massa Ndong [3], Ibra Dioum [4], Idy Diop[4], and Mohamed A.El Ghany [5]

[1] *University Cheikh Anta Diop of Dakar, 5005, Dakar, Senegal*
[2]*Department of FEMTO-ST Institute/UMR CNRS 6174 Montbeliard, France*
[3] *University Virtual of Senegal, Dakar, Senegal*
[4] *Higher Polytechnic School Cheikh Anta Diop University of Dakar, 5005, Dakar, Senegal*
[5] *German University in Cairo, 3611, Cairo, Egypt*

Abstract

The indoor localization problem consists in identifying the Cartesian coordinates of an object or a personal asset in the buildings, malls, hospitals, campuses, factories, etc. To solve this problem, we consider a Wi-Fi-based localization method called fingerprinting, a two-step process, where a radio map of the monitored area is constructed by collecting signal strength from known locations. An unknown location is then predicted using this radio map as a reference. In this paper, we first propose an adapted Random Forest (RF) and Extreme Gradient Boosting (XGB) algorithms. This adaptation, combined with Minimum Mean Square Error (MMSE), improves the accuracy problem caused by the change of environment and extends the concept by adding a signal processing functionality as an edge cloud feature to address a dynamic cooperation clustering. By embedding the Wi-Fi Access Point (WAP) with multiple antennas, the signals sent by the Mobile User Equipment (MUE) can be processed to improve the accuracy of the bootstrap. Adding Minimum Mean Square Error (MMSE) is a kind of data-centric approach because it yields high-quality data as input. The noise inherent in the location data is reduced and thus the performance of the MMSE-aided RF and XGB improved. This enhancement is further extended by sharing data between WAPS. Thus, the MMSE processing and the sharing of such processed data between WAPS enhance the positioning model performance. The performance of these methods is evaluated through robust and extensive experiments in real-time indoor areas, with regular and reproducible scenarios. We found an interesting outcome that the proposed approach can offer better time-2-market compared to the traditional, non-Machine-Learning-based indoor positioning system approach.

Keywords

Indoor Positioning, Wi-Fi signals, Fingerprinting approach, Machine Learning, Extreme Gradient Boosting (XGB), Random Forest (RF), Received Signal Strength Indicator (RSSI), Data-Centric Artificial Intelligence, Minimum Mean Square Error (MMSE).

1. Introduction

The rapid growth of the Internet of Things (IoT), resulted in a wide range of services, including Location Based Services (LBS). Generally, localization refers to the process of obtaining the same region or the geographical location of a user or a device. Enabling accurate location-based services depends on the availability of location information. Localization systems can be categorized into

IPIN 2022 WiP Proceedings, September 5 - 7, Beijing, China
EMAIL: mariame.niang@gmail.com (M. Niang); philippe.canalda@femto-st.fr (P. Canalda); francois.spies@univ-fcomte.fr (F. Spies); massandong@mail.com (M. Ndong); ibra.dioum@esp.sn (I. Dioum); idy.diop@esp.sn (I. Diop); moh_salim@hotmail.com (M. A. El Ghany)
ORCID:0000-0003-2577-1437 (M. Niang); 0000-0002-6477-3673 (P. Canalda); 0000-0002-9964-2745 (F. Spies); 0000-0001-5773-7589 (M. Ndong); 0000-0002-2586-3908(I. Dioum); 0000-0002-9143-196X (I. Diop); 0000-0002-6282-773 (M. A. El Ghany).

outdoor localization and indoor localization. The Global Positioning System (GPS) is the main technology used to determine the position in outdoor localization. However, its accuracy deteriorates in the indoor environment due to the poor penetration of GPS signals inside buildings, a lot of power consumption, and the multipath effects on the propagating signals [1]. There is an urgent need to address precise indoor localization. Nowadays, indoor localization is highly used in our daily life. It is used in tracking the location inside a building, malls, hospitals, campuses, factories, etc. Several techniques are employed for localization parameter measurements, including Time of Arrival (ToA) [2], Time Difference of Arrival (TDoA) [3], Received Signal Strength Indicator RSSI [4], Angle of Arrival (AoA), and Time of Flight (ToF) [5]. These approaches suffer from many challenges, including poor accuracy, high computational complexity, multipath effect, shadowing, fading, and delay distortion. The fingerprints method achieves great attention recently due to its promising results with various ways of making predictions. In fingerprinting, a database is first built with data collected from a thorough measurement of the field in the offline stage. Then, the position of a mobile user can be estimated by comparing the newly received test data with that in the database, the online phase.

Besides, Wi-Fi fingerprinting localization is one of the methods based on RSSI [6,7,8,9,10], Euclidean distance [11], based on RSSI ranging [12], trilateration [13], etc. Compared to other indoor localization methods, Wi-Fi fingerprinting localization technology has some advantages including low hardware requirements and wide scope of application. At the same time, the technology needs to cooperate with more advanced algorithms to ensure higher positioning precision [14]. However, indoor localization using Wireless Local Area Network (WLAN) fingerprinting faces several challenges including propagation effects, which degrades the localization accuracy [15].

The rest of the paper is organized as follows. Section 2 gives a brief about the state of the art. Section 3 presents our proposed localization methods. Section 4 presents the localization performance of the algorithm in different ways, and section 5 the conclusion of the work.

