IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 21, 2019, accepted September 30, 2019, date of publication October 4, 2019,
date of current version October 16, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2945539*

# Handling Device Heterogeneity and Orientation Using Multistage Regression for GMM Based Localization in IoT Networks

**ANKUR PANDEY, (Student, IEEE), RAGHU VAMSI, AND SUDHIR KUMAR, (Member, IEEE)**
Indian Institute of Technology Patna, Patna 801106, India

Corresponding author: Ankur Pandey (1821ee12@iitp.ac.in)

This work was supported in part by the Department of Science and Technology, Government of India, under Grant NRDMS/UG/S.Kumar/IIT Patna/e-04/2019.

**ABSTRACT** Location estimation of heterogeneous smart devices is needed for the Internet of Things (IoT) based location services. Device orientation and heterogeneity are the bottlenecks in accurate location estimation, which are not addressed together in the existing methods. Also, most of the state-of-the-art Received Signal Strength (RSS) based localization methods consider a single Gaussian model instead of a mixture of Gaussians. In this paper, we propose to solve both these issues with a combination of multistage linear regression and Gaussian Mixture Model (GMM) method. Additionally, the proposed method detects the malicious data in the IoT network and estimates the location in case of sensor faults. The performance of the proposed method is tested using Wi-Fi signals in an indoor environment.

**INDEX TERMS** Device heterogeneity, device orientation, localization, IoT.

## I. INTRODUCTION

Cloud-based IoT location services find applications in areas such as health, security, transport, automation, weather, and agriculture monitoring contributing to the implementation of smart cities [1]–[7]. Global Positioning System (GPS) enabled systems are viable solutions for outdoor localization, but for indoor positioning, the localization accuracy degrades. It is because of the unavailability of the satellite signals and Non-line-of-sight (NLOS) propagation of radio signals. Additionally, GPS consumes high power and is costly to equip in every smart device [8]. Hence, we need methods for indoor localization, which use Wi-Fi, Bluetooth, geomagnetic, visible light, acoustic, FM radio, and RFID signals due to their pervasive nature in an indoor environment [9].

### A. MOTIVATION AND RELATED WORK

RSS based localization methods do not require any separate infrastructure or hardware equipment. They use already existing signals such as Wi-Fi or Bluetooth to name a few for location estimation, thereby reducing the cost [8]–[13]. The heterogeneity of smart devices in the IoT network affect the indoor localization as discussed in [14]–[17]. Most of the state-of-the-art localization methods assume that the homogeneous devices are used for constructing the offline fingerprinting map as well as the online location estimation. In the IoT environment, heterogeneous devices record different RSS values at the same location, thereby providing high localization error. Orientation diversity is another issue which increases the localization error due to the orientation mismatch during offline and online phases. The human body also reflects and blocks the Wi-Fi signals [18]. Hence, users holding the smart device closer to the body in different orientations, cause a difference in the Wi-Fi RSS measurements. Also, a variation of 2.5-10 dBm is observed when a user changes its direction from facing the access point. Most of the state-of-the-art localization methods address device heterogeneity, but there are limited works on orientation issues [19]–[23]. The existing orientation compensation techniques are mainly based on fingerprint matching and do not establish any relationship among the orientations. Therefore, these methods are prone to performance degradation if the RSS data from exact test orientations do not exist in the offline database. Hence, the orientations and heterogeneity of the smart devices create RSS diversity, which is a bottleneck for localization in IoT networks.

In order to address these issues, various calibration methods are proposed such as least squares, histogram equalization, machine learning and navigation states from dead

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

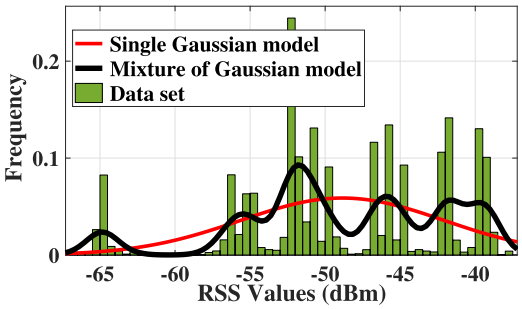


**FIGURE 1.** Wi-Fi RSS distribution following mixture of Gaussians instead of single Gaussian.

reckoning (DR) [14], [20], [24]–[26]. In the literature, several calibration-free methods such as differential RSS (DRSS), pairwise differential, ZU-mean, Mean Differential Fingerprinting (MDF) and regression [15], [16], [27]–[29] are also proposed. However, as discussed in [30], most of these methods suffer from RSS variations due to variable power of access points even after considering the ratio or difference of RSS values. Another issue with these methods is that they do not address both the heterogeneity and orientation issues together. In DR based methods, the accumulated error in sensor measurements of human activities further increases the localization error. Further, the existing methods do not provide any relationship between the RSS of heterogeneous smart devices with different orientations. In order to address the device heterogeneity, transformation methods such as [14], [15], [31], [32] are also proposed. A drawback of these methods is the requirement of the prior knowledge of the device type. The device type information is utilized to develop a pairwise relationship (only two devices at a time). Hence, at a time, only two devices are used, which increases the storage demand of the smart device. For cases where the device type is unknown, the method fails. Additionally, these transformation methods do not address the orientation diversity issue.

Most of the discussed RSS based localization methods assume a single Gaussian distribution of RSS data. The state-of-the-art RSS localization methods [33]–[37] show that a single Gaussian distribution can not model the RSS data. It is because of the multipath, device heterogeneity, and device orientation issues in the indoor environment. However, the RSS data distribution follows a mixture of Gaussian [38], which is also evident from Fig. 1 for the RSS data observed in our experimental setup. Therefore, single Gaussian can not provide accurate location estimation, and we need a multi Gaussian based method for localization [39]. Now, the fingerprinting pattern matching can be performed with two methods, namely probabilistic and non-probabilistic methods [12], [13]. Further, the probabilistic pattern matching algorithms utilize a RSS probability distribution function from the training RSS data, and then the Maximum Likelihood (ML) method is used for smart device localization. On the contrary, the non-probabilistic methods do not require creating a parametric model and are suitable for an indoor IoT environment as it has small training sets [40], [41].

Finally, the RSS values can be affected by the malicious and erroneous smart devices in the IoT network [42]–[45]. Most of the existing methods that address device heterogeneity and orientation do not test the localization performance in the presence of erroneous RSS values. In this paper, it is shown that the proposed method's performance is robust to these faults and the multiple Gaussian method outperforms the single Gaussian method in terms of the localization error.

### B. CONTRIBUTIONS

In this paper, a localization system using multistage regression and GMM is proposed, that addresses the device heterogeneity and orientation diversity issues. Additionally, the method is robust to erroneous RSS data. The main contributions of this paper are summarized as follows

1) A multistage regression method that addresses the device heterogeneity and orientation issues for smart device localization in an indoor loT environment.
2) Investigating the effect of the choice of the dependent variables, that is, device type and orientation, for a regression method to compensate the device heterogeneity and orientation effect.
3) Leveraging multiple Gaussian instead of a single Gaussian method resulting in high localization accuracy.
4) The proposed method is robust to malicious RSS data.

The remainder of the paper is organized as follows: Section II presents the proposed multistage regression and GMM localization method to address both the device heterogeneity and orientation issues. Section III describes the experimental setup and results of the proposed method. Finally, Section IV concludes the paper with future research challenges.

## II. GMLOC: MULTISTAGE REGRESSION METHOD FOR SMART DEVICE LOCALIZATION

In the first stage, the data is collected and stored at a cloud server where the device information such as Media Access Control (MAC) address is available. Fig. 2 explains the overview of the proposed method for location estimation in an indoor environment. Missing RSS values due to hardware issues or connectivity failure can lead to a high localization error. We address this issue by estimating the missing values using a regression method. Further, a two-stage regression is performed on orientations and smart devices respectively to obtain a transformed RSS vector. Finally, the GMM parameters are estimated using the transformed RSS vectors for localization. The GMLoc method is explained in detail in the ensuing subsections.

