IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2020, accepted June 1, 2020, date of publication June 12, 2020, date of current version June 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001912

A Comparative Performance Evaluation of Machine Learning Algorithms for Fingerprinting Based Localization in DM-MIMO Wireless Systems Relying on Big Data Techniques

WALAA Y. AL-RASHDAN[1] AND ASHRAF TAHAT[2], (Senior Member, IEEE)
[1]Department of Electrical Engineering, School of Engineering, Princess Sumaya University for Technology, Amman 11941, Jordan
[2]Department of Communications Engineering, School of Engineering, Princess Sumaya University for Technology, Amman 11941, Jordan

Corresponding author: Ashraf Tahat (tahat@psut.edu.jo)

ABSTRACT Mobile terminal (MT) localization based on the fingerprint approach is a strong contender solution for utilization in microcells urban environments and indoor settings that suffer from severe multipath and signal degradation. In this paper, we investigate and evaluate the performance of thirteen machine learning (ML) algorithms (including multi-target algorithms) employed in conjunction with fingerprint based MT localization for distributed massive multiple input multiple-output (DM-MIMO) wireless systems configurations. The fingerprints will rely solely on the received signal strengths (RSS) from the single-antenna MT collected at each of the receive antenna elements of the massive MIMO base station. The performance is evaluated through numerical simulations incorporating practical millimeter-wave signal propagation models suited for 5G wireless systems in combination with ray-tracing techniques, and in conjunction with the 3D OpenStreetMap to replicate real-life environments. In addition, the ML computational platform, and implementation of the proposed framework was selected with a focus on efficiently handling the anticipated big data that could be generated from a typical 5G network with expected large subscriber cell density (1 million/km^2). To that end, an Apache Spark based ML platform is proposed and employed. Several DM-MIMO system topologies and configuration parameters combinations affecting MT localization were investigated to analyze performance. Numerical simulation results demonstrated that the location of a MT could be effectively predicted by means of a subset of the collection of considered ML algorithms. The obtained results of MT localization performance evaluation metrics served to identify an optimum ML algorithm and methodology for employment in DM-MIMO systems.

INDEX TERMS Fingerprint, localization, positioning, RSS, 5G, MIMO, artificial neural network, SVM, random forest, decision tree, KNN, gradient boosted, Gaussian process, Bayesian ridge, kernel ridge regression, big data, machine learning.

I. INTRODUCTION

Over the last two decades, location based services (LBSs) have attained a great deal of popularity, where most consumer gadgets and goods are equipped with user location feature [1]. Simultaneously, there is an exponential increase of applications incorporating user location awareness on smart mobile terminals (MTs), and are equally important as traditional vehicular navigation. Meanwhile, autonomous cars and the wireless Internet-of-Things (WIoT) MTs localization requirements resulted in unparalleled focus on and enthusiasm for localization research [1]–[3]. Nevertheless, the majority of localization applications stem in urban settings, where the commonly utilized global positioning satellite systems endure deterioration in accuracy as a consequence of diminished satellite signals in the absence of line-of-sight (LOS) propagation to MTs in the vicinity of high-rise buildings and shadowed locale, in addition to the large consumption of power on a MT. Massive MIMO, which is a fundamental framework for the preeminent 5G technology, is established on the utilization of large-scale antenna arrays at base stations (BSs) or access points (APs).

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

This architecture is an enabler of enhanced cellular communications that offer higher spectral density, and reduced complexity, in combination with millimeter-wave (mm-Wave) communications that exploit a wider spectrum [4], [5]. Although, the attention was centered primarily on enhanced communications as the main advantage of massive MIMO, high-accuracy localization will definitely be a principal feature in such systems [4]–[6]. Within wireless communication systems, localization is attracting greater interest because of the broad applicability and relative cost efficiency, where there are no MTs hardware specifications need to be imposed. Localization based on fingerprinting approaches relies on the fact that the signals propagating from MTs to BSs or APs are remarkably decided by the scattering surroundings in the vicinity of MTs locations [7], [8]. Hence, we can extract from the wireless channel a group of corresponding unique traits associated to each location identified as a fingerprint. Subsequently, the localization process can be regarded as a pattern recognition problem encompassing the extraction of fingerprints, fingerprint matching, and ultimately the prediction of location [7]–[9].

This paper analyzes and comparatively evaluates performance of ML algorithms (including multi-target ML algorithms) for MT localization based on the received signal strengthen (RSS) fingerprinting technique for distributed massive MIMO (DM-MIMO) outdoor cellular systems [7], [8]. The utilized framework for analysis and evaluation represents an environment that mimics real-life mm-Wave 5G networks. A collection of ML algorithms are employed in predicting MT locations under various network configuration parameters, topologies and weather conditions. Also, the employed big data ML platform, Apache Spark, was selected for its capability to accommodate designs suited for the anticipated 5G networks cell user density and traffic (1 million subscriber/km^2) as specified by the 3GPP Release 15. This is because Apache Spark is one of the latest big data handling engines for a ML framework [25]. Within this framework, we compared the performance of thirteen ML models within the described structures above to identify optimal trends for DM-MIMO system configurations and their associated parameters.

The rest of the paper is organized as follows: Section II discusses relevant foundation, background and associated literature. We present in Section III the employed proposed models and system architecture. Simulations results are presented and discussed in section IV. Finally, conclusions are drawn in Section V.

II. RELEVANT BACKGROUND

This section provides a review and discusses relevant background of main components and underlying principles of this framework that constitutes this work.

A. LOCALIZATION TECHNIQUES

Localization techniques are divided into four main categories: proximity-based, angle-based, range-based, and fingerprinting-based. The angle-based technique is based on angle-of-arrival (AoA). The signal angle of arrival is the direction from which the radio signal is received. However, this method fails when there is no LOS (NLOS) communication, and starts producing huge errors in position [9], [10]. The proximity-based technique is the simplest of the four techniques, where it mainly relies on checking whether or not the object to be positioned is present in a particular radio coverage area. It requires many base stations to implement and is not suited for large areas or areas with low density populations [11]. In range-based techniques, the range (distance) between the MT and a minimum of three base stations (BSs) (or 1 base station if the BS antennas are separate wide enough) is required to be computed to estimate the MT location by using trilateration. This can be accomplished through time-of-arrival (TOA), or received signal strength (RSS) measurements. The TOA method is known for its complexity and low performance in NLOS environments. The RSS approach has been proved useful in outdoor non-urban environments because the path loss is predictable as it is expected to decrease steadily with distance. On the other hand, in urban environments, its localization accuracy significantly deteriorates. To mitigate this performance issue, the RSS approach is enhanced when employed in conjunction with fingerprinting [3], [7]. Fingerprinting is a technique based on using a set of measurements (such as RSS data) to identify the current location using pre-recorded data called fingerprints. Fingerprints are collected at known positions on the map and saved in a database. This database of locations and the associated fingerprints are called the radio map. This database is used to predict and identify the current location of one or more MTs. This method is capable of working in highly-cluttered multipath environments [3], [7]. In recent years, localization in massive MIMO is attracting significant attention because of the evolving 5G wireless technology.

In [12], it was proposed by the authors to estimate the locations of a MT from TOA data gathered within multiple massive MIMO base stations, where they relied on the method of compressed sensing. In [13]–[15], the AOA information was utilized for localization of MTs in massive MIMO, where in [14] a mm-Wave massive MIMO system was the focus under LOS scenario. The authors of [16] investigated an mm-Wave architecture operating at 60 GHz employing a method for environment sensing. However, all of the schemes mentioned above are established relying on the information captured in a co-located MIMO base station structure, where the antenna elements are adjacent to each other in an array. Nevertheless, in DM-MIMO systems with system architectures similar to the ones we investigate in this work, where each of the MTs employs a single-antenna element, MT localization estimates will be undependable. In the relevant works, studies of [7] and [8] relate most to our investigation, which explore MT localization based on RSS measurements in DM-MIMO systems. Although, in [7] and [8] the focus of the analysis was limited to the Gaussian Process (GP) ML algorithm. Moreover, the simulations and

models adapted in [8] addressed current LTE MTs operating in frequency bands below 6GHz (i.e., not the mm-Wave frequency DM-MIMO systems). We expand on the works presented in [7] and [8] by investigating the performance of thirteen ML regression algorithms within variant mm-Wave DM-MIMO network configurations and under different realistic environment conditions on MT localization accuracy. To that end, we utilized a pertinent millimeter-wave signal propagation model in combination with ray-tracing techniques in conjunction with the 3D OpenStreetMap within our constructed 5G cell service area for coverage simulation, and to generate a MT fingerprinting datasets. Hence, our MT fingerprint datasets replicate real-life RSS values as a consequence of incorporating service area topography and surface clutter. Finally, in our investigations, the fingerprinting dataset sample sizes required for training the ML algorithms in the learning phase was also varied to assess computational cost, system complexity and efficiency of the 13 ML algorithms under consideration.