2. State of the art of the previous works

With the rapid growth in Machine Learning (ML) systems, similar approaches need to be developed in the context of ML engineering, which handles the unique complexities of the practical applications of ML. This is the domain of MLOps. It is a set of standardized processes and technology capabilities for building, deploying, and operationalizing ML systems rapidly and reliably. In recent years, ML algorithms such as K-Nearest Neighbor (KNN) [16], Random Forest (RF) [17], XGB [18,19], Support Vector Machine (SVM) [20,21], KNN, a rules-based classifier (JRip), Decision Tree (DT), RF, and SVM [22], KNN, WKNN [23], RF, and XGB [24] have been applied to the RSSI fingerprinting positioning technique and have achieved better location results.

When the structure and layout of the indoor environment change, the indoor wireless communication environment also changes, which leads to a large gap between the new environment and the established positioning fingerprint. However, the establishment process of the fingerprint is very time-consuming and laborious. It is not economical or realistic to update all positioning fingerprints regularly and frequently, which will greatly improve the maintenance cost of the RSSI location fingerprinting system. Several methods to reduce the inaccuracies in location measurements are proposed in the literature [25]. There is no regular test in the work we have seen. In our previous work [24], reproducing these tests can bias the experiments. To assess the bias of machine learning methods, carrying out more regular and reproducible tests will make it possible to resolve these questions.

It is possible to improve the position system performance by using fingerprint techniques that employ multipath information in an ML framework, which operates a dataset generated in real-time using MMSE. In this work, we consider the RSSI between the transmitter and the receiver as the localization attribute. This is because the RSSI-based approach poses minimum requirements on the Wi-Fi technology of the requisite modules. RF and XGB algorithms combined with MMSE are proposed to minimize both the measurement noise and resolve the accuracy problem caused by the change of environment for indoor localization tasks. The method first uses RF and XGB algorithms to establish an indoor positioning model, which can achieve indoor positioning. When the environment

changes, a further MMSE method is used to improve the initial positioning. However, Data-centric approaches to solving AI problems have been dominant in applications where large and high-quality datasets are available. Such approaches aim to improve model performance through the development of more complex architectures.

3. Experiments

The fingerprint map is built where it contains the data points covering the whole area to be used by the algorithms to predict the position. Each data point has the RSSI values from four fixed APs and their position. The whole area is 9.5 m x 9.25 m, as shown in Figure 1. A point was taken every 0.2 m from the x-axis and every 0.5 m from the y-axis starting from the origin unless there were obstacles like walls or furniture that prevented taking the point. This approach for the fingerprint map resulted in having 700 data points covering the whole area. Our approach was to increase the number of data points and decrease the spacing between them to increase the accuracy in predicting the location. We have as input a list of 700 points. For each measurement point, we have 20 RSSI values then we calculate the mean of the 20 points as RSSI. (m). However, the RSSI values are very fluctuating so the mean is not enough to characterize the precision. To improve accuracy, the mean (m) and the MMSE are combined. We performed a point density analysis for the different scenarios. For this, we carried out different scenarios depending on the size of the training and testing. First, we divided our data at 10 %, we have 70 for training and 630 testing points evenly distributed along with x coordinates at 0.2 m doing 1 of 2 along x and by doing 1 out of 5 according to the y coordinates at 0.5 m to respect the pitch homogeneously, that is to say, take the diagonal. At 33 %, we divided our database by 3 by doing 1 out of 3 along x and 1 out of 1 along y which gives 233 for training and 467 testing points respecting the step between the coordinates x and y. At 66 %, we used 2/3 of our database, i.e.467 for training and 233 testing points. At 80 %, we divided our base by 4/5 using the fourth points for training and fifth points for testing resulting in 560 for training and 140 testing points. Then, we added a random positioning algorithm as a reference algorithm to compare the quality of our proposal compared to the random one. For this, we took a random point among the 700 and we calculate the distance of this point from real coordinates which gives us a distance of 7.5 m. We also used the midpoint algorithm, another benchmark algorithm. The midpoint is the central point which corresponds to the 350 points of our database and we calculate the distance from this point to the 699 remaining points then we calculate the average. We found a distance of 3.5 m for the midpoint. Finally, we calculated the Confidence Interval (IC) for each test point, a statistical result by calculating the mean and the standard deviation. For this, we give a confidence interval on these values. We used the following formula to calculate the IC. If X is a random variable defined on Ω of unknown expectation m and standard deviation σ and if $\bar{x}$ is the mean of the values observed on a sample of size n, IC at the confidence threshold α for the parameter m is:

$$I_\alpha = [\bar{x} - t\frac{\sigma}{\sqrt{n}}, \bar{x} + t\frac{\sigma}{\sqrt{n}}] \text{ where } \pi(t) = \frac{\alpha+1}{2} \quad , \qquad (1)$$

In MLOps, the model training lets efficiently and cost-effectively run powerful algorithms for training RF and XGB with MMSE models. Model training should be able to scale with the size of both the models and the datasets that are used for training. The testing model capability lets us understand how newly trained models perform. It enhances the reliability of our ML releases by helping to decide whether to reject poorly performing models and promote well performing ones. In the process of serving predictions, once our model is deployed to an indoor environment, the model service starts accepting prediction requests and providing responses with predictions.The testing data is used to evaluate the predictions generated by the ML model. The predicted locations will be compared to the actual positions of the test points able to evaluate the performance of different algorithms.