### A. HANDLING DEVICE HETEROGENEITY AND ORIENTATION ISSUES

#### 1) DEVICE HETEROGENEITY

Consider a Redmi Note 5 Pro for building the offline fingerprinting map, and in the online phase, the user has

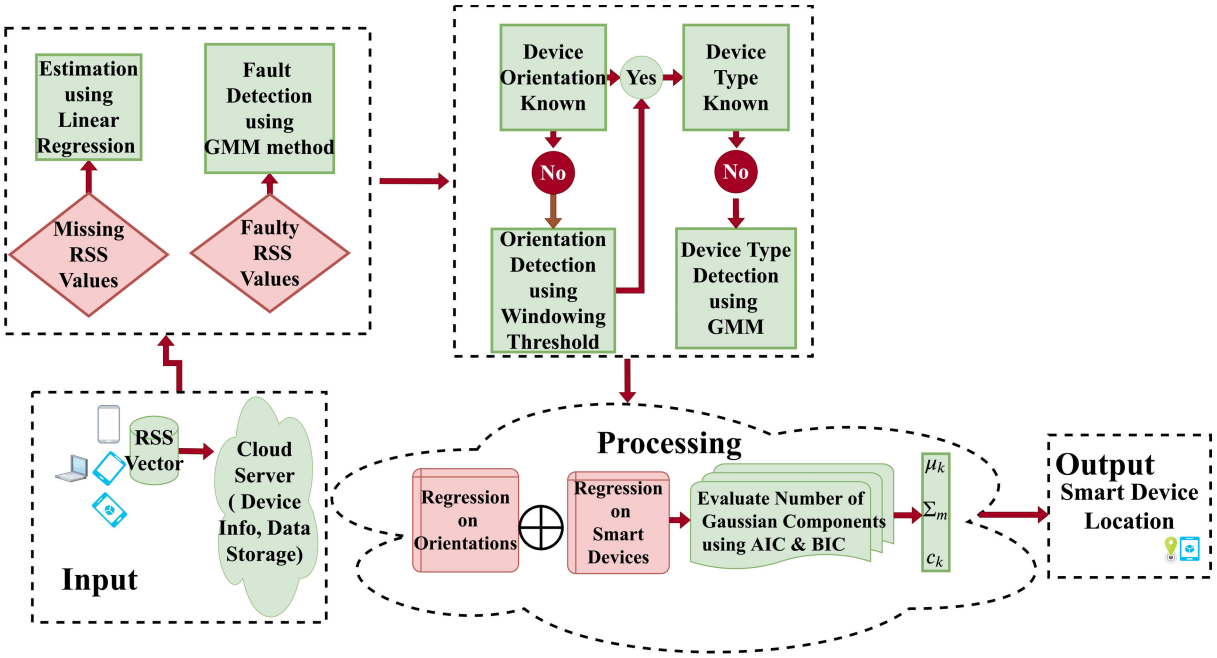


**FIGURE 2.** Overview of the proposed indoor localization method when device types and multiple orientations are known/unknown at the cloud server. GMM = Gaussian Mixture Model, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

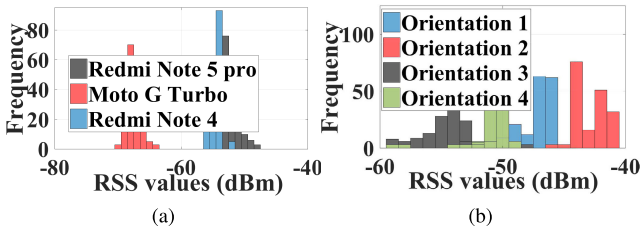


**FIGURE 3.** Histogram of RSS data for different (a) smart devices. (b) orientations.

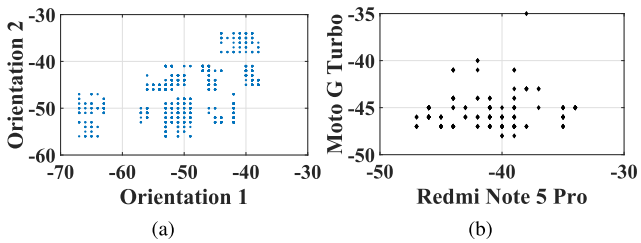


**FIGURE 4.** RSS data between two different (a) orientations. (b) smart devices.

Moto G Turbo. We demonstrate the device diversity challenge using recorded experimental dataset, as shown in Fig. 3(a). It is observed that at a particular location, in terms of variance, there is a difference of approximately 14 dBm from one smart device to another. Hence, it increases the localization error. Also, signal processing operations such as shifting and scaling do not provide higher localization accuracy as smart devices of the same characteristic also perform differently [46]. The primary reasons for this variation are different receiver antenna designs of variable size, chip design materials with different absorption coefficients for RF signals, linear and non-linear receiver circuit design and support for a different number of frequency bands [47].

### 2) ORIENTATION DIVERSITY

The RSS data is much affected with device orientations also. We collected the RSS data in four orientations, namely, 0°, 90°, 180° and 270° at the same location. Fig. 3(b) shows the histogram for four different orientations. A maximum variance of 13 dBm is observed between multiple orientations. This variation leads to a higher localization error. It is also discussed in [48], [49] that there is a variation of approximately $2.5 - 7.6$ dBm due to the change in the orientation of the smart device. In order to develop a relationship between multiple orientations and devices' RSS data, we use a two-stage multiple regression method instead of the regression technique used in [15] that considers only two smart devices at a time. The advantage of multiple regression is that it models the relationship between many independent variables and a dependent variable by linear modeling of the observed RSS data.

### 3) MULTISTAGE LINEAR REGRESSION METHOD

We establish a relationship to address both the device heterogeneity and orientations. It is observed that there exist a linear pattern between multiple orientations as well as heterogeneous smart devices. Fig. 4 shows the RSS values of one orientation with another and similarly one smart device

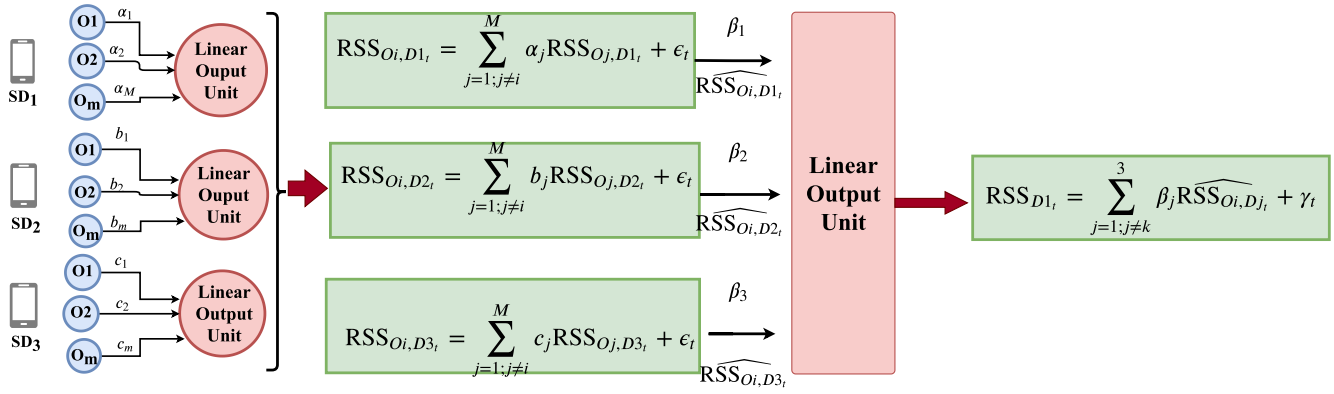


**FIGURE 5.** Multistage linear regression method to compensate device heterogeneity and orientation.

with another. A linear relationship is observed through these plots.