B. MASSIVE MIMO AND DM-MIMO

Massive multiple-input multiple-output (MIMO) is considered to be one of the principal technologies constituting 5G wireless cellular networks for their huge conceivable spectral efficiency and power efficiency coupled with reduced signal processing [17]. The large scale antenna array provides the advantage of acquiring multipath characteristics with high resolution at base stations, which prompts for employing fingerprinting based localization in wireless massive MIMO systems. In-turn, this will establish connections to machine learning applications for MT localization within this domain. Because of the large number of antenna elements, base stations are able to record large vectors of RSS on the massive MIMO uplink that are suited for ML.

We consider in our work a DM-MIMO system, where multiple single-antenna MTs are served simultaneously. DM-MIMO provides a more convenient deployment option as it relies on breaking down the base station into segments, utilizing a large number of distributed single-antennas connected to a central unit, to enhance the spatial diversity and to enjoy the versatility to choose from different antenna arrangements [18], [19]. Achieved DM-MIMO sum-throughput or average throughput data rates can surpass those of co-located massive MIMO network rates, and the overall capacity of the network [20]. Nevertheless, these benefits are attached to higher costs of deployment.

Today, 5G cellular communication systems are already being deployed in some countries like China, the United States and some European countries. Candidate mm-Wave frequency bands lie in the range of 24.5 GHz up to 100 GHz, in addition to sub 6 GHz frequency networks [21]. Although, a strong contender is the 28 GHz frequency band.

C. PROPAGATION MODEL

In [21]–[23], it was demonstrated that the close-in (CI) wireless signal propagation model is the most pertinent model to employ in the context of wideband mm-Wave frequencies communications for the outdoor urban environments that implement macro or micro cells. This is due to the fact that it is simple, accurate, with performance of remarkable sensitivity, and noting the actuality that, in outdoor environments, measured path loss manifests little reliance on frequency. These results were obtained with tests performed over vast frequencies that fall in the microwave and mm-Wave ranges, distances and schemes, while the used models were simple incorporating fewer parameters. In addition, it was proven that they were easily implemented facilitating replacement of existing ones. For the above mentioned reasons, we have chosen to utilize within our works here the CI propagation model. The CI large-scale propagation path loss model is expressed by the following equation as:

$$PL^{CI}(f, d)\,[\text{dB}] = FSPL(f, 1\text{ m}) + 10n\log_{10}(d) + \chi_{\sigma}^{CI} \quad (1)$$

In (1), parameter n denotes the path loss exponent (PLE) of the single parameter model. The interpretation of $10n$ is in units of dB of path loss over distance in decades, which starts at 1 m. d is the 3D T-R separation distance, with the carrier frequency of operation f. Also, $\chi_{\sigma}^{CI} \sim N(0, \sigma^2)$ is a random variable with normal distribution representing the shadowing fading effects. The free-space path loss (FSPL) is defined as:

$$FSPL(f, 1\text{ m})\,[\text{dB}] = 20\log_{10}\left(\frac{4\pi f}{c}\right) \quad (2)$$

where c is the speed of light in (2). Intrinsically, the CI model has an inherent dependence on frequency of path-loss incorporated into the 1 m free-space amount of path-loss. The adapted the CI propagation model parameters for an omnidirectional antenna as specified in [23] at the mm-Wave frequency band of 28GHz, where the path loss exponent $n = 2.1$ with $\sigma = 3.6$dB for LOS and $n = 3.4$ with $\sigma = 9.7$dB for NLOS. The CI path loss model was used in combination with the ray-tracing technique and in conjunction with the 3D OpenStreetMap within our constructed 5G cell service area for coverage simulation to generate the MT fingerprints datasets.

D. MACHINE LEARNING

The process of choosing the most suitable ML algorithm involves several elements [24], which can influence our decision since we will not be able to identify a single approach that will be most effective for all scenarios. As in our localization scenario we utilize the real GPS coordinates (i.e., latitude and longitude) in locating the MTs in the service area under study, a collection of thirteen supervised regression ML algorithms were employed and investigated. This is because the aim is to provide numerical estimates (i.e., GPS coordinates) of MTs locations relying on previously collected fingerprints in the vicinity of each of the current MT actual location. Their performance was evaluated and compared in estimating the MTs locations under variant network topologies and environment variables.

Thus, having an output variable y, a vector of input variables $\boldsymbol{x}$ and using a training set $\{(\boldsymbol{x_1}, y_1), .., (\boldsymbol{x_n}, y_n)\}$ of known values of $\boldsymbol{x}$ and associated values of y, the ML algorithm estimates or predicts unknown values of y, when provided an arbitrary vector of input variables $\boldsymbol{x}$.

The employed big data ML platform, Apache Spark, was selected for its capability to accommodate designs suited for the anticipated 5G networks cell user density and traffic (1 million subscriber/km^2) as specified by the 3GPP Release 15. This is because Apache Spark is one of the latest big data handling engines for a ML framework, and is a unified engine for large-scale data analysis and processing. Due to the fact that it is characterized by very high speed, in addition to scalability combined with its ease of use, it is very popular in the data science world. Apache Spark platform has been put to use by many of the prominent Internet giants including Netflix, Yahoo, eBay, …etc, which have adapted Spark at an extensive scale [25].

The following 13 ML algorithms were selected, and implemented in the comparative evaluation framework using Python object-oriented, high-level programming language:

1. Linear Regression
2. K-Nearest Neighbor
3. Decision Tree
4. Decision Tree Multi-Output
5. Random Forest
6. Random Forest Multi-Output
7. Gaussian Process
8. Support Vector Machines
9. Bayesian Ridge
10. Gradient boosted
11. Kernel Ridge
12. Stochastic Gradient Descent
13. Artificial Neural Network

We present below a brief introduction to the underlying approach and principles of operation to each of the employed ML algorithms in this paper.

1) LINEAR REGRESSION

Linear regression (LR) [24] is one of the most common and fundamental machine learning algorithms. The principles behind linear regression are establishing a linear relationship of the data points using a line of best fit. This approach models the relationship in a linear fashion when applied to scenarios of predicting output value (dependent variable) according to new input features (independent variables).

2) K-NEAREST NEIGHBOR

The k nearest neighbors (KNN) is one of the simplest, yet effective, forms of ML algorithms [26], [27]. It can be used for classification and regression problems. The principal idea is the use of feature proximity to predict values of new data points. This means a point in a dataset will exist in close similarly, based on defined distance metrics, to another point appearing in the training dataset that has similar feature or properties. The KNN algorithm measures the distance between a numerical target parameter and a set of parameters in the data set. In KNN regression, the expected result is a property value of the object. This property value is the average of the values of the k nearest neighbors. An alternate scheme utilizes an inverse distance weighted average of the k nearest neighbors. The distance is computed using a distance function d(x, y), where x the true is value and y is the predicted value. The distance equations used are:

Euclidian distance:

$$\mathrm{d}\,(x_i, y_i) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{k}} (x_i - y_i)^2} \tag{3}$$

Manhattan Distance:

$$\mathrm{d}\,(x_i, y_i) = \sum_{\mathrm{i}=1}^{\mathrm{k}} |x_i - y_i| \tag{4}$$

Minkowski Distance:

$$\mathrm{d}\,(x_i, y_i) = (\sum_{\mathrm{i}=1}^{\mathrm{k}} (|x_i - y_i|)^q)^{1/q} \tag{5}$$

Minkowski is a generalized distance equation, in which, q could take any value, however it is set to 1 for Manhattan distance, 2 for Euclidean distance [28].