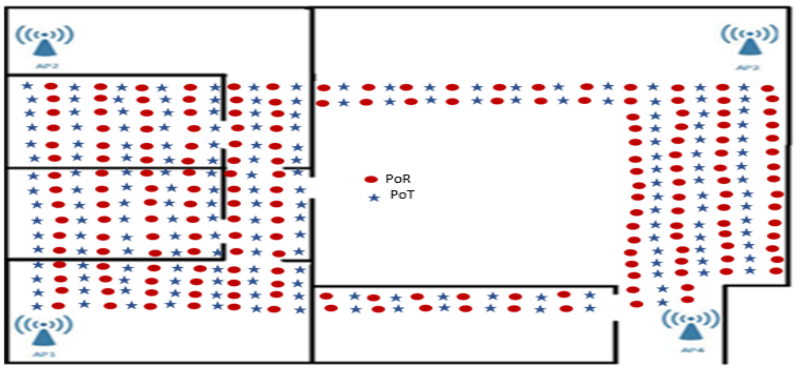

Figure 1: Area of indoor localization test:700 Point of Reference (PoR)/ Point of Test (PoT), real indoor evaluation room based on ML with various ratios (for example here 75 % training and 25 % testing with fairly regular topology).

3.1. Hardware

The offline phase is divided into different parts. Firstly, the RSSI reading was taken using an android app called Wi-Fi Fingerprint installed on HTC One X9. This RSSI value can be fluctuated due to the shadowing effect. Adding MMSE an approach of data-centric AI at each WAP mitigate the effect of environmental variation by reducing the noise in the data. This new fingerprint map was saved in an excel sheet CSV file to be used by the algorithm and sent to Python. Secondly, in the online phase, a Wi-Fi module ESP can read the values from APs and send this value to Firebase. Firebase database is specifically used because it is easy to be integrated with the Wi-Fi module and has also a library defined in Python making it easy to deal with the Firebase [26] database. Finally, Python IDE 'Spyder'was used to access the data in the excel sheet. The dataset is divided into training and testing. The training data is used to train the machine learning model to predict the position and the testing data is used to evaluate the predictions generated by the machine learning model, as shown in Figure 2.

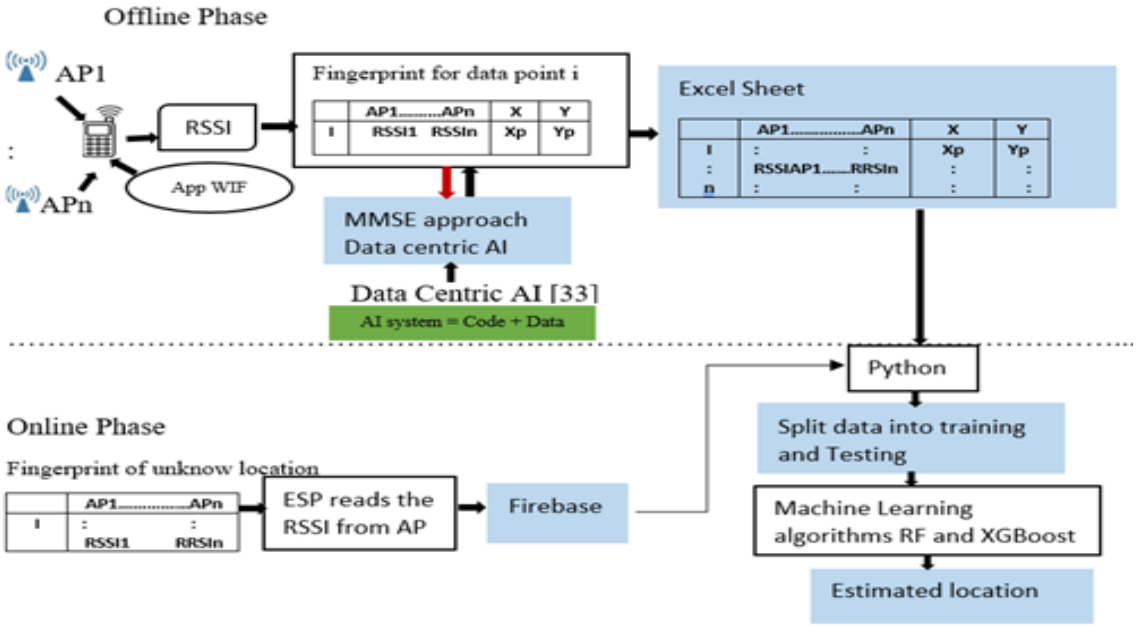

Figure 2: Steps of Fingerprint positioning using machine learning.

3.2. MMSE Estimation

A variety of speech enhancement approaches have been proposed. They differ in the statistical model, distortion measure, and in the manner in which the signal estimators are being implemented. Perhaps the simplest scenario is obtained when the signal and noise are assumed statistically independent Gaussian processes, and the MSE distortion measure is used. For this case, the optimal estimator of the clean signal is obtained by the Wiener filter. Since speech signals are not strictly stationary, a sequence of Wiener filters is designed and applied to vectors of the noisy signal. MMSE estimation under Gaussian assumptions leads to linear estimation in the form of Wiener filtering. Noise Reduction using MMSE can be used where the enhancement of noisy speech signals is essentially an estimation problem in which the clean signal is estimated from a given sample function of the noisy signal. The goal is to minimize the expected value of some distortion measure between the clean and estimated signals. For this approach to be successful, a perceptually meaningful distortion measure must be used, and a reliable statistical model for the signal and noise must be specified. At present, the best statistical model for the signal and noise, and the most perceptually meaningful distortion measure, are not known.