Fig. 5 shows the proposed two-stage linear regression method. Let us consider three smart devices for illustration. It can be observed from Fig. 5 that the smart devices $SD_1$, $SD_2$ and $SD_3$ have multiple orientations $O_1, O_2, \ldots, O_m$ each respectively. In the first stage, a relationship is derived using regression between RSS of multiple orientations to generate a RSS vector for the three smart devices represented as $\text{RSS}_{O1,D1}$, $\text{RSS}_{O1,D2}$ and $\text{RSS}_{O1,D3}$ respectively, considering orientation 1 as dependent variable. The RSS vector $\text{RSS}_{O1,D1}$ for first smart device with $T$ samples can be obtained from Equation 1.

$$\begin{aligned}
&\mathbf{Y}_{T\times 1} = \mathbf{X}_{T\times M} \times \alpha_{M\times 1} + \epsilon_{T\times 1} \\
&\mathbf{Y}_{T\times 1} = \begin{bmatrix}\text{RSS}_{O1,D1_1} & \cdots & \text{RSS}_{O1,D1_T}\end{bmatrix}^T \\
&\mathbf{X}_{T\times M} = \begin{bmatrix}1 & \text{RSS}_{O2,D1_1} & \cdots & \text{RSS}_{O(M-1),D1_1} \\ 1 & \text{RSS}_{O2,D1_2} & \ldots & \text{RSS}_{O(M-1),D1_2} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & \text{RSS}_{O2,D1_T} & & \text{RSS}_{O(M-1),D1_T}\end{bmatrix} \\
&\boldsymbol{\alpha}_{M\times 1} = \begin{bmatrix}\alpha_0 \ \ldots \ \alpha_{(M-1)}\end{bmatrix}^T; \quad \boldsymbol{\epsilon}_{T\times 1} = \begin{bmatrix}\epsilon_0 \ \ldots \ \epsilon_T\end{bmatrix}^T \qquad (1)
\end{aligned}$$

Further, Equation 2 is the generalization of Equation 1 where $\text{RSS}_{Oi,Dk_t}$ represents RSS value of the $i$th ($i \in \{1, 2, \cdots, M\}$) orientation selected as dependent variable from $k$th smart device and $t \in \{1, 2, \cdots, T\}$ represents the index of the set of $T$ samples of the $i$th orientation where $T > M$. $k \in \{1, 2, \cdots, K\}$ represents the index of the total $K$ smart devices with $T > K$. $\alpha_j$ denotes the regression coefficient for $j$th orientation and $\epsilon$ is the residual term with distribution as $\mathcal{N}(0, \sigma^2)$, representing the difference between the observed and modeled RSS observation.

$$\text{RSS}_{Oi,Dk_t} = \sum_{j=1;j\neq i}^{M} \alpha_j \text{RSS}_{Oj,Dk_t} + \epsilon_t \qquad (2)$$

The regression coefficient using least square method can be estimated as
$\hat{\alpha} = \left(\text{RSS}_{Oj,Dk_t}{}^T \ \text{RSS}_{Oj,Dk_t}\right)^{-1} \text{RSS}_{Oj,Dk_t}{}^T \ \text{RSS}_{Oi,Dk_t}$ [50].

Therefore, we estimate $\widehat{\text{RSS}}_{Oi,Dk_t}$ using the estimated regression coefficient vector $\hat{\alpha}$.

Further, in the second stage the device heterogeneity is addressed using Equation 3 with the estimated orientation RSS vector $\widehat{\text{RSS}}_{Oi,Dk_t}$ obtained at the first stage as

$$\text{RSS}_{Dk_t} = \sum_{j=1;j\neq k}^{K} \beta_j \widehat{\text{RSS}}_{Oi,Dj_t} + \gamma_t \qquad (3)$$

Here, $\text{RSS}_{Dk_t}$ is the $t$th RSS value of the $k$th smart device selected as dependent variable. $\beta_j$ denotes the regression coefficient for $j$th smart device and $\gamma$ is the residual term with distribution as $\mathcal{N}(0, \sigma^2)$ representing the difference between the observed and linear modeling of RSS data. We estimate the regression coefficient of smart devices $\hat{\beta}$ using the least square method discussed before. Finally, the estimated RSS vector $\widehat{\text{RSS}}_{Dk_t}$ is obtained using the estimated regression coefficient for smart devices given as $\hat{\beta} = \left(\widehat{\text{RSS}}_{Oi,Dj_t}{}^T \ \widehat{\text{RSS}}_{Oi,Dj_t}\right)^{-1} \widehat{\text{RSS}}_{Oi,Dj_t}{}^T \ \text{RSS}_{Dk_t}$

#### 4) SELECTION OF DEPENDENT VARIABLE FOR MULTISTAGE REGRESSION

In order to improve the localization accuracy, we choose the orientation and smart device as the dependent variable. Therefore, the RMSE is evaluated between $\widehat{\text{RSS}}_{Oi,Dk_t}$ and $\text{RSS}_{Oi,Dk_t}$ for $i$th orientation as $\sqrt{\frac{\sum_{t=1}^{T}\left(\widehat{\text{RSS}}_{Oi,Dk_t} - \text{RSS}_{Oi,Dk_t}\right)^2}{T}}$. The orientation for which least RMSE is obtained is chosen as the dependent variable. Same procedure is followed for choosing the smart device as dependent variable by computing the RMSE between $\widehat{\text{RSS}}_{Dk_t}$ and $\text{RSS}_{Dk_t}$. Algorithm 1 explains the proposed multistage regression method for handling the device heterogeneity and orientation issues.

### B. FINGERPRINTING BASED LOCALIZATION

Multipath propagation and shadow fading effects occur due to the human movements in an indoor environment along with the reflections and scattering due to the furniture and

**Algorithm 1** Multistage Regression Method to Compensate Heterogeneity and Orientation Effect

**Data**: $\text{RSS}_{O1,D1_t}, \text{RSS}_{O2,D1_t}, \cdots, \text{RSS}_{OM,D1_t}$, $\text{RSS}_{O1,D2_t}, \cdots, \text{RSS}_{OM,DK_t}$
**Result**: $\widehat{\text{RSS}_{Dk_t}}$
**begin**
1) Estimate the missing data using linear regression method for a particular smart device and orientation if any.
2) Evaluate $\widehat{\text{RSS}_{Oi,Dk_t}}$ using the estimated regression coefficient $\hat{\alpha}$ in Equation 2.
3) Choose the appropriate orientation as dependent variable based on least RMSE by comparing $\widehat{\text{RSS}_{Oi,Dk_t}}$ with $\text{RSS}_{Oi,Dk_t}$.
4) Compute $\widehat{\text{RSS}_{Dk_t}}$ using $\hat{\beta}$ in Equation 3.
5) Choose the smart device as dependent variable based on least RMSE by comparing $\widehat{\text{RSS}_{Dk_t}}$ with $\text{RSS}_{Dk_t}$.

**end**

building structure. It affects the performance of location estimation methods [1], [2], [8], [9]. There are two phases in the fingerprinting method, namely, the offline training and online location estimation phases. The user records RSS data from the multiple training locations, and then the online phase compares the RSS data of an unknown location to the recorded fingerprinting map, thereby estimating the unknown location using the nearest fingerprint [12].

$$\hat{l} = \arg\max_{l} P\left(\mathbf{RSS}_{\text{train}_l}|\mathbf{RSS}_{\text{test}}\right) = \arg\max_{l} P\left(\mathbf{RSS}_{\text{train}_l}\right) \prod_{k=1}^{K} P\left(\mathbf{RSS}_{\text{test}_k}|\mathbf{RSS}_{\text{train}_{k,l}}\right) \quad (4)$$

where $\hat{l}$ is the estimated location corresponding to the $\mathbf{RSS}_{\text{test}}$, $\mathbf{RSS}_{\text{train}_l}$ represents the offline training RSS values at location $l$, $\mathbf{RSS}_{\text{train}_{k,l}}$ denotes the offline training RSS values from the $k$th smart device at location $l$, $P(.)$ represents the probability density function. Further, as we use the Gaussian method for fingerprinting, $P\left(\mathbf{RSS}_{\text{test}_k}|\mathbf{RSS}_{\text{train}_{k,l}}\right) = \frac{1}{\sqrt{2\pi\sigma_{k,l}^2}} \exp\left(-\frac{\left(\mathbf{RSS}_{\text{test}_k}-\mu_{k,l}\right)^2}{2\sigma_{k,l}^2}\right)$ where $\mu_{k,l}$ and $\sigma_{k,l}^2$ are the mean and variance of RSS values from the $k$th smart device at the particular location $l$ in the offline database.