3) DECISION TREE REGRESSION

As mentioned in [29], Decision Tree (DT) algorithm is a very common technique that is used regularly to construct classification and regression models in the shape of trees. The models created are sequential models that run a sequence of simple tests following logical binary rules. This means that the decision tree is created by asking questions with true or false answers and based on their answers, the trees narrow down until the model is confident to give a prediction. The orders and content of these questions are specified by the model. The algorithm used to build decision tress is ID3 (Iterative Dichotomiser 3) [30].

4) DECISION TREE MULTI-OUTPUT

Multi-output regression, or Multi-target regression (MTR), assigns each sample a set of target values [31]. Hence, it implies that for every data-point prediction is performed for a set of properties (e.g., the direction of wind at a certain location and its magnitude). MTR performs predictions of more than one variable concurrently given that a common set of attributes is provided [32]. Decision tree multi-output uses (DTM) the decision tree methodology; however it takes more than one parameter as labels. Those labels are usually not mutually exclusive. The standard one output (label) decision tree algorithm approach treats each output independently, whereas multi-output algorithm may treat the multiple labels simultaneously, accounting for correlated behavior among them [33].

5) RANDOM FOREST REGRESSION

The random forest (RF) technique works by building several independent decision trees in parallel with no interaction

among those trees when constructing them. It is called ''forest'' because it is made of many decision trees, and it is random because instead of averaging the prediction output of each tree, it uses random sampling of the training data.

Many samples are created from training data. Then a model is generated for each group of data samples. When new data is given to the random forest algorithm, each of the created models make a prediction on the new data, and the predictions are averaged to give an approximation of the right output value [34].

6) RANDOM FOREST MULTI-OUPUT

Random forest multi-output (RFM) regressor uses the random forest main algorithm with the support of multi-label parameters. It predicts several labels that are not mutually exclusive and that are related in a way or another. It generates multi-valued predictions instead of a single prediction based on a relation between the given predicted labels. It is commonly used in classification problems because of its capabilities in dealing with data that belongs to multiple classes, however its performance is studied in regression problems in this research [35].

7) GAUSSIAN PROCESS REGRESSION

Gaussian process (GP) is a Bayesian ML algorithm. Gaussian process regression (GPR) [36], [37] assumes an unknown nonlinear random function. This method is nonparametric technique for regression, where it is capable of providing probabilities for the output. Because of its computational complexity, it was not popular to use until early twenty first century.

8) SUPPORT VECTOR MACHINE

The support vector machine (SVM) ML algorithm is used in classification and regression tasks. It is a nonparametric scheme based on convex optimization. SVM may be utilized for regression problems since it is capable of modeling any nonlinear relationship. Because the estimation relies on a subset of the training instances (the support vectors), it is more efficient than alternate nonparametric algorithms. it uses a margin of tolerance where only the points that are within this boundary will be considered (the points with the least error rate) [38], [39].

9) BAYESIAN RIDGE REGRESSION

Ridge regression [40] is suited for the analysis of multiple regression data suffering from the existence among the independent variables of near-linear relationships. When those non-linear relationships occur, the least squares are unbiased; however, because their large enough variances they allow drift away from the true values. When a degree of bias is added to the regression estimates, a reduction in the standard errors was demonstrated by ridge regression, and this is where Bayesian ridge regression (BRR) evolved from [41].

10) GRADIENT BOOSTED

Gradient boosting (GB) is a dynamic technique for predictive models construction [42]. It was created based on a statistical framework that cast boosting as a problem of numerical optimization with the aim of minimizing the loss of the model via including weak learners by means of a gradient descent like procedure. It operates as an additive model in stage-wise fashion, since a single new weak learner is added at a time, and already existing weak learners are fixed and left unaltered in the model.

Thus, having an output variable y and a vector of input variables x and using a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ of known values of $\boldsymbol{x}$ and corresponding values of y, the gradient boosting estimation or approximation function is calculated as follows:

$$\mathrm{F}_J(x) = \mathrm{F}_{j-1}(x) - \gamma_j \sum_{i=1}^{n} \nabla_{F_{j-1}} L\left(y_i, \mathrm{F}_{\mathrm{j-1}}(x_i)\right) \quad (6)$$

where J is the number of iterations, γ_j is the step length and L is the loss function [43].

11) KERNEL RIDGE

Kernel ridge regression (KRR) is a non-parametric form of ridge regression. Ridge regression, classification, and the kernel trick are all combined within KRR. By minimizing a squared loss of a squared norm regularization term [44], in the space induced by the respective kernel, k, the objective is to learn a desired function.

Formulated in closed form, we can express the solution as:

$$\alpha = (k + \tau \mathrm{I})^{-1} y \quad (7)$$

where in the space induced by the kernel, k is the kernel matrix, and vector of weights is represented by α, τ is the conditioning factor, and I is the identify matrix. The learned function may be found as:

$$f(x) = \sum_{i=1}^{n} \propto_i k(x, x_i) \quad (8)$$

12) STOCHASTIC GRADIENT DESCENT

In the stochastic gradient descent (SGD) ML algorithm, the goal of the gradient descent process is to minimize an error function by updating, in an iterative fashion, a set of parameters. Samples from the training data, in random fashion, are selected every run within the framework of gradient descent in order to update the parameters during optimization, and hence the word stochastic.

A solution to the equation $f'(\boldsymbol{x}) = 0$, evaluating the gradient of the function, may lead to finding a minimum of a line $f(\boldsymbol{x})$. Gradient descent carries-out, with a step size, a number of iterations. In each iteration, to minimize the error in Eq. (8), all of the coefficients (weights) will be updated. Where w_j represents the selected weight, the step size is η, h_j is the feature associated with w_j, and the gradient is manifested in the last part is [45].

$$\mathrm{w}_j^{(j+1)} \leftarrow \mathrm{w}_j^{(j)} + 2\eta \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{h}_j(\mathrm{x}_i)(\mathrm{y}_i - \hat{\mathrm{y}}_i(\mathrm{w}^{(j)})) \quad (9)$$

13) ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are defined in [46] as "a mathematical model that is based on biological neural networks and therefore is an emulation of a biological neural system" [47]. The perceptron is the starting point in the procedure of forming a neural network. Simply put, the inputs are received by perceptron, where they are multiplied by a set of weights, and subsequently passed to employed activation function (e.g., identity, relu, logistic, tanh,) to generate an output. To form a neural network, the perceptron layers are cascaded together, recognized as a multilayer perceptron model (MLP). Every neural network has three types of layers, which are the input, hidden, and output layers. The data is candidly received by the input layer, while the required output is produced by the output layer. The commonly known hidden layers are situated in between the input and output layers, where the transitional computations are performed.

III. PROPOSED LOCALIZATION MODEL

This investigation targets realistic RSS fingerprinting-based techniques in conjunction with ML learning algorithms for localization in a multi-user DM-MIMO system. Our objective is to estimate the location of MTs utilizing their respective received signals at the DM-MIMO base station (BS). Hence, employing M antennas at the BS, we examine the received signals of a multi-user DM-MIMO system on the uplink, where each of MTs employs a single-antenna, each MT will have a vector of received signal strengths (RSS) as $\mathbf{r} = (rss_1, rss_2, rss_3, \ldots rss_M)$. Those values will be used to predict the user geographical location (latitude and longitude) on the map, i.e., GPS coordinates. So in this 2-dimensional, we denote the location by (x, y). Although the proposed system works by taking multiple users' received signals strengths vectors and producing their location estimates, we consider a single MT at a time. Therefore, the regression problem we need to solve is stated in equation form as follows:

$$x = f_x(\mathbf{r}) + \varepsilon_x \quad \text{and } y = f_y(\mathbf{r}) + \varepsilon_y \tag{10}$$

where each of $f_x(\cdot)$ and $f_y(\cdot)$ is a non-linear function of the input received signal strengths vector $\mathbf{r}$, and each of ε_x and of ε_y is a Gaussian random variable ($\varepsilon_x, \varepsilon_y \sim \mathcal{N}(0, \sigma^2)$) that represents the error. We evaluate the performance of the above mentioned 13 ML algorithms and techniques described above in attaining solutions in the formulated problem of MT location.