Due to the shadowing effect which deteriorates the MSE of localization. The MMSE estimation of Wireless Sensor Networks (WSN) is investigated. This MMSE algorithm can be used to locate the coordinates of unknown node values and also minimize location errors. Their simulation results show that the distance variance of distances between reference nodes and unknown nodes increases the MSE of localization [27]. In this paper, to calculate the MMSE, we use the method proposed in [28] by using for APs with their coordinates such as AP_1 (x_1 , y_1), AP_2 (x_2 , y_2), AP_3 (x_3, y_3), AP_4 (x_4 , y_4) and M (x, y) the coordinates of the mobile user

$$\begin{aligned} (x-x_1)^2 + (y-y_1)^2 &= d_1^2 \quad (A) \\ (x-x_2)^2 + (y-y_2)^2 &= d_2^2 \quad (B) \\ (x-x_3)^2 + (y-y_3)^2 &= d_3^2 \quad (C) \\ (x-x_4)^2 + (y-y_4)^2 &= d_4^2 \quad (D) \end{aligned} \quad ,(2)$$

After subtraction of the equations $(A)\ et\ (B)$ then $(C)\ et\ (D)$, we will have the following systems:

$$\{x_1^2 - x_2^2 - 2x\,(x_1 - x_2) + y_1^2 - y_2^2 - 2y\,(y_1 - y_2) = d_1^2 - d_2^2$$

$$\{x_2^2 - x_3^2 - 2x\,(x_2 - x_3) + y_2^2 - y_3^2 - 2y\,(y_2 - y_3) = d_2^2 - d_3^2$$

This can be written as a linear equation and becomes:

bX=a such as b= $\begin{bmatrix} x \\ y \end{bmatrix}$; a= $\begin{bmatrix} x_1^2 - x_2^2 + y_1^2 - y_2^2 - d_1^2 - d_2^2 \\ x_2^2 - x_3^2 + y_2^2 - y_3^2 - d_2^2 - d_3^2 \end{bmatrix}$;X= $\begin{bmatrix} 2\,(x_1 - x_2) & 2y\,(y_1 - y_2) \\ 2\,(x_2 - x_3) & 2\,(y_2 - y_3) \end{bmatrix}$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2\,(x_1 - x_2) & 2\,(y_1 - y_2) \\ 2\,(x_2 - x_3) & 2\,(y_2 - y_3) \end{bmatrix}^{-1} \begin{bmatrix} x_1^2 - x_2^2 + y_1^2 - y_2^2 - d_1^2 + d_2^2 \\ x_2^2 - x_3^2 + y_2^2 - y_3^2 - d_2^2 + d_3^2 \end{bmatrix} ,(3)$$

Distance measurements can be disturbed by noise or obstacles, which makes distances instead, distances are used with measurement errors and the equation becomes:

$$\widehat{d_i} = \sqrt{(x_i - \hat{x})^2 + (y_i - \hat{y})^2} \quad , \quad (4)$$

for i=1,.,n. n is the number of AP.

The Squaring and rearranging these terms yields the following equation for each access point measurement

$$\begin{aligned} (\hat{x}-x_1)^2 + (\hat{y}-y_1)^2 &= \widehat{d_1^2}(1) \\ (\hat{x}-x_2)^2 + (\hat{y}-y_2)^2 &= \widehat{d_2^2}(2) \\ (\hat{x}-x_3)^2 + (\hat{y}-y_3)^2 &= \widehat{d_3^2}(3) \end{aligned} \quad ,(5)$$

$$(\hat{x}-x_4)^2+(\hat{y}-y_4)^2=\widehat{d_4^2}(4)$$

$$\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}=\begin{bmatrix}2(x_1-x_2) & 2(y_1-y_2)\\ 2(x_2-x_3) & 2(y_2-y_3)\end{bmatrix}^{-1}\begin{bmatrix}x_1^2-x_2^2+y_1^2-y_2^2-\widehat{d_1^2}+\widehat{d_2^2}\\ x_2^2-x_3^2+y_2^2-y_3^2-\widehat{d_2^2}+\widehat{d_3^2}\end{bmatrix}, \quad (6)$$

The difference between equations (6) and (3) gives:

$$\begin{bmatrix}\hat{x}-x\\ \hat{y}-y\end{bmatrix}=\begin{bmatrix}2(x_1-x_2) & 2(y_1-y_2)\\ 2(x_2-x_3) & 2(y_2-y_3)\end{bmatrix}^{-1}\begin{bmatrix}(\widehat{d_1^2}-d_1^2)+(d_2^2-\widehat{d_2^2})\\ (\widehat{d_2^2}-d_2^2)+(d_3^2-\widehat{d_3^2})\end{bmatrix}, \quad (7)$$

$$A=\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix},\ W=\begin{bmatrix}2(x_1-x_2) & 2y(y_1-y_2)\\ 2(x_2-x_3) & 2(y_2-y_3)\end{bmatrix},Z=\begin{bmatrix}x_1^2-x_2^2+y_1^2-y_2^2-\widehat{d_1^2}+\widehat{d_2^2}\\ x_2^2-x_3^2+y_2^2-y_3^2-\widehat{d_2^2}+\widehat{d_3^2}\end{bmatrix}$$

A is solved using the Moore-Penrose generalized matrix inverse solution for the MMSE [29], [30].

$$A=(W^TW)^{-1}W^TZ \quad (8)$$

However, Federated learning (FL) is a distributed learning framework. As described in [31], FL requires end-users' devices with low computation power to send in their local pretrained machine learning model to a sink. The sink will concatenate the models into a global model to perform ML tasks. The models received at the sink are affected by noise, and the sink needs to mitigate the noise to effectively use the local models. Similarly, MMSE is used in our proposed approach to Data-centric AI to suppress the noise of the received measurement used in the fingerprinting.