Therefore, to analyze the effect of device diversity on this fingerprinting method, it is considered that at the same location there are two sets of test vectors $\mathbf{RSS}_{\text{test}_{1,l}}$ and $\mathbf{RSS}_{\text{test}_{2,l}}$, from two smart devices at location $l$. We denote the device diversity effect as $RSS_{1,2,l}$ arising because of the two smart devices' RSS vectors at the same location $l$. The new estimated location index using this device diversity is denoted as $\hat{l}^{\star}$ and is represented by the following equation

$$\hat{l}^{\star} = \arg\max_{l} P\left(\mathbf{RSS}_{\text{train}_{1,l}}\right) \times \prod_{j=1}^{K} P\left(\left(\mathbf{RSS}_{\text{test}_{1,l}} + \mathbf{RSS}_{1,2,l}\right)|\mathbf{RSS}_{\text{train}_{1,l}}\right) \quad (5)$$

Hence, to compensate the bias $RSS_{1,2,l}$ because of the device heterogeneity, we use the regression method. The regression between the smart devices' RSS values $\mathbf{RSS}_{\text{test}_{1,l}}$ and $\mathbf{RSS}_{\text{test}_{2,l}}$ gives a single output $\mathbf{RSS}'_{\text{test}}$. Results show that using the $\mathbf{RSS}'_{\text{test}}$, localization error is reduced. Hence, the new location index $\hat{l^{\star\star}}$ using the regression RSS values can be computed as

$$\hat{l^{\star\star}} = \arg\max_{l} P\left(\mathbf{RSS}_{\text{train}_{1,l}}\right) \prod_{j=1}^{K} P\left(\left(\mathbf{RSS}'_{\text{test}_l}\right)|\mathbf{RSS}_{\text{train}_{1,l}}\right) \quad (6)$$

The results show that $\hat{l^{\star\star}}$ is close to the actual $\hat{l}$ as the diversity is addressed with the regression output $\mathbf{RSS}'_{\text{test}}$ in Equation 6. A similar analysis can be shown for orientation diversity affecting the estimated location index $\hat{l}$.

### C. GMM BASED LOCALIZATION

GMM, as compared to single Gaussian, is a better fit for the probability distribution modeling of the RSS values that contain a weighted mixture of Gaussian, as shown in Fig. 1. A GMM for the Wi-Fi RSS fingerprints, contains the superposition of $Q$ Gaussian densities [51] with $x$ as a $D$-dimensional RSS values given by $\sum_{m=1}^{Q} \frac{c_m}{\sqrt{(2\pi)^D \Sigma_m}} \exp\left(\frac{-1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)\right)$ where each Gaussian density is denoted as a component of the mixture with mean $\mu_m$ and covariance $\Sigma_m$. $c_m$, the mixing coefficient is given as $\sum_{m=1}^{Q} c_m = 1$, where $0 \leq c_m \leq 1$. The model parameter for location $l$ is then given by $\psi_l = \{c_{l,m}; \mu_{l,m}; \Sigma_{l,m}\}$. Therefore, using the Maximum Likelihood parameter estimation method, the model parameters $\hat{\psi}_l$ at a location $l$ are learned through the Equation 4. Further, for optimizing the parameters, the Expectation-Maximization (EM) algorithm is used [52]. The number of GMM components are estimated using AIC and BIC criteria using the Elbow method discussed next.

#### ELBOW METHOD FOR THE SELECTION OF NUMBER OF GAUSSIAN COMPONENTS USING AIC AND BIC

AIC provides the measure of goodness of fit of a statistical model such as GMM [53]. The AIC is defined as $\text{AIC} = 2h - 2\ln(L)$ where $h$ is the number of parameters and $L$ is the likelihood function. The model with the lowest AIC score is preferred. BIC helps to choose between two different models with different numbers of parameters by selecting the one which gives the lowest BIC score given as $\text{BIC} = \ln(N)h - 2\ln(L)$ where $N$ is the number of RSS data points. The elbow test is a heuristic method that helps in determining the values for AIC and BIC, after which the decrease in

**Algorithm 2** GMM Based Localization

**Data**: $\widehat{\text{RSS}_{Dk_t}}$ vector after multistage regression
**Result**: Estimated location $l$ of $\widehat{\text{RSS}_{Dk_t}}$ vector.
**begin**
- Input the $\widehat{\text{RSS}_{Dk_t}}$ vectors for training the GMM.
- Estimate the number of Gaussian components using the AIC and BIC criteria.
- Determine the GMM parameters $\{\mu_k, \Sigma_m, c_k\}$.
- Evaluate the probabilistic location for online RSS vectors using fingerprinting (Equation 6).
- **return** *Estimated location index l of the* $\widehat{RSS_{Dk_t}}$ *vector.*

**end**

values are small so that after this point adding another Gaussian component does not help in minimizing the AIC/BIC scores [54], [55]. Using the values of AIC and BIC, we find the number of Gaussian components in the mixture. Algorithm 2 summarizes the localization method after the orientation and device diversity are addressed through the regression method.

### D. LOCALIZATION WITH DIFFERENT USER CASES

The location of the smart devices is estimated for three possible user cases, as shown in Fig. 2. In the first case, both the orientation type, as well as device type or MAC ID, are known. Second, when only the device type is available but not the type of orientation. For scenarios where orientation is not known, we propose to detect the orientation leveraging the accelerometer sensor of the smart device along with a windowing threshold of RSS data. As the accelerometer sensor may provide erroneous orientation information due to the accumulation of gyroscope errors [56], we use windowing threshold of RSS data to confirm the orientation change. In this method, the moving average is computed using a sliding window, where the mean of the samples for a particular length of the sliding window is computed. The orientation change is detected using the moving average $\text{RSS}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\text{RSS}_i$, where $i = 1$ to $N$ (sliding window size). We define

$$\Delta\text{RSS}_i = \text{RSS}_{\text{avg}_i} - \text{RSS}_{\text{new}_i} \tag{7}$$

where $\text{RSS}_{\text{new}_i}$ is the next RSS value due to the new or existing orientation of the smart device for the sliding window after $N$ samples. We test the proposed orientation detection method with a sliding window size of the unit sample ($N = 1$) and then keep appending the detected orientations into the sliding window to create the RSS fingerprints. A threshold value of $\pm 4.5$ dB is computed from the offline database as most of the orientation change is observed for a minimum change of $\pm 4.5$ dB. Also, as discussed in [19], [49] due to change in the orientation the RSS values vary from $2.5 - 7.6$ dBm. Therefore, the detection $\Delta_d$ is defined as

$$\Delta_d = \begin{cases} 1 \rightarrow \text{ Orientation change} & \text{if } \Delta\text{RSS}_i \geq \text{ Th} \\ 0 \rightarrow \text{ No change} & \text{if } \Delta\text{RSS}_i < \text{ Th} \end{cases}$$

Hence, using the proposed method along with the accelerometer sensor information from the smart device provides an accurate information about the orientation change.

Finally, the proposed method is tested for the scenario where neither device type nor its orientation is known. In this case, the localization error is higher than the previous two cases. In order to reduce the localization error, device identification is performed using the GMM method. The RSS vector received at the smart devices is fed to the GMM, which classifies the input RSS vector received from different smart devices using mean, variance, and mixing coefficient as features corresponding to a particular smart device, as explained in Section II. Further, when the smart device is identified, the proposed orientation detection method is applied. Therefore, the localization error is reduced when both the device type and orientation are detected.

### E. ROBUSTNESS OF THE PROPOSED GMM METHOD

RSS values in a smart device are also affected by chipset degradation, noise in the IoT environment or router malfunctioning, and hence, this can affect the localization performance of the system. The RSS data may encounter an error in measurements such as precision degradation, offset, stuck-at-fault, and missing data [43]–[45], [57]. We analyze the performance of the proposed method for these type of faults. The first two faults can be modeled as $\text{RSS}_{\text{m}} = \text{RSS}_{\text{a}} + \beta + \eta$, where $\text{RSS}_{\text{m}}$ shows the measured RSS value (erroneous) from a smart device, $\text{RSS}_{\text{a}}$ is the actual RSS value from the smart device without any fault, $\beta$ is the additive offset constant and $\eta$ is the external noise. Precision degradation fault can occur because of the chipset failure in the smart device, thereby increasing the noise. This failure is modeled with $\beta = 0$ and the Gaussian noise parameter $\eta$ [45]. The offset fault arises due to the calibration errors in the sensors, thereby generating the sudden deviations from the normal data with an additive constant $\beta$. We model it with $\eta = 0$. Stuck-at-fault arises due to device failures, external attacks, or connection failures, and the device generates a constant reading with no variation over a period of time. Fig. 6 represents the type of faults induced in the normal RSS vectors to test the proposed GMLoc method. Finally, missing data is a common problem where, for a certain duration, no data is received in a network. This is mainly due to the packet drop or configuration mismatch. In this paper, the missing data issue is addressed using the same regression method, where a regression model is developed when the data is available (not missing). Hence, we obtain the regression model between RSS values' time index and corresponding RSS values. Therefore, when a missing value is detected at a particular time index, the previously obtained regression model is used to estimate the missing RSS data.