A. THE GENERAL FRAMEWORK OF INVESTIGATION

The components and architecture of the framework of our investigative analysis and experiments are depicted in Fig. 1. It is comprised of interconnected modules to achieve the objectives enumerated in the previous sections. Each phase of our framework will be illustrated in this section.

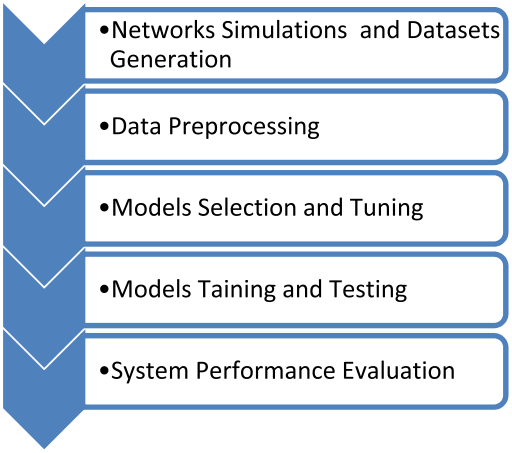

FIGURE 1. Framework and system design process.

FIGURE 2. PSUT campus area.

1) NETWORKS SIMULATION AND DATA GENERATION

The first phase in our investigative research was to datasets generation (fingerprinting), that are comprised of the received signals on the uplink of the single-antenna MTs in the a multi-user DM-MIMO system. The MATLAB software package was utilized to construct simulation environments that resemble a realistic mm-Wave 5G DM-MIMO network architecture to generate the desired datasets of fingerprints. This is because 5G is not yet implemented in Jordan. The Princess Sumaya University of Technology (PSUT) campus in Amman, Jordan was chosen as the center for the virtual area of experimentation as depicted in Fig. 2.

Nevertheless, this simulation can be arbitrarily implemented in any geographical area of interest.

In our work, we have constructed customized simulations to implement twelve different network configurations to generate the corresponding fingerprinting datasets. A sample 64 antenna element DM-MIMO BS configuration at the specified service area grid is illustrated in Fig. 3. The generated fingerprints and corresponding geographic area are depicted in Fig. 4. We pinpoint the major steps that were performed in order to accomplish the construction of these simulations and to resemble realistic environments:

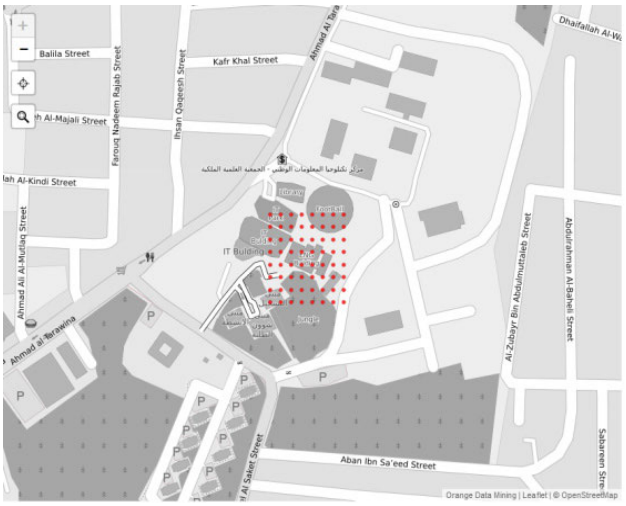

FIGURE 3. An example 64 antenna element arrangement in the simulated DM-MIMO BS at PSUT campus.

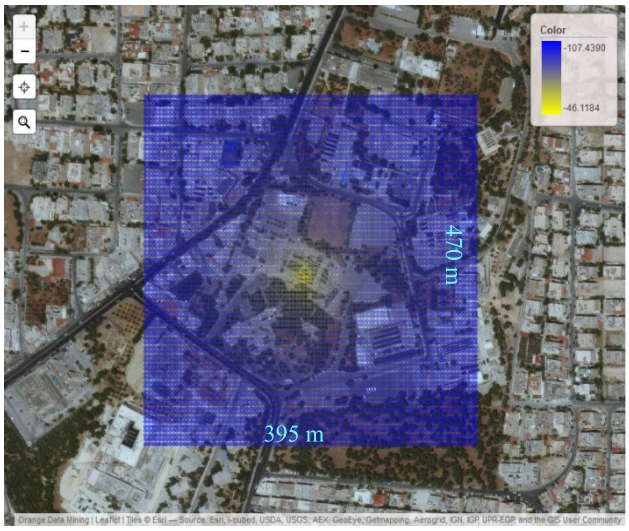

FIGURE 4. Sample DM-MIMO mm-Wave network service area fingerprints grid.

i. A satellite map of the area under consideration was obtained from OpenStreetMap open source maps and enhanced by embedding a 3D Building layer.
ii. Terrain gmted2010 dataset was used for computation and elevation information, that had been built-in within the satellite map.
iii. A set of realistic 5G FR2 MT transmitters of 1W (30dBm) transmitting power, and DM-MIMO antenna receivers were created. Their corresponding fingerprints were plotted on the area map as shown in Fig. 4.
iv. The CI propagation model [21]–[23] was employed for mm-Wave signal propagation in our simulated 5G networks in conjunction with ray-tracing to predict and generate the RSS at each receiving antenna of the DM-MIMO BS.
v. The generated fingerprinting datasets for the desired network configurations and variant parameters were exported in the CSV file format.

In the 3GPP Release 15, for the FR2 type UE operating in the 28 GHz frequency band and of power class 3 (i.e., a mobile terminal of the Smartphone type,..etc.), the maximum permitted EIRP is 43 dBm, while the minimum peak is specified at 22.5 dBm [48]–[50]. In our simulation, we set the EIRP of the MT to be 1W (30 dBm). The DM-MIMO BS antenna array was placed in the center of a cell with a square geographical grid format, with the generated fingerprints placed around them in a rectangular grid format as depicted in Fig. 3 and Fig. 4. The RSS values on antenna element 1 of the DM-MIMO BS array is also depicted on Fig. 3. Initially, twelve MT RSS fingerprints datasets containing 7600 instances each were generated, where each dataset correspond to a specific network configuration and MT fingerprints location resolution in clear skies weather conditions scenarios. The resolution of the fingerprints grid is the minimum distance separation between a MT fingerprint location and the nearest adjacent one. The variant network configurations and parameters used in generating the datasets are discussed in the following sections. Each instance (record) within each of the datasets corresponding to one MT fingerprint location has the following format: [rss_1, rss_2, rss_3, ... rss_M, *latitude*, *longitude*],where the M is the number of antenna elements in the DM-MIMO antenna array of BS, and the RSS values are expressed in dBm.

2) ML MODELS TRAINING AND TESTING

To conduct our investigations of comparative performance evaluation of the previously introduced ML algorithms, each of the datasets was divided into two partitions, training data and testing data. Training data is used to train the algorithm and the testing data is used to test and evaluate the performance of each of the ML algorithms. In the initial evaluation phase, the splitting ratio of training to testing data instances of each of the datasets was 70 percent for training while the remaining 30 percent was used for testing the models. Varying this splitting ratio was later evaluated to analyze the effect of fingerprint dataset size on the efficient and effective operation of this approach within the context of DM-MIMO localization accuracy. The splitting of datasets enables evaluation of performance of the ML algorithms on data that they have never encountered.

3) PERFORMANCE EVALUATION

In order to classify performance optimality and identify best ML algorithm(s) for MT localization in conjunction with preferred DM-MIMO network configurations, a set of evaluation metrics were employed [51]. In the discussion that follows, the true or *actual* value is designated by t, while p is the *predicted* value of t, and $\bar{t}$ is the mean value of t.