3.2.1. Data-centric AI with MMSE

Due to the training datasets which impact the performance of the ML, this paper explores the concept of data-centric explanations for ML systems that describe the training data to the end-user. Their results show that data-centric explanations have the potential to impact how users judge the trustworthiness of a system and to assist users in assessing fairness [32]. A data-centric approach to AI provides a systematic way to improve data, build data consensus, and clean up inconsistent data. This is usually overlooked and data collection is treated as a one-time task. The data-centric approach is more rewarding and calls for a move towards data centrism. To make MLOps systematic, it uses firstly a model-centric view to collect what data it can develop a model good enough to deal with the noise in the data and hold the data fixed and iteratively improve the model. Secondly, it uses a data-centric view witch the consistency of the data is paramount. However, using tools to improve the data quality will allow multiple models to do well but to hold the code fixed and iteratively improve the data. MLOps' most important task is to make high-quality data available through all stages of the ML project lifecycle example prediction serving [33]. In wireless signal processing applications, where the RSSIs values are usually noisy, a potentially more fruitful approach is MMSE as an approach to data-centric AI one that focuses on improving the data to make simpler wireless network locations perform better. The idea is to enhance signal data by improving removing noise. This idea can be extended to include transforming signals into a wireless network where key features become more prominent and easier to use. However, with a data-centric view, there is significant room for improvement in problems with noise.

3.2.2. Random Forest MMSE

RF contains several DTs on various subsets of the given dataset and takes the average to predict the location and the accuracy of the dataset compared to other algorithms in ML such as SVM, KNN, etc. During training, a set of labeled training points can be used to optimize the parameters of the tree, and for testing the same unlabeled test input data is pushed through each component tree. At each internal position, a test is applied and the data point is sent for a prediction. To extend the concept by adding a signal processing functionality as an Edge cloud feature to implement a dynamic cooperation clustering, the MMSE algorithm at each WAP to enhance the quality of the bootstrapped data and share that

enhanced bootstrap with the neighboring WAPS in demand, and this MMSE is combined to the random forest.

Proposed RF. (MMSE) algorithm for dynamic cooperation clustering.

1. For k=1 to B
    - Draw N sample points from the collected data from the MUEs and the neighboring WAPS to form a bootstrap at the designated WAP
    - Applied the MMSE to the data collected from the MUEs to reduce the noise
    - Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum size $nmin$ is reached
        - Select m variables at random from the p variables
        - Pick the best variable/split-point among the m (iii) Split the node into two daughter nodes

2. Output the ensemble of trees $\{T_b\}_1^B$.

The prediction of a new location from the u=input data x is given by the regression

$$\hat{f}_{RF}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$$

The classification is given by the majority vote as follows: Let $C_b(x)$ be the class prediction of the b-th random forest tree, then

$$C_{RF}^B(x) = majority\ vote\ \{C_b(x)\}_1^B.$$

With the proposed RF. (MMSE) the algorithm, each WAP applies the RF locally using its data and the data received from the neighboring WAPS to construct the bootstrap. The contribution to this scheme is the sharing of data by the WAPS which enables a dynamic cooperation clustering. The data shared between WAPs is already processed with MMSE to reduce the noise. It further makes the size of the bootstrap variable at each WAP. The cluster of WAPS exchanging data is of a variable size too.

3.2.3. XGBoost MMSE

XGB is a software library. It split the X and Y data into a learning and testing set. The training set will be used to prepare the XGB model and the testing set will be used to make the predictions, from which it can evaluate the performance of the model. For this, it will use the train test split function from the scikit-learn library. It also specifies a seed for the random number generator so that we always get the same split of data each time. The format of the positions of the training data also needs to be modified for the fit function to work Finally. To improve the location accuracy caused by the change in environment, we propose to use XGB. (MMSE). The method first uses the XGB algorithm to establish an indoor positioning model. When the environment changes, further combine the MMSE method to improve the initial positioning.

4. Evaluation of performance

The performance of our developed system is evaluated in terms of localization accuracy. In MLOps, to evaluate the performance capability let's assess the effectiveness of our model, interactively during experimentation. For this, we need to visualize and compare performances of different models, compute pre-defined or custom evaluation metrics for our model on different slices of the data and track trained-model predictive performance across different continuous-training executions. This can help to enable model behavior interpretation using various explainable AI techniques. To evaluate the performance, the different localization algorithms are tested in simulation and compared, as shown in table 1. In all cases, the same training data was used to make the machine learning model. The MSE is used to measure the accuracy of the localization algorithms.

MSE= $\frac{1}{n}\sum (Y - \hat{Y})^2$,(9) where Y and $\hat{Y}$ are the actual and estimate coordinates at n-th references point.

4.1.1. Simulation description

For the simulation, we took all the test points for each percentage to sweep the whole space. That is to say take 630 test points for 10 %, as shown in Figure 3, 467 test points for 33 %, as shown in Figure 4, 233 testing points for 66 %, as shown in Figure 5 and 140 test points for 80 %, as shown in Figure 6. So, for testing, we have other possibilities for each percentage. We have 9 possibilities at 10 %, 3 possibilities at 33 %, 2 possibilities at 66 %, and 2 possibilities at 80 %. These experimental results show that at 10 %, the accuracy between RF. (m) and RF. (MMSE) is improved by 66 % and 48 % between XGB. (m) and XGB. (MMSE). At 33 %, there is a 79 % improvement in accuracy between RF. (m) and RF. (MMSE) and 80 % between XGB. (m) and XGB. (MMSE). At 66 %, there is a 22 % improvement in accuracy between RF. (m) and RF. (MMSE) and 28 % between XGB. (m) and XGB. (MMSE). At 80 %, there is a 27 % improvement in accuracy between RF. (m) and RF. (MMSE) and 29 % between XGB. (m) and XGB. (MMSE).