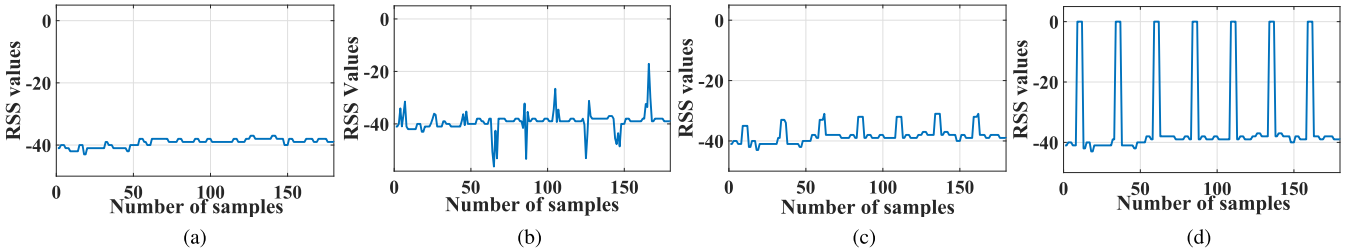


**FIGURE 6.** (a) Normal RSS vector received at a particular location. (b) Precision degradation. (c) Offset fault. (d) Stuck-at fault.

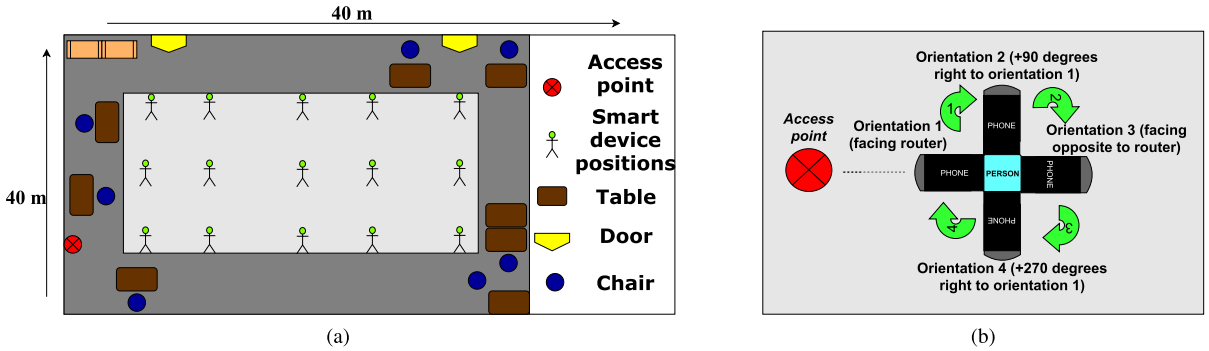


**FIGURE 7.** (a) The experimental setup for collecting Wi-Fi RSS from different devices and orientations. (b) Orientations observed from multiple smart devices.

Finally, localization is performed using the proposed method on the complete RSS data obtained after replacing the missing data values. It is observed that the proposed method is robust to such errors and can provide accurate location of the smart device even in such adverse scenarios.

## III. PERFORMANCE EVALUATION

In this section, we discuss the experimental setup for Wi-Fi RSS collection. The performance of GMLoc is discussed for both muliple and single Gaussian.

### A. DATASET PREPARATION

Fig. 7(a) shows the experimental setup for the data collection. The data is collected in an indoor environment considering reflections, scattering, and absorptions from furniture, walls, and human bodies. The setup is as per [58], because collections of RSS values for single grid points train the algorithm better and linear regression method can be used at a particular location. In order to consider the device heterogeneity, three different types of smartphones, namely Moto G Turbo, Xiaomi Note 5 Pro and Xiaomi Note 4 are used. Further, the RSS data is also collected in four orientations with 0° (facing towards the access point), 90°, 180° (facing opposite to the access point) and 270° [21] as shown in the Fig. 7(b). The other orientations can also be considered. An Android-based application is used to collect the data for each orientation. The experimental dataset consists of a RSS vector with parameters $T = 181$, $K = 3$ and $M = 4$.

### B. THE CROSS-VALIDATION METHOD AND PERFORMANCE METRIC

In order to evaluate the performance of the proposed method, the cross-validation method is used to calculate the localization error of the test RSS data. The collected RSS samples are divided into two parts. One part is used as the training set to learn the model parameters for the GMM localization method. The second part contains test RSS vectors from an unknown location whose location is to be estimated [59]. Therefore, for 10-fold cross-validation method, the Wi-Fi RSS dataset is partitioned into ten equal size samples. One part is used for validation purpose, and the rest nine parts are used for training. Further, this same procedure is repeated ten times randomly to cover the whole dataset. Finally, the results of all the folds are averaged. The advantage of the method is that it uses the complete dataset randomly for both the training and testing purpose and avoids overfitting.

The performance of any localization algorithm is based on the localization error, which is the Euclidean distance between the actual and the estimated smart device's location. Therefore, the localization error is calculated as Localization Error $= \sqrt{(Y_P - Y_A)^2 + (X_P - X_A)^2}$ where $(X_P, Y_P)$ are the estimated coordinates and $(X_A, Y_A)$ are the actual coordinates of the smart device location.

### C. EXPERIMENTAL RESULTS

This subsection presents the results for the estimated Gaussian parameters, the effect of regression on orientation

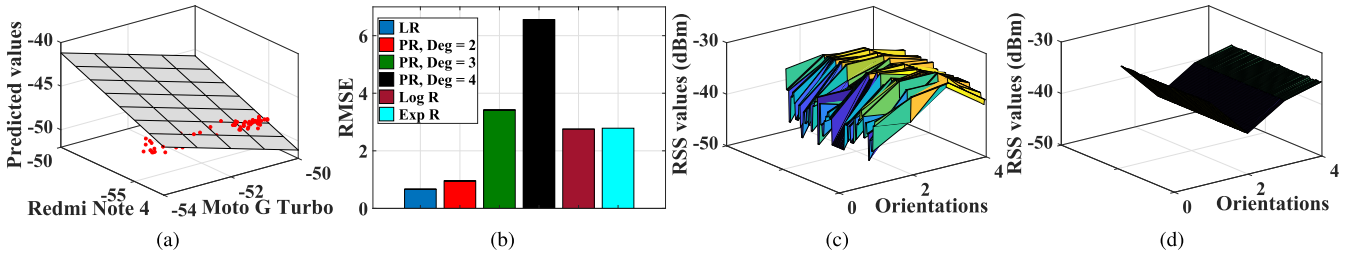


**FIGURE 8.** (a) The plane formed after the linear regression for three smart devices. (b) Comparison of RMSE for different types of regression models. LR = Linear regression, PR = Polynomial regression, Log R = Logarithmic regression, Exp R = Exponential regression and Deg = Degree. (c) The plot of RSS values for multiple orientations. (d) The plot of transformed orientation after regression.

and device diversity, and the performance of multiple Gaussian method over single Gaussian.

### 1) MULTISTAGE REGRESSION TO MITIGATE THE EFFECT OF DEVICE HETEROGENEITY AND ORIENTATION

As discussed in the methodology section, multistage regression compensates the device and orientation diversity. Fig. 8(a) shows the regression output obtained between the three smart devices considering Redmi Note 5 as the dependent variable. The transformed RSS data is used as the input to the GMM localization method. Fig. 8(b) shows the Root Mean Square Error (RMSE) values for goodness fit of the RSS data values. It is observed that the linear model produces the least RMSE, and hence, there exist a linear relationship between the orientations as well as devices' RSS values. Fig. 8(c) shows the original data with orientation diversity and Fig. 8(d) shows the transformed orientation using the linear regression method. Fig. 9(a) shows the AIC and BIC scores computed for a particular location. It is observed from the elbow diagram that the number of Gaussian in the mixture is two as the difference between two consecutive AIC/BIC remains almost constant after two Gaussian. The Gaussian parameters estimated at a particular location are given as $\mu_1 = -51.62$, $\mu_2 = -52.72$, $\Sigma_1 = 0.057$ and $\Sigma_2 = 0.46$ with mixing coefficient $c_1 = 67$ and $c_2 = 33$. These results are shown for illustration purpose, and similarly, the GMM parameters are obtained for other locations. This is in accordance with the number of Gaussian components obtained with the heuristic AIC/BIC criteria. The regression coefficients obtained at a particular location are $\beta_0 = -40.54$, $\beta_1 = 0.012$ and $\beta_2 = -0.15$ respectively with a RMSE of 0.19. In a similar way, the regression coefficients at other locations are estimated. It is observed that the RMSE obtained is low and desirable for establishing a model; hence, the linear regression model is most suitable.