The most common error measure for regression ML algorithms is the root mean squared error (*RMSE*) over a set of N data instances.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - p_i)^2} \tag{11}$$

The *MAE* is the mean-absolute error, which is expressed in equation form as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|t_i - p_i| \tag{12}$$

The median absolute error (*MedAE*) is particularly of interest in ML because it is robust to outliers. The error is measured by taking the median of all absolute differences between the true value and the prediction value.

$$MedAE = median\left(|t_1 - p_1|, \ldots, |t_i - p_i|\right) \tag{13}$$

IV. NUMERICAL RESULTS AND DISCUSSION

Our goal is to estimate the locations of the MT test locations distributed in a random fashion across the square cell service area understudy as depicted in Fig. 4. To that end, we present the numerical results and computations of performance metrics for comparative evaluations of localization accuracy of the ML algorithms under study within the variant DM-MIMO network configurations and with different parameters. We consider the example DM-MIMO setup, as illustrated in Fig. 3, and Fig. 4 with $M = 16$, 36, 64, and 100 antennas and 7600 MT RSS fingerprints locations instances distributed over a service area of 395m x 470m spaced in a grid like fashion.

The generated RSS MT fingerprints are the noise-free steady-state received signal strength values produced using the Matlab® channel link analysis using ray-tracing in the mm-Wave frequency band [52]. The CI propagation model [22], [23] was incorporated in this ray-tracing simulation in conjunction with the downloaded 3D OpenStreetMap [53] of the cell service area under consideration. We adapted the CI propagation model parameters for an omnidirectional antenna as specified in [23] at the mm-Wave frequency band of 28GHz, where the path loss exponent $n = 2.1$ with $\sigma = 3.6$dB for LOS and $n = 3.4$ with $\sigma = 9.7$dB for NLOS. This comprehensive setup of the simulation will lead to near real-life values that are repeatable since large scale surface effects, terrain, and shadowing are incorporated in the RSS estimates. The MT transmitter placed at each of the RSS fingerprints locations is of the FR2 type UE operating in the 28 GHz frequency band and of power class 3 (i.e., Smartphone type) [48]–[50]. In our simulation, we set the EIRP of the MT to be 1W (30 dBm).

The DM-MIMO BS antenna elements separation was analyzed on the system's localization performance by considering inter-element spacing distances of 2m, 5m, 10m, 15m, while fixing the fingerprinting resolution at 5m in a subset of datasets. Also, the MT fingerprints locations resolution was varied as 2m, 5m, 10m, 15m, while setting the number of antenna elements to 64 and inter-element spacing distances of 5 m in a subset of datasets. Moreover, the effect of changing weather conditions, operation frequency band change, and training fingerprints sample size reduction (i.e., training and testing splitting ratio) were also investigated, and analyzed their effect on the DM-MIMO system's localization accuracy.

TABLE 1. Common simulations parameters.

Parameters	Values
Frequency	28GHz
Location	PSUT region
Fingerprints count	7600
Fingerprints resolution	5 meters
Antennas resolution	2 meters
Propagation Model	CI
Weather	Clear skies

A. COMPARATIVE STUDY OF THE EFFECT OF CHANGING THE NUMBER ANTENNA ELEMENTS ON LOCALIZATION ACCURACY

This part analyzes the effect of the DM-MIMO BS antennas count on the MT location prediction accuracy. Four 5G DM-MIMO BS antenna array configurations were simulated with common simulations parameters as listed in Table 1. The considered antenna arrays are: 4×4, 6×6, 8×8, 10×10.

Performance comparison of the 13 ML in terms of localizations accuracy expressed as MAE distance from the actual location is plotted in Fig. 5 for different antennas count at BS.

It can be seen that all estimators' results go, hand in hand, except for ANN, showing that an increase in antenna number enhances the cellular positioning performance. The resulting, a MAE of 7 meters has been reported when using 100 antennas at the BS. The KNN demonstrated the best performance in localization accuracy, followed by Kernel Ridge, then Random Forest, followed by the rest of the studied algorithms.

It is noticed that neural network algorithm came last in positioning accuracy with different behavior at some points. Neural networks follow complicated models with highly sensitive learners and highly customizable parameters. Some of the fundamental problem with deep learning and neural networks in general is that the solutions that fit training data are infinite. This means, there are truly endless possibilities when training the data to get the required results. Also, neural networks don't always give the most optimum results, following the no-free-lunch theorem [43], that suggests that there can be no single method which performs best on all datasets. Therefore, it is natural that ANN will perform better in some cases and worse in some other. It all depends on the datasets and the generalization capabilities. For each specific problem with a specific dataset format and presentation, it is needed to apply different algorithms in order to choose the one that performs the best and this is what we did in this research.

B. COMPARATIVE STUDY OF THE EFFECT OF CHANGING THE ANTENNA ARRAY INTER-ELEMENT SEPARATION DISTANCE ON LOCALIZATION ACCURACY

The effect of DM-MIMO BS antenna elements separation distance on the accuracy of MT localization was studied. To that end, comparisons were conducted relying on four different antenna array configurations at the 28 GHz band with respective generated fingerprints datasets that were fed

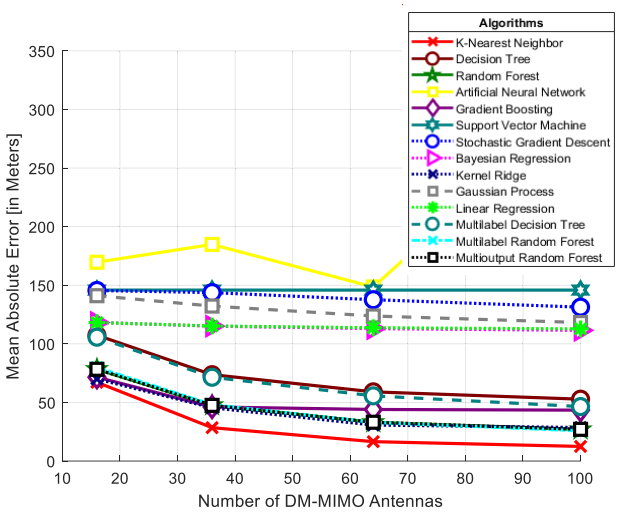

FIGURE 5. Effect of Changing the number of antenna elements on localization accuracy using MAE.

TABLE 2. Effect of antenna array inter-element separation distance on MT localization accuracy using MAE.

	MAE (meters)			
ML Algorithm	2	5	10	15
KNN	16.54	7.50	4.29	3.27
DT	59.04	36.12	33.59	32.07
RF	33.04	19.40	18.17	17.96
ANN	148.50	124.76	127.25	130.94
GB	44.00	41.08	40.04	39.48
SVM	145.86	145.86	145.86	145.86
SGD	137.67	118.58	103.49	93.49
BR	112.96	105.67	93.40	78.58
KRR	30.71	17.01	8.10	4.81
GP	123.76	109.83	98.29	88.22
LR	113.74	106.41	93.92	78.88
DTM	55.62	34.71	30.60	32.13
RFM	33.04	16.15	13.48	12.73

into the ML algorithms under investigation. The DM-MIMO BS antennas are arranged in a square format at the center of the service area grid as depicted in Fig. 2 and Fig. 3. The fingerprinting resolution is set at 5 meters. We considered inter-element spacing distances of 2, 5, 10 and 15 meters. We summarize in Table 2 the obtained MT localization performance results in terms of the MAE metric in meters.

The MT localization accuracy was a MAE of 16.5 m at 2 meter antenna inter-element spacing, 7.5 m at 5 m spacing, 4.3 m at 10 m spacing and has reached 3.3 m at 15 meters antennas separation, respectively.

Also, the evaluation metrics (MAE, RMSE and MedAE) results for this study are plotted in the figures of Fig. 7, Fig. 8 and Fig. 9. It is evident from the three figures that there is an overall enhancement in cellular MT localization when increasing the inter-element spacing distance among

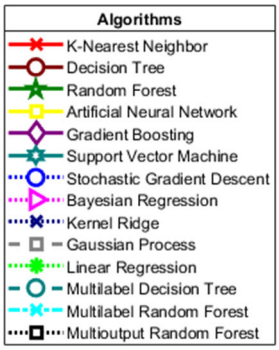

FIGURE 6. Common legend for considered ML algorithms in all figures.