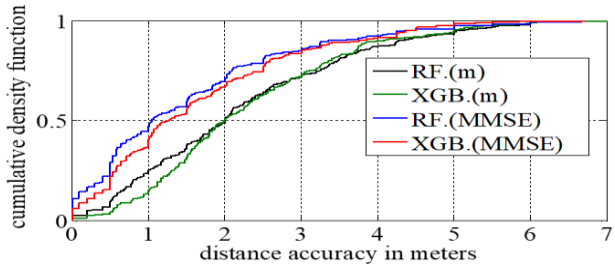

Figure 3: CDF of RF.(m), XGB. (m), RF.(MMSE), XGB. (MMSE) at 10 %

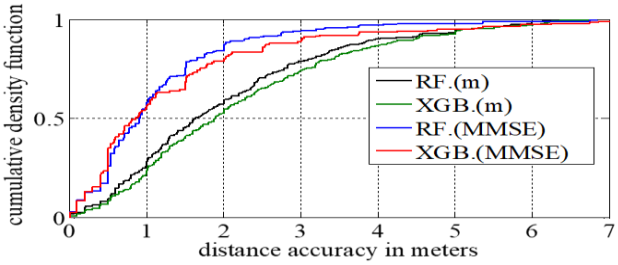

Figure 4:CDF of RF.(m), XGB. (m), RF.(MMSE), XGB. (MMSE) at 33 %

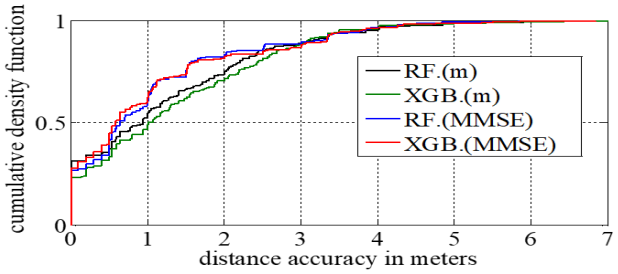

Figure 5: CDF of RF.(m), XGB. (m), RF.(MMSE), XGB. (MMSE) at 66 %

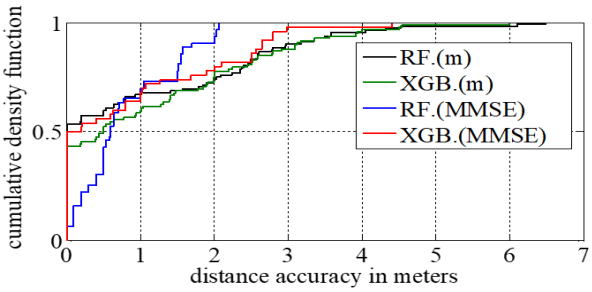

Figure 6: CDF RF, XGB, RF-MMSE, XGB-MMSE at 80 %

Table 1

Representation of positioning errors for the Egypt room from the data of a composition of elements

%	Scenario	RF.(m)	XGB. (m)	RF.(MMSE)	XGB.(MMSE)
10 %	T=70	2.26	2.36	1.60	1.88
	A=630	2.33	2.38	152	1.75
		2.21	2.31	1.55	1.78
		2.28	2.35	1.59	1.86
		2.23	2.30	1.61	1.80
		2.25	2.37	1.50	1.84
		2.24	2.39	1.57	1.79
		2.32	2.41	1.54	1.77
		2.34	2.40	1.56	1.73
	I_α at 95 %	[2.19;2.35]	[2.27;2.45]	[1.50;1.62]	[1.72;1.90]
33 %	A=233	2.01	2.17	1.22	1.37
	T=467	2.09	2.20	1.19	1.40
		2.02	2.19	1.17	1.35
	I_α at 90 %	[2;2.09]	[2.16;2.20]	[1.12;1.26]	[1.29;1.45]
66 %	A=467	1.25	1.30	1.03	1.08
	T=233	1.22	1.33	1.01	1.05
	I_α at 97 %	[1.20;1.25]	[1.28;1.33]	[1;1.04]	[1.04;1.08]
80 %	A=560	1	1.11	0.73	0.82
	T=140	1.02	1.13	0.70	0.79
		1.01	1.15	0.71	0.81
		1.03	1.12	0.72	0.80
	I_α at 97 %	[1;1.04]	[1.10;1.15]	[0.70;0.74]	[0.80;0.83]

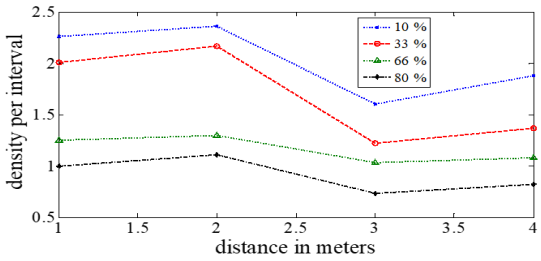

Figure 7: Percentages of different tests using density per interval.

4.1.2. Discussion of the experimental results

Analysis of our experimental data revealed that most location errors occurred due to attribution of too much relevance for low RSSI values, that is to say, corresponding to a weak reception, which would present fluctuations that can be further amplified by the presence of interior obstacles, can cause the coordinates of a point of distant affect the estimation. We compared the performance of the XGB and RF algorithm by using MMSE with the state-of-the-art in terms of accuracy. The experiment is done in a real-time environment, with a regular and reproducible scenario. Different scenarios of the test are done with different training and testing with regular distribution. The accuracy of RF.(m), XGB. (m), RF. (MMSE) and XGB. (MMSE) are respectively 2.26 m, 2.36 m, 1.60 m, and 1.88 m at 10 %.