### 2) CHOICE OF DISTANCE METRIC, CROSS-VALIDATION FOLDS AND DEPENDENT VARIABLE

It is important to use the correct distance metric in fingerprinting based localization. Fig. 9(b) shows the error for the different type of distance metric used for the fingerprinting method. We observe that the minimum localization error of 0.57 m is obtained for the Euclidean distance metric

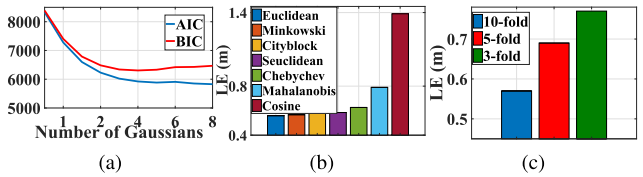


**FIGURE 9.** (a) Variation of AIC and BIC scores for estimating the number of Gaussian components. (b) Localization error for different distance metrics. (c) Localization error for different folds of cross validation; LE = Localization error.

and a maximum of 1.35 m for the Cosine distance metric. Hence, the Euclidean distance metric is chosen for the localization. This is also suggested in [60] that for higher dimensional dataset Mahalanobis and Euclidean distance are preferred. Further, the cross-validation method with least localization error is also analyzed. It is observed that the 10-fold cross-validation method gives the least localization error of 0.57 m, as shown in Fig. 9(c). It is shown in Fig. 10 that a random selection of the dependent variable leads to a higher RMSE, which further leads to a higher localization error. It is observed that for our experimental RSS data, orientation 3 as a dependent variable provides least average RMSE. Hence, in the proposed method, we choose $O_3$ as the dependent variable for multiple regression, as shown in Fig. 10(a). Similarly, Fig. 10(b) shows that $SD_1$ should be chosen as the dependent variable as the RMSE is least for $SD_1$. The effect of the choice of the dependent variable on localization error is shown in Fig. 10(c). It is observed that orientation 3 provides the best localization error of 0.27 m. Further, for the user case, when the device type is not known, smart device detection is performed as it helps in better location estimation. Fig. 10(d) shows the accuracy of the smart device detection. We find that the multiple Gaussian method detects the type of smart device with a high accuracy of 94% as the RSS data from multiple smart device follows a mixture of GMM.

### 3) LOCALIZATION ERROR PERFORMANCE

Fig. 11(a) shows that error in distance estimation of the smart device from the access point using multiple Gaussian method is 0.25 m as compared to 0.49 m of single Gaussian considering orientation 3. Further, it can be observed from Fig. 11(b) that multiple Gaussian method outperforms the single Gaussian for all the four orientations. Also, the localization error

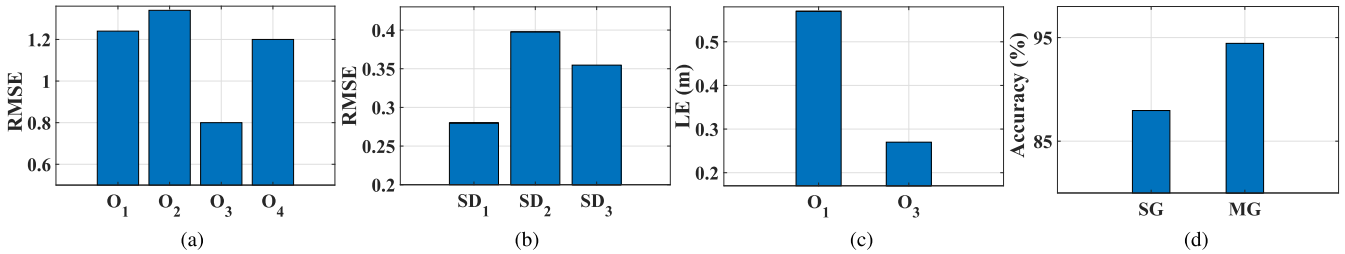


**FIGURE 10.** (a) RMSE considering all the four orientations as the dependent variable.(b) RMSE considering three smart devices as the dependent variable. (c) Comparison of localization error betwen two orientations. (d) Detection accuracy for the type of smart device using the proposed method.

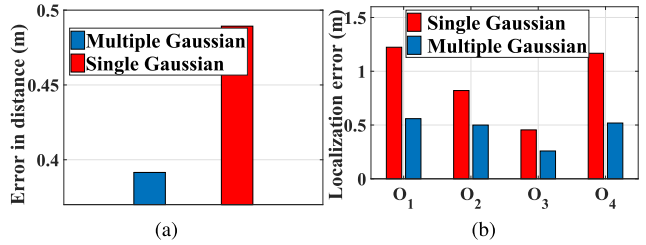


**FIGURE 11.** (a) Error in distance estimation using single and multiple Gaussian method. (b) Comparison of localization error for single and multiple Gaussian method for different orientations as dependent variable.

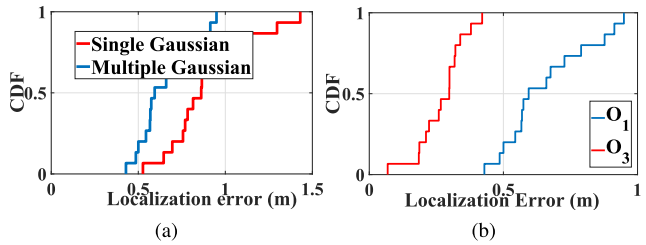


**FIGURE 12.** (a) Comparison of CDF of localization error for single and multiple Gaussian method. (b) Comparison of CDF of localization error for two different orientations as the dependent variable.

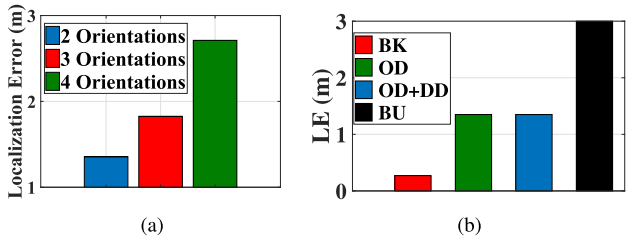


**FIGURE 13.** (a) Results of localization error with different number of orientations detected using the threshold method. (b) Performance of the proposed method for different user cases. BK = Both orientation and device type known, OD = Orientation detection through threshold method when the device type is known, OD+DD = Both orientation detection and device type detection using the proposed method and BU = Both orientation and device type unknown.

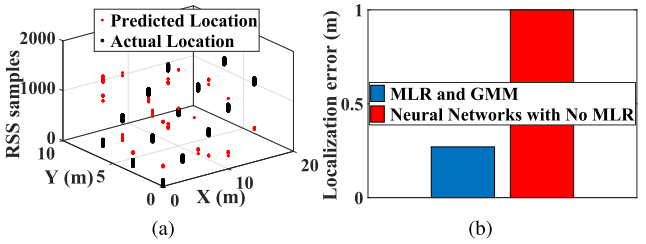


**FIGURE 14.** (a) The actual and estimated locations for test RSS values using GMLoc. (b) Comparison of the proposed GMLoc with neural networks. MLR = Multi level regression.

is least for orientation 3 as the RMSE is least for the same. Cumulative distribution function (CDF) plot of localization error is shown in Fig. 12(a). It is observed that in terms of localization error, for all the locations, multiple Gaussian method outperforms the single Gaussian method. This is because the RSS data from smart devices with multiple orientations in an IoT network follows a multi Gaussian model instead of a single Gaussian. Further, the localization error is computed for the three user cases. It is observed in Fig. 12(b) that when the system has the prior knowledge of the type of orientation and smart device, it provides the least localization error of 0.27 m with the dependent variable as orientation $O_3$.