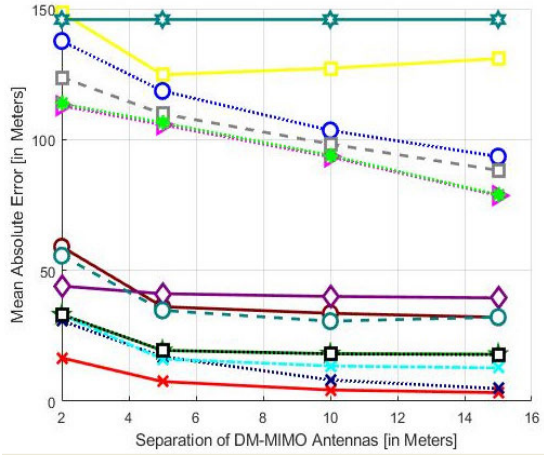

FIGURE 7. Effect of antenna array inter-element separation distance on MT localization accuracy using MAE.

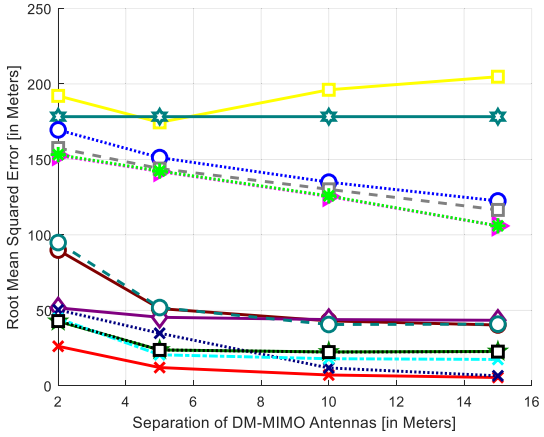

FIGURE 8. Effect of antenna array inter-element separation distance on MT localization accuracy using RMSE.

DM-MIMO antennas irrespective of the utilized ML algorithm except for the ANN.

The best MedAE attained is 1.88 m precision of predicting MT locations at 15 m DM-MIMO spacing was realized using the KNN algorithm. This is due to the fact that the KNN algorithm logic is well suited for the nature of the MT RSS fingerprints in this research works. This due to the fact that the logic of KNN says in simple terms that similar things exist in the vicinity of each other, which holds true in the case of RSS at a specific location to follow closely in value the RSS at a location just next to it.

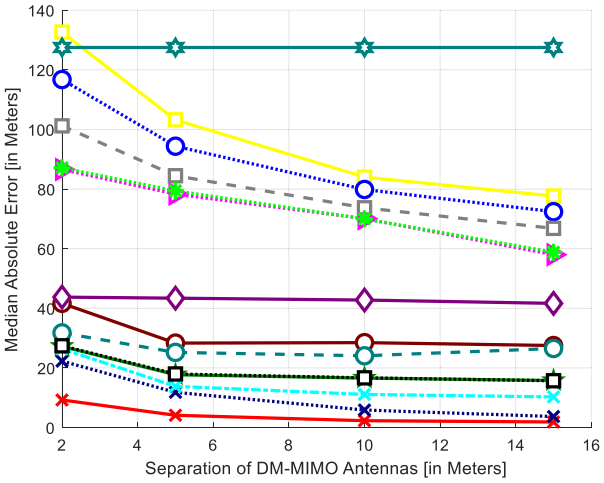

FIGURE 9. Effect of antenna array inter-element separation distance on MT localization accuracy using MedAE.

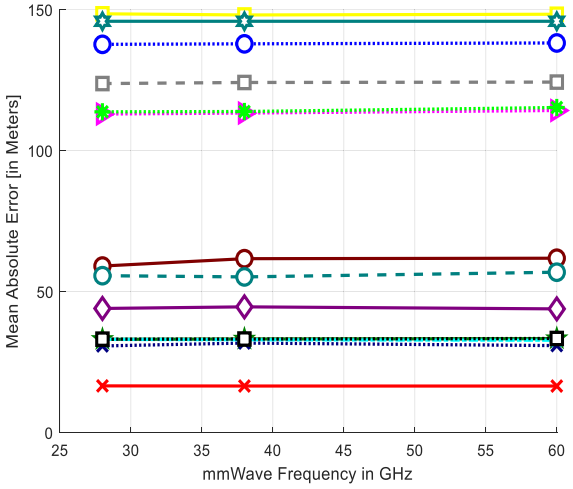

FIGURE 10. Effect of changing the mm-Wave frequency of operation on localization accuracy in MAE.

Moreover, it is demonstrated that this trend in obtained results is consistent in all three performance metrics. Therefore, we will show only the MAE performance metric in the following investigations.

Due to the resulting large size of common legends of the figures that contain the generated curves associated with the ML algorithms and topologies, we depicted the common legend in Fig. 6 for the aim of clarity in the subsequent Fig. 7 through Fig. 12.

C. COMPARATIVE STUDY ON THE EFFECT OF CHANGING FREQUENCY OF OPERATION ON LOCALIZATION ACCURACY

The DM-MIMO system was simulated with three different mm-Wave operation frequencies (28 GHz, 38 GHz, and 60 GHz) to investigate the MT Localization accuracy at the selected frequency band. To ensure a fair test, the following variables were fixed for the four network environments. The common simulations parameters are listed in Table 3. A plot of the MAE performance metric is shown Fig. 10. As illustrated in the figure, varying the frequency in the mm-Wave range didn't affect the ML algorithms performance in predicting MT locations, noticeably. That is in agreement with the discussion presented in the literature [48]–[50], that the CI propagation model measured path loss exhibits little dependence on frequency in outdoor environments. The KNN kept providing best obtained results in location prediction of the MTs, followed by kernel Ridge and Random Forest multi-out estimator.

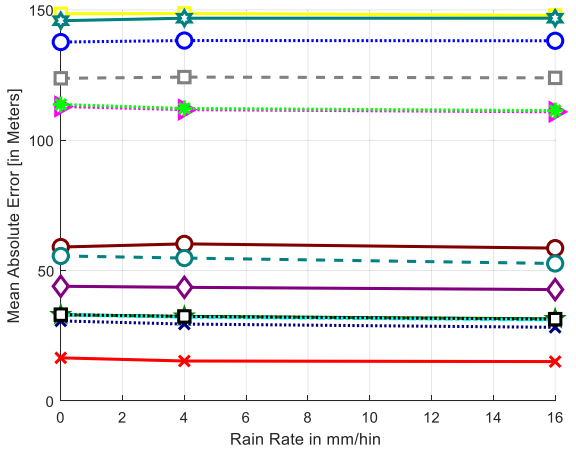

FIGURE 11. Effect of Rain on MT localization accuracy in MAE.

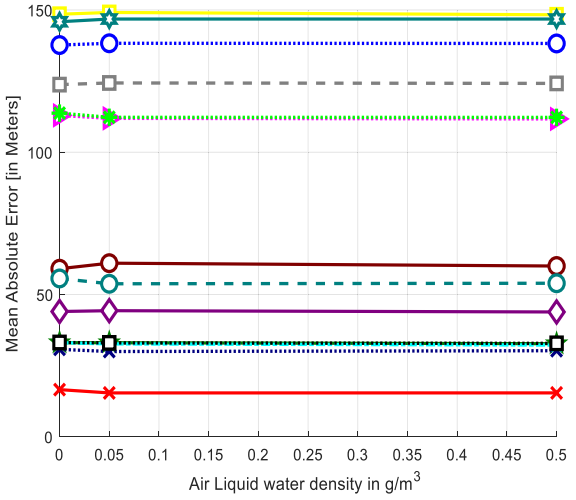

FIGURE 12. Effect of Fog on MT localization accuracy in MAE.