At 33 %, we have 233 for training and 467 for testing, the accuracy of RF.(m), XGB. (m), RF. (MMSE) and XGB. (MMSE) are 2.01 m, 2.17 m,1.22 m, and 1.37 m respectively. At 66 %, we have 467 for training and 233 for testing. The accuracy of RF.(m), XGB. (m), RF. (MMSE) and XGB. (MMSE) are respectively 1.25 m,1.30 m,1.03 m, and 1.08 m. At 80 %, this means that we divided our data into 560 for training and 140 for testing. The accuracy of RF.(m), XGB. (m), RF. (MMSE) and XGB. (MMSE) are respectively 1 m, 1.11 m, 0.73 m and 0.82 m. These results show that RF. (MMSE) and XGB. (MMSE) give the highest accuracy than RF.(m), XGB. (m). These results confirm the interest of ML. But, the analysis of knowing which is the most efficient algorithm varies according to the training set compared to the testing set that is needed at 70 %, this is where we obtain the best result. such algorithms using RF or XGB vary, we do not have the same performance and above all the quality of the accuracy is really different. What seems more reasonable is the results we obtain today rather than in the initial test which according to the non-reproducible tests we have a bias which is very important of 2 % compared to the previous paper.

5. Conclusion

In this work, we performed an implementation, evaluation, and analysis of machine learning algorithms such as Random Forest and Extreme Gradient Boosting in an indoor environment. These algorithms are combined with MMSE, a data-centric approach to AI, to reduce the noise data and improve accuracy. This indoor location approach resulted in having 700 data points by using an app called Wi-Fi Fingerprint installed on the phone. Various regular and reproducible test sets were carried out. These regular tests are useful to evaluate the ML algorithms and to have a more real and reproducible. As part of an indoor experiment, XGB and RF combined with MMSE give better results at 80 % or 560 learning data and 140 test data with an accuracy of 0.72 m and 0.80 m respectively. The experimental results show that the proposed algorithms RF. (MMSE) and XGB. (MMSE) still achieve good positioning effect even in environmental changes compared to other algorithms, which makes it a good algorithm for the indoor location.

6. Acknowledgements

This work is the results of the research project funded by the International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA), Artificial

Intelligence for Development (AI4D) Africa Scholarship Fund Manager- Africa Center for Technology Studies (ACTS). This work was supported by the French government's "Eiffel excellence scholarship", program. [grant number N° P769615J-2021]

7. References

[1] A. S. Paul, and E. A. Wan, "Wi-Fi Based Indoor Localization and Tracking Using Sigma-Point Kalman Filtering Methods," Position, Location and Navigation Symposium, 2008 IEEE/ION, pp. 646-659, United States of America, 5-8 May 2008.

[2] D. Liu, Y. Wang, P. He, Y. Zhai, and H. Wang, "TOA localization for multipath and NLOS environment with virtual stations," EURASIP Journal on Wireless Communications and Networking, 2017.

[3] W. Gerok, J. Peissig, "TDOA assisted RSSD based localization using UWB and directional antennas," Leibniz Universität Hannover, Thomas Kaiser, Universität DuisburgEssen, Germany, 2013.

[4] A. Kokkinis, L.Kanaris, A.Liotta, S.Stavrou, "RSS Indoor Localization Based on a Single Access Point, " Sensors 2019, 19, 3711. https://doi.org/10.3390/s19173711.

[5] A.U.Ahmed, R.Arablouei, F.D.Hoog, B.Kusy, R.Jurdak, and N. Bergmann, "Estimating Angle-of-Arrival and Time-of-Flight for Multipath Components Using WiFi Channel State Information, " Sensors 2018, 18, 1753. https://doi.org/10.3390/s18061753.

[6] Y. Duan, K.Y. Lam, V.C. S. Lee, W. Nie, K. Liu, H. Li, and C. J. Xue, "Data Rate Fingerprinting: A WLAN-Based Indoor Positioning Technique for Passive Localization," IEEE Sensors Journal, Aug. 2019.

[7] A. Kokkinis, L. Kanaris, A. Liotta and S.Stavrou "RSS Indoor Localization Based on a Single Access Point," Department of Electrical Engineering, Eindhoven University of Technology, 5600 Eindhoven, The Netherlands, Journal Sensors 2019.

[8] A.Zhang, Y.Yuan, Q. Wu, S.Zhu and J.Deng, "Wireless Localization Based on RSSI Fingerprint Feature Vector,'' College of Computer and Information Engineering, Xiamen University of Technology, China, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 528747, 7 pages http://dx.doi.org/10.1155/2015/528747.

[9] P.Bahl, V.N Padmanabhan, " Radar: An in-building RF-based user location and tracking system '', In Proc. IEEE Infocom, Israel; 2000. p. 775–784.

[10] M.Youssef and A. Agrawala, "The Horus WLAN locationdetermination system'', Conference: Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Seattle, Washington, USA, June 2005. 7]

[11] W.Xue, Q. Li, X. Hua, K. Yu, W. Qiu, and B. Zhou, "A New Algorithm for Indoor RSSI Radio Map Reconstruction," Department of Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, School of Geodesy and Geomatics and Collaborative Innovation Center for Geospatial Technology, Wuhan University, School of Environmental Science and Spatial Informatics, China University of Mining and Technology, 2018

[12] Y. Huang, J. Zheng, Y. Xiao, and M. Peng, "Robust Localization Algorithm Based on the RSSI Ranging Scope," School of Electronic Information Engineering, Suzhou Vocational University, Publishing Corporation International Journal of Distributed Sensor Networks, China, Jan.2015.