The CDF plot shows that for more than 90% of the test RSS samples, the localization error is 0.3 m. Also, when the base orientation is not chosen with least RMSE, for example with the dependent variable as $O_1$ the localization error increases to 0.7 m. This result clearly shows the importance of the choice of the dependent variable. We also evaluate the performance of GMLoc when a different number of orientations are detected using the proposed threshold method.

Fig. 13(a) shows that the proposed method estimates the location with an error of 1.35 m with the knowledge of a minimum of only two orientations. Further, Fig. 13(b) shows the comparison of localization error for three different user cases. It is observed that when no prior information of device type and orientation is available (BU case), the method yields a comparatively high localization error of 3 m. For such cases, the proposed method provides an improved localization error of 1.35 m by identifying the device type and orientation. Also, when we have the prior knowledge of the device type and orientation with the appropriate dependent variable, the proposed method achieves the least localization error of 0.27 m. Hence, the proposed method is suitable for all type of user cases in IoT network.

Finally, Fig. 14(a) shows the performance of the proposed algorithm for the test RSS vectors. The black dots show the actual location, and the red dots show the estimated location with the average localization error of 0.39 m.

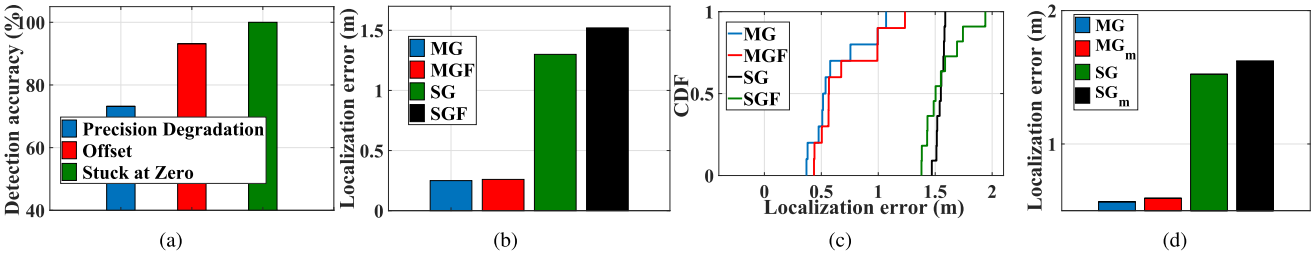


**FIGURE 15.** (a) Performance of GMLoc for anomaly detection. (b) Localization error comparison of single and multiple Gausssian method for normal and erroneous RSS values. (c) CDF comparison of localization error on normal and erroneous RSS values. SGF = Single Gaussian on erroneous data and MGF = Multiple Gaussian on erroneous data. (d) Localization error with missing data for single and multiple Gaussian where $MG_m$ denotes the multiple Gaussian on missing data and $SG_m$ for single Gaussian.

The performance of GMLoc is compared with machine learning methods such as neural networks where the data is fed without any regression, knowing the device type and the orientation. Fig. 14(b) shows that GMLoc outperforms the neural network classification method. Therefore, it is concluded that multistage regression with GMM addresses the device and orientation diversity issues better than the state-of-the-art single Gaussian and neural network method.

#### 4) PERFORMANCE OF GMLOC ON ERRONEOUS RSS VALUES

The performance of GMLoc is also evaluated when the system encounters the attack, or the smart devices' sensors generate erroneous RSS readings. First, anomaly detection is performed with the proposed method. Fig. 15(a) shows the performance of the proposed method on the faults. The detection accuracy is high for stuck-at fault and the offset fault and marginal for the precision degradation fault. Further, the performance of the single Gaussian method over multiple Gaussian is compared for the erroneous data. In order to test the robustness of the proposed method, a maximum of 20% of the RSS values are made erroneous with $\eta = 5$ and $\beta = 10$. It is observed that the proposed method is still robust for erroneous RSS data and multiple Gaussian method outperforms single Gaussian. Fig. 15(b) shows the comparison of the GMLoc on normal and erroneous data for single and multiple Gaussian methods. It is concluded that the multiple Gaussian method outperforms the single Gaussian. This is because the erroneous data also follows a mixture of Gaussian instead of the single Gaussian; therefore, the location estimation is better with the multiple Gaussian method even in the case of erroneous RSS values.

Fig. 15(c) shows the variation in the Cumulative Distribution Function (CDF) of the localization error. We find that 90% of the test RSS samples are located with an error of 1.2 m and more than 60% of the test RSS vectors are located within 1 m range. On the contrary, the single Gaussian method can locate 90 % of the test RSS vectors with approximately 2 m error. Further, Fig. 15(d) shows the performance of the proposed method on the RSS test data that contain missing data. We evaluated the performance of the proposed method with 15% missing data and imputed the RSS values using a linear regression method. It is observed that the multiple Gaussian method provides approximately the same results as on the normal RSS samples. Therefore, we conclude that the localization with multiple Gaussian method addresses the faulty data with higher accuracy than the single Gaussian method. The experiments are performed using Intel Core i5 processor having 8 GB RAM. It is observed that the multiple Gaussian method takes 0.35 seconds more than the single Gaussian method.

## IV. CONCLUSION AND FUTURE RESEARCH CHALLENGES

The paper proposed a novel method to address both device heterogeneity and orientation issues in the IoT network using a multistage regression and GMM method. The effect of the choice of the dependent variable for orientation and device type on localization is also investigated. The results obtained using the GMM method achieve better localization as compared to single Gaussian methods. Also, the proposed method is robust to erroneous RSS data in the IoT network. Additionally, the proposed method detects the device type with an accuracy of 95%, which improves the localization error to 1.35 m even when both the device type and orientation are unknown. The future work includes the development of a mobile application for the proposed method.

A low complexity localization algorithm which is accurate, handles device diversity and orientation, and uses opportunistic signals is desirable for time-vayring IoT networks. The localization method needs to be robust to environment changes and the type of signals. Further, new methods using deep neural networks can be used to increase the localization accuracy for time-varying IoT networks. Channel State Information (CSI) based device-free localization method in the presence of fading effect is a potential future work for a smart environment.

## REFERENCES

[1] F. Khelifi, A. Bradai, A. Benslimane, P. Rawat, and M. Atri, "A survey of localization systems in Internet of Things," *Mobile Netw. Appl.*, vol. 24, no. 3, pp. 761–785, 2019.

[2] S. Kumar and R. M. Hegde, "A review of localization and tracking algorithms in wireless sensor networks," 2017, *arXiv:1701.02080*. [Online]. Available: https://arxiv.org/abs/1701.02080

[3] I. Al Ridhawi, M. Aloqaily, A. Karmouch, and N. Agoulmine, "A location-aware user tracking and prediction system," in *Proc. Global Inf. Infrastruct. Symp.*, Jun. 2009, pp. 1–8.