D. COMPARATIVE STUDY ON THE EFFECT OF VARYING WEATHER CONDITIONS ON LOCALIZATION PERFORMANCE

Localization of MTs performance is analyzed under varying weather conditions. The weather conditions included in the investigation are clear skies weather, fog and rain. The ML algorithms were trained using clear skies weather, and then tested using MT RSS fingerprints under clear skies weather (reference), under rainy weather (moderate and heavy) RSS fingerprints, and foggy (moderate and heavy) RSS fingerprints. Thus, our simulation of the DM-MIMO service area was run with the corresponding weather condition configuration to produce the respective dataset. In foggy weather conditions, air liquid water density of 0.005 g/m^3 reading is

TABLE 3. Common simulations parameters.

Parameters	Values
Location	PSUT region
Fingerprints count	7600
Fingerprints resolution	5 m
Antenna separation	2 m
Antennas count	64
Propagation Model	Close-in
Weather	Clear Skies

TABLE 4. Common simulations parameters.

Parameters	Values
Frequency	28GHz
Fingerprints count	7600
Fingerprints resolution	5 meters
Antennas resolution	2 meters
Antennas count	64
Propagation Model	Close-in + Rain/Fog
Weather	Clear skies/Rain/Fog

used for medium fog, and 0.5 g/m^3 is used for heavy fog. As for rainy weather conditions, medium rain is measured at a rain rate of 4 mm/h, and heavy rain is measured at 16 mm/h. The common simulations parameters are listed in Table 4.

The results for the localization performance in rainy days are shown in Fig. 11, while the effect of fog (moderate and heavy) on localization accuracy is illustarted in Fig 12.

In Fig. 11 and Fig. 12, the consistent trend and steady performance of MT location prediction of the considerd ML algorithms is confirmed using the proposed DM-MIMO system. It is also evident that MT location prediction accuarcy is unaffected by variant weather conditions. The KNN algorithm enjoys best MT localization accuatrcy desipite the differing weather conditions.

E. COMPARATIVE STUDY OF THE EFFECT OF CHANGING MT FINGERPRINTS LOCATIONS GRID RESOLUTION ON MT LOCALIZATION PERFORMANCE

Performance of location estimation of MTs is studied with respect to varying the grid resolution of MT RSS fingerprints. The grid resolutions considered are 2 m, 5 m, 15 m. The common simulations parameters are listed in Table 5. The obtained results are illustrated in Fig. 13, which verify the expected enhancement in MT localization accuracy when the fingerprints grid resolution is higher (i.e., distance separation is smaller at 2 m between a pair of fingerprints). The KNN algorithm maintained the best prediction accuracy in this study.

F. INVESTIGATION INTO THE EFFECT OF REDUCING THE SPLITTING RATIO OF TESTING AND TRAINING DATA SIZE ON LOCALIZATION ACCURACY

We investigate the effect of changing the datasets splitting ratio of the testing data to the remaining training data of the

TABLE 5. Common simulation parameters.

Parameters	Values
Frequency	28GHz
Fingerprints count	7600
Antennas spacing	2 m
Antennas count	64
Propagation Model	Close-in
Location	PSUT region
Weather	Clear skies

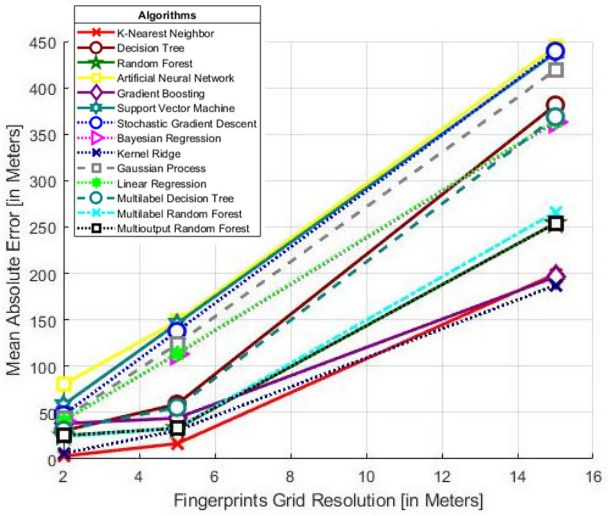

FIGURE 13. Effect of changing the fingerprinting grid resolution on localization accuracy in MAE.

TABLE 6. Common simulation parameters.

Parameters	Values
Frequency	28GHz
Fingerprints count	7600
Fingerprints resolution	5 meters
Antennas separation	2 meters
Antennas count	64
Propagation Model	Close-in
Weather	Clear skies

datasets studied. To that end, we evaluated the performance of MT localization in our DM-MIMO using the common simulation parameters of Table 6, while we change the training data size fed to the ML algorithms. The splitting ratios considered are as follows: a) testing data 10% and training data 90% b) testing data 20% and training data 80%. c) testing data 30% and training data 70%. d) testing data 50% and training data 50%. e) testing data 70% and training data 30%. In Fig. 14., the obtained MT localization performance expressed in terms of MAE is depicted for each of the ML algorithms when varying the splitting ratio. The figure demonstrates that for most of the ML estimators, MT localization performance is not significantly affected when the training data percentages were reduced incrementally from at 90% to 70% of the total of 7600 instances in each of the datasets. Then it exhibits a continued degradation in performance towards the 30%

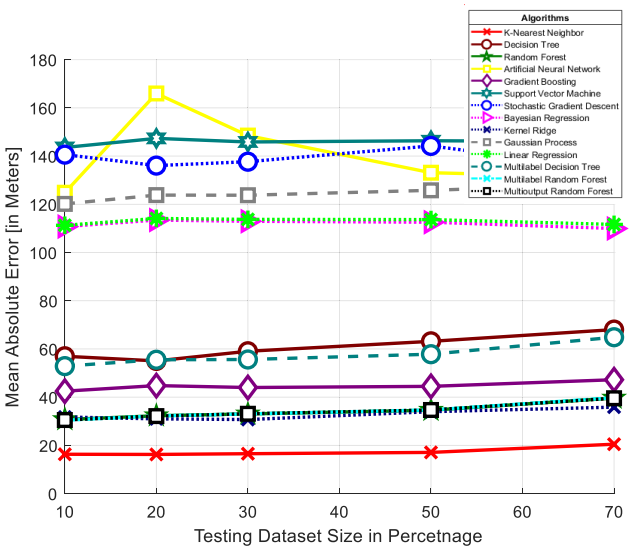

FIGURE 14. Effect of reducing the splitting ratio of testing to training data size on localization accuracy in MAE.

training ration only of the whole dataset provided. Most algorithms performance was stable, while KNN had optimal performance in predicting locations of MT in all considered different dataset splitting ratios.

V. CONCLUSION

In this work, we investigated MT localization in DM-MIMO systems based on the RSS fingerprinting approach in conjunction with ML algorithms utilizing Apache Spark big data computational platform, to be employed in microcells urban environments. We evaluated the performance of 13 ML algorithms (including multi-target algorithms) for this task in variant DM-MIMO wireless system configurations and conditions. The RSS fingerprints from the single-antenna MT collected at each of the receive antenna elements of the DM-MIMO base station. Practical mm-Wave signal propagation models suited for 5G wireless systems we incorporated in the numerical simulations that used ray-tracing techniques in conjunction with the 3D OpenStreetMap to mimic real-life environment propagation conditions. Several DM-MIMO system topologies and configuration parameters combinations affecting MT localization were investigated to analyze performance. Produced results demonstrated that the location of a MT could be effectively predicted by means of a subset of considered ML algorithms. Performance evaluation metrics served to identify optimum ML algorithm and methodology for utilization in DM-MIMO systems. It was evident that the KNN was the leading ML algorithm performer within the framework of this comparative analysis, followed by the KRR and RF in all investigated scenarios. For future investigations within this framework, the following factors would be worth consideration and further analysis: (a) the advantage (if any) of using other DM-MIMO systems antenna array geometries and study their effect on MT localization. (b) the effect of different geographic areas and topography on localization accuracy and the size of training fingerprints datasets.

ACKNOWLEDGMENT

The authors acknowledge that this research work was completed as part of the thesis requirements for the degree of M.Sc. in electrical engineering during the studies of Mrs. Walaa Y. Al-Rashdan at Princess Sumaya University for Technology.