[13] O.Pathak, P.Palaskar, R.Palkar, M.Tawari," Wi-Fi Indoor Positioning System Based on RSSI Measurements from Wi-Fi Access Points –A Tri-lateration Approach,'' International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014.

[14] Q.Yang, S.Zheng, M. Liu and Y. Z.Yang, "Wi-Fi indoor positioning in a smart exhibition hall based on received signal strength indication,'' EURASIP Journal on Wireless Communications and Networking'' (2019) 2019:275 https://doi.org/10.1186/s13638-019-1601-3

[15] A.Khalajmehrabadi, N.Gatsis and D.Akopian, IEE, "Modern WLAN Fingerprinting Indoor Positioning Methods and Deployment Challenges,'' IEEE Communications Surveys & Tutorials, Oct. 2016.

[16] M.S. Choi, B.Jang, "An Accurate Fingerprinting based Indoor Positioning Algorithm,'' Department of Computer Science, Sangmyung University, Seoul, South Korea, International

Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 1 (2017) pp. 86-90.
[17] E. Jedari, Z. Wu, R. Rashidzadeh, M. Saif, “Wi-Fi Based Indoor Location Positioning Employing Random Forest Classifier,” Department of Electrical and Computer Engineering, University of Windsor 401 Sunset Ave. Windsor, Alberta, International Conference on Indoor Positioning and Indoor Navigation (IPIN), Canada, Oct.2015.
[18] M.Luckner, B.Topolski, M.Mazurek, “ Application of XGBoost Algorithm in Fingerprinting Localisation Task, ” 16th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM), Bialystok, Poland, Jun 2017, pp. 661- 671, ff10.1007/978-3-319-59105-6_57ff. ffhal-01656240.
[19] W.Qiao, X.Kang, M.Li, “ An Improved XGBoost Indoor Localization Algorithm,’’ International Conference on Computer Intelligent Systems and Network Remote Control (CISNRC 2020), 2020.
[20] E.Schmidta, D.Akopiana, “Indoor Positioning System Using WLAN Channel Estimates as Fingerprints for Mobile Devices,’’ Department of Electrical Engineering, One UTSA Circle, San Antonio, TX 78249, Preprint of the 2015 IS&T/SPIE Electronic Imaging Conference, CA, February 8 - 12, 2015.
[21] Y. Tifani, B. Lee, E. Jeong, “A Patient’s Indoor Positioning Algorithm Using Artificial Neural Network and SVM,” Department of Computer Engineering, Catholic Kwandong University, Journal of Theoretical and Applied Information Technology, South Korea, Aug.2017.
[22] A.H. Salamah, M. Tamazin, M.A. Sharkas, M.Khedr, “An Enhanced Wi-Fi Indoor Localization System Based on Machine Learning,” Department of Electronics and Communications Engineering, Collège of Engineering and Technology, Arab Academy for Science, Technology and Maritime Transport, Alexandria, University of Alcalá, Madrid, Spain, International Conference on Indoor Positioning and Indoor Navigation (IPIN), Egypt,4-7 Oct.2016.
[23] O.G. Coast, Z. Kai, L. Binghao, A. Dempster, “A Comparison of algorithms adopted in fingerprinting indoor positioning systems,” School of Surveying and Spatial Information Systems University of New South Wales Sydney, International Global Navigation Satellite Systems Society (IGNSS) Symposium Australia, 16-18 Jul. 2013.
[24] M. Niang, M. Ndong, I. Dioum, I. Diop, M. Mashaly and M. A. A. E. Ghany, “Comparison of Random Forest and Extreme Gradient Boosting Fingerprints to Enhance an indoor Wifi Localization System’’, International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC),2021, pp.143,148, doi:10.1109/MIUCC52538.20 21.9447676.
[25] Q.Yang, S.Zheng, M. Liu, and Y.Zhang, “ Research on Wi-Fi indoor positioning in a smart exhibition hall based on received signal strength indication, ” J Wireless Com Network 2019, 275 (2019). https://doi.org/10.1186/s13638-019-1601-3.
[26] https://pypi.org/project/python-firebas
[27] Y.F. Huang, Y.T.Jheng and H.C.Chen, “Performance of an MMSE based indoor localization with wireless sensor networks, ’’ The 6th International Conference on Networked Computing and Advanced Information Management, 2010, pp. 671-675.
[28] J. Arnold, N. Bean, M. Kraetzl and M. Roughan, “Node Localisation in Wireless Ad Hoc Networks,’’ 2007 15th IEEE International Conference on Networks, 2007, pp. 1-6, DOI: 10.1109/ICON.2007.4444052.
[29] L. Hogben, “Handbook of Linear Algebra’’, N.W.: Chapman and Hall/CRC, 2007. pp 5.12-5.16.
[30] E. W. Weisstein, “Moore-Penrose matrix inverse,” From Math World– A Wolfram Web Resource,2002, https://mathworld.wolfram.com.
[31] Q. Lan, D. Wen1, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, K. Huang, “What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence’’, Department of Electrical and Electronic Engineering, University of Hong Kong, Journal of Communications and Information Networks 1 Oct 2021
[32] A.I. Anik, A. Bunt, “Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency,’’ University of Manitoba, Winnipeg, Canada, CHI ’21, Yokohama, Japan, May 08–13, 2021.
[33] A. Ng, “MLOps-From-Model-centric-to-Data-centric-AI’’