REFERENCES

[1] Y. Liu, X. Shi, S. He, and Z. Shi, "Prospective positioning architecture and technologies in 5G networks," *IEEE Netw.*, vol. 31, no. 6, pp. 115–121, Nov./Dec. 2017.

[2] A. H. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based wireless location: Challenges faced in developing techniques for accurate wireless location information," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 24–40, Jul. 2005.

[3] A. Tahat, G. Kaddoum, S. Yousefi, S. Valaee, and F. Gagnon, "A look at the recent wireless positioning techniques with a focus on algorithms for moving receivers," *IEEE Access*, vol. 4, pp. 6652–6680, 2016.

[4] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.

[5] W. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 106–112, Apr. 2014.

[6] F. Rusek, D. Persson, B. Kiong Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[7] V. Savic and E. G. Larsson, "Fingerprinting-based positioning in distributed massive MIMO systems," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–5.

[8] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Machine learning methods for RSS-based user positioning in distributed massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8402–8417, Dec. 2018.

[9] C. Gao, G. Zhao, and H. Fourati, *Cooperative Localization and Navigation: Theory, Research, and Practice*, 1st ed. Boca Raton, FL, USA: CRC Press, 2019.

[10] M. Schüssel, "Angle of arrival estimation using WiFi and smartphones," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, 2016, pp. 1–2.

[11] J. Larsson, "Distance estimation and positioning based on Bluetooth low energy technology," Student thesis, TRITA-ICT-EX, Inf. Commun. Technol., Kth Roy. Inst. Technol., Stockholm, Sweden, 2015, p. 27.

[12] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich, and M. Coulon, "Direct localization for massive MIMO," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2475–2487, May 2017.

[13] A. Hu, T. Lv, H. Gao, Z. Zhang, and S. Yang, "An ESPRIT-based approach for 2-D localization of incoherently distributed sources in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 996–1011, Oct. 2014.

[14] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "5G position and orientation estimation through millimeter wave MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[15] A. Guerra, F. Guidi, and D. Dardari, "Position and orientation error bound for wideband massive antenna arrays," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 853–858.

[16] T. Wei, A. Zhou, and X. Zhang, "Facilitating robust 60 GHz network deployment by sensing ambient reflectors," in *Proc. 14th USENIX Symp. Netw. Syst. Design Implement (NSDI)*, Mar. 2017, pp. 213–226.

[17] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[18] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[19] P. Harris, S. Zang, A. Nix, M. Beach, S. Armour, and A. Doufexi, "A distributed massive MIMO testbed to assess real-world performance and feasibility," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–2.
[20] J. Wang and L. Dai, "Asymptotic rate analysis of downlink multi-user systems with co-located and distributed antennas," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3046–3058, Jun. 2015.
[21] S. Sun, T. S. Rappaport, T. A. Thomas, A. Ghosh, H. C. Nguyen, I. Z. Kovacs, I. Rodriguez, O. Koymen, and A. Partyka, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2843–2860, May 2016.
[22] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, and J. Jarvelainen, "Propagation path loss models for 5G urban micro- and macro-cellular scenarios," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–6.
[23] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[24] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[25] The Apache Software Foundation. *Apache Spark*. Accessed: May 19, 2020. [Online]. Available: https://spark.apache.org/
[26] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory, and Methods*. Hoboken, NJ, USA: Wiley, 1998.
[27] N. S. Altman, "An introduction to kernel and nearest-neighbor non-parametric regression," *Amer. Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992.
[28] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, "An empirical study of distance metrics for K-nearest neighbor algorithm," in *Proc. 2nd Int. Conf. Ind. Appl. Eng.*, 2015, pp. 1–6.
[29] X. Niuniu and L. Yuxun, "Review of decision trees," in *Proc. 3rd Int. Conf. Comput. Sci. Inf. Technol.*, 2010, pp. 105–109.
[30] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[31] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 5, no. 5, pp. 216–233, Sep. 2015.
[32] A. Dogan, D. Birant, and A. Kut, "Multi-target regression for quality prediction in a mining process," in *Proc. 4th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2019, pp. 639–644.
[33] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, 2008.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[35] X. Wu, Y. Gao, and D. Jiao, "Multi-label classification based on random forest algorithm for non-intrusive load monitoring system," *Processes*, vol. 7, no. 6, p. 337, Jun. 2019.
[36] F. Perez-Cruz, S. Van Vaerenbergh, J. J. Murillo-Fuentes, M. Lazaro-Gredilla, and I. Santamaria, "Gaussian processes for nonlinear signal processing: An overview of recent advances," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 40–50, Jul. 2013.
[37] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[38] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[39] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 9, 1997, pp. 155–161.
[40] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.
[41] T. C. Hsiang, "A Bayesian view on ridge regression," *Statistician*, vol. 24, no. 4, p. 267, Dec. 1975.
[42] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.
[43] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[44] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012, ch. 14.4.3, pp. 492–493.
[45] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.
[46] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.
[47] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.
[48] Y. Huo, X. Dong, and W. Xu, "5G cellular user equipment: From theory to practical hardware design," *IEEE Access*, vol. 5, pp. 13992–14010, Aug. 2017.
[49] *NR User Equipment (UE) Radio Transmission and Reception: Part 2: Range 2 Standalone*, document TS V1520, 3GPP, Jun. 2018.
[50] Y. Sano, S. Okuyama, N. Iizasa, T. Takada, K. Ando, and N. Fujimura, "5G radio performance and radio resource management specifications," *NTT DOCOMO Techn. J.*, vol. 20, no. 3, pp. 79–95, Jan. 2019.
[51] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdiscipl. J. Inf., Knowl., Manage.*, vol. 14, pp. 045-076, 2019, doi: 10.28945/4184.
[52] MathWorks, Natick, MA, USA. *MATLAB and Antenna Toolbox*. Accessed: May 16, 2020. [Online]. Available: https://www.mathworks.com/help/antenna/ug/urban-channel-link-analysis-and-visualization-using-ray-tracing.html
[53] *OpenStreetMap Contributors, Planet Dump*. Accessed: Dec. 2019. [Online]. Available: https://planet.osm.org and https://www.openstreetmap.org/#map=9/310447/35.9898
[54] Y. C. Ho and D. Pepyne, "Simple explanation of the no-free-lunch theorem and its implications," *J. Optim. Theory Appl.*, vol. 115, no. 3, pp. 549–570, 2002.

WALAA Y. AL-RASHDAN received the B.Sc. degree in computer engineering from the Jordan University of Science and Technology (JUST), Irbid, Jordan, and the M.Sc. degree in electrical engineering from the Princess Sumaya University of Technology (PSUT), Amman, Jordan. From 2004 to 2006, she was a Networks Engineer at Omnix International, Qatar. From 2006 to 2010, she was an Information Management Engineer at Schlumberger, Qatar. She is a member of the Jordan Engineers Association.

ASHRAF TAHAT (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Illinois Institute of Technology (IllinoisTech), Chicago, IL, USA, and the Ph.D. degree from IllinoisTech, in 2002, with a focus on communications and signal processing. He joined the Princess Sumaya University for Technology (PSUT), in 2005, where he participated in establishing the Department of Communications Engineering and also served as the Head of the Department, from 2010 to 2012. He was also a Visiting Professor with the Department of ECE, McGill University, Montreal, QC, Canada, where he conducted research on modern communications systems, from 2012 to 2013. From 2002 to 2003, he was an Adjunct Professor with IllinoisTech. Before completing his doctoral studies, he was with the Concept Center, Bell Labs-Lucent Technologies Inc. (Nokia Networks), working on first generation ADSL modem prototypes. He was also a Senior Staff Engineer with the Personal Communications Division, 3Com Corporation, IL, USA, from 1998 to 2000. He is currently an Associate Professor with the Department of Communications Engineering, PSUT. He is a Senior Member of the Eta Kappa Nu and the Tau Beta Pi honor societies. He is also the Vice-Chairman of the IEEE Jordan Section.