[4] I. Al Ridhawi, M. Aloqaily, B. Kantarci, Y. Jararweh, and H. T. Mouftah, "A continuous diversified vehicular cloud service availability framework for smart cities," *Comput. Netw.*, vol. 145, pp. 207–218, Nov. 2018.
[5] Y. Al Ridhawi, I. Al Ridhawi, A. Karmouch, and A. Nayak, "A context-aware and location prediction framework for dynamic environments," in *Proc. IEEE 7th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2011, pp. 172–179.
[6] Y. Al Ridhawi, I. Al Ridhawi, L. Bruno, and A. Karmouch, "Policy-based personalized context dissemination for location-aware services," in *Proc. Int. Conf. Mobile Ubiquitous Syst., Comput., Netw., Services*. Berlin, Germany: Springer, 2010, pp. 366–371.
[7] V. Balasubramanian, M. Aloqaily, F. Zaman, and Y. Jararweh, "Exploring computing at the edge: A multi-interface system architecture enabled mobile device cloud," in *Proc. IEEE 7th Int. Conf. Cloud Netw. (CloudNet)*, Oct. 2018, pp. 1–4.
[8] R. C. Shit, S. Sharma, D. Puthal, and A. Y. Zomaya, "Location of Things (LoT): A review and taxonomy of sensors localization in IoT infrastructure," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2028–2061, 3rd Quart., 2018.
[9] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2568–2599, 3rd Quart., 2019.
[10] N. Patwari, A. O. Hero, M. Perkins, N. S. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2137–2148, Aug. 2003.
[11] C. Yang and H.-R. Shao, "Wi-Fi-based indoor positioning," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 150–157, Mar. 2015.
[12] S. He and S.-H. G. Chan, "Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 466–490, Jan. 2015.
[13] P. Davidson and R. Piché, "A survey of selected indoor positioning methods for smartphones," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1347–1370, 2nd Quart., 2017.
[14] J.-G. Park, D. Curtis, S. Teller, and J. Ledlie, "Implications of device diversity for organic localization," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 3182–3190.
[15] L. Zhang, L. Ma, Y. Xu, and C. Li, "Linear regression algorithm against device diversity for indoor WLAN localization system," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.
[16] F. Dong, Y. Chen, J. Liu, Q. Ning, and S. Piao, "A calibration-free localization solution for handling signal strength variance," in *Mobile Entity Localization and Tracking in GPS-Less Environnments*. Berlin, Germany: Springer, 2009, pp. 79–90.
[17] G. Lui, T. Gallagher, B. Li, A. G. Dempster, and C. Rizos, "Differences in RSSI readings made by different Wi-Fi chipsets: A limitation of WLAN localization," in *Proc. Int. Conf. Localization GNSS (ICL-GNSS)*, Jun. 2011, pp. 53–57.
[18] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D tracking via body radio reflections," in *Proc. 11th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, 2014, pp. 317–329.
[19] F. D. Rosa, M. Pelosi, and J. Nurmi, "Human-induced effects on RSS ranging measurements for cooperative positioning," *Int. J. Navigat. Observ.*, vol. 2012, Sep. 2012, Art. no. 959140.
[20] S.-H. Fang, C.-H. Wang, and Y. Tsao, "Compensating for orientation mismatch in robust Wi-Fi localization using histogram equalization," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5210–5220, Nov. 2015.
[21] Z.-A. Deng, Z. Qu, C. Hou, W. Si, and C. Zhang, "Wi-Fi positioning based on user orientation estimation and smartphone carrying position recognition," *Wireless Commun. Mobile Comput.*, vol. 2018, Apr. 2018, Art. no. 5243893.
[22] A. Papapostolou and H. Chaouchi, "Orientation-based radio map extensions for improving positioning system accuracy," in *Proc. Int. Conf. Wireless Commun. Mobile Comput., Connecting World Wirelessly*, 2009, pp. 947–951.
[23] Y. Li, Z. He, Y. Li, Z. Gao, R. Chen, and N. El-Sheimy, "Enhanced wireless localization based on orientation-compensation model and differential received signal strength," *IEEE Sensors J.*, vol. 19, no. 11, pp. 4201–4210, Jun. 2019.
[24] S. Lee, B. Cho, B. Koo, S. Ryu, J. Choi, and S. Kim, "Kalman filter-based indoor position tracking with self-calibration for RSS variation mitigation," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 8, 2015, Art. no. 674635.
[25] A. W. Tsui, Y.-H. Chuang, and H.-H. Chu, "Unsupervised learning for solving RSS hardware variance problem in Wi-Fi localization," *Mobile Netw. Appl.*, vol. 14, no. 5, pp. 677–691, 2009.
[26] S. He, S.-H. G. Chan, L. Yu, and N. Liu, "SLAC: Calibration-free pedometer-fingerprint fusion for indoor localization," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1176–1189, May 2018.
[27] X. Zhang, A. K.-S. Wong, C.-T. Lea, and R. S.-K. Cheng, "Unambiguous association of crowd-sourced radio maps to floor plans for indoor localization," *IEEE Trans. Mobile Comput.*, vol. 17, no. 2, pp. 488–502, Feb. 2018.
[28] S. Kumar and S. K. Das, "ZU-mean: Fingerprinting based device localization methods for IoT in the presence of additive and multiplicative noise," in *Proc. Workshop Program 19th Int. Conf. Distrib. Comput. Netw.*, 2018, Art. no. 15.
[29] C. Laoudias, P. Kolios, and C. Panayiotou, "Differential signal strength fingerprinting revisited," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2014, pp. 30–37.
[30] Y. Shu, Y. Huang, J. Zhang, P. Coué, P. Cheng, J. Chen, and K. G. Shin, "Gradient-based fingerprinting for indoor localization and tracking," *IEEE Trans. Ind. Electron.*, vol. 63, no. 4, pp. 2424–2433, Apr. 2015.
[31] C. Figuera, J. L. Rojo-Álvarez, I. Mora-Jiménez, A. Guerrero-Curieses, M. Wilby, and J. Ramos-López, "Time-space sampling and mobile device calibration for Wi-Fi indoor location systems," *IEEE Trans. Mobile Comput.*, vol. 10, no. 7, pp. 913–926, Jul. 2011.
[32] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proc. 10th Annu. Int. Conf. Mobile Comput. Netw.*, 2004, pp. 70–84.
[33] J. Luo and X. Zhan, "Characterization of smart phone received signal strength indication for WLAN indoor positioning accuracy improvement," *J. Netw.*, vol. 9, no. 3, pp. 739–747, 2014.
[34] M. Alfakih, M. Keche, and H. Benoudnine, "Gaussian mixture modeling for indoor positioning Wi-Fi systems," in *Proc. 3rd Int. Conf. Control, Eng. Inf. Technol. (CEIT)*, May 2015, pp. 1–5.
[35] A. Goswami, L. E. Ortiz, and S. R. Das, "WiGEM: A learning-based approach for indoor localization," in *Proc. 7th Conf. Emerg. Netw. Exp. Technol.*, 2011, Art. no. 3.
[36] C. H. Tseng and J.-S. Yen, "Enhanced Gaussian mixture model of RSSI purification for indoor positioning," *J. Syst. Archit.*, vol. 81, pp. 1–6, Nov. 2017.
[37] N. A. Dieng, M. Charbit, C. Chaudet, L. Toutain, and T. B. Meriem, "Indoor localization in wireless networks based on a two-modes Gaussian mixture model," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*., Sep. 2013, pp. 1–5.
[38] D. Wu, Y. Zhang, L. Bao, and A. C. Regan, "Location-based crowdsourcing for vehicular communication in hybrid networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 837–846, Jun. 2013.
[39] Y. Zhang, S. Xing, Y. Zhu, F. Yan, and L. Shen, "RSS-based localization in WSNs using Gaussian mixture model via semidefinite relaxation," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1329–1332, Jun. 2017.
[40] T.-N. Lin and P.-C. Lin, "Performance comparison of indoor positioning techniques based on location fingerprinting in wireless networks," in *Proc. Int. Conf. Wireless Netw., Commun. Mobile Comput.*, vol. 2, Jun. 2005, pp. 1569–1574.
[41] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[42] P. Richter, M. Valkama, and E. S. Lohan, "Attack tolerance of RSS-based fingerprinting," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.
[43] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor network data fault types," *ACM Trans. Sensor Netw.*, vol. 5, no. 3, p. 25, 2009.
[44] T. Muhammed and R. A. Shaikh, "An analysis of fault detection strategies in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 78, pp. 267–287, Jan. 2017.

[45] J. Ludeña-Choez, J. J. Choquehuanca-Zevallos, and E. Mayhua-López, "Sensor nodes fault detection for agricultural wireless sensor networks based on NMF," *Comput. Electron. Agricult.*, vol. 161, pp. 214–224, Jun. 2018.
[46] J. Jun, L. He, Y. Gu, W. Jiang, G. Kushwaha, A. Vipin, L. Cheng, C. Liu, and T. Zhu, "Low-overhead Wi-Fi fingerprinting," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 590–603, Mar. 2018.
[47] Y. Ye, B. Wang, X. Deng, and L. T. Yang, "On solving device diversity problem via fingerprint calibration and transformation for RSS-based indoor localization system," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–8.
[48] P. J. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Mining Knowl. Discovery*, vol. 12, no. 1, pp. 29–45, 2006.
[49] Y. Chapre, P. Mohapatra, S. Jha, and A. Seneviratne, "Received signal strength indicator and its analysis in a typical WLAN system (short paper)," in *Proc. 38th Annu. IEEE Conf. Local Comput. Netw.*, Oct. 2013, pp. 304–307.
[50] G. Strang, *Introduction to Linear Algebra*, vol. 3. Wellesley, MA, USA: Wellesley Cambridge Press, 1993.
[51] N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imag.*, vol. 16, no. 4, 2007, Art. no. 049901.
[52] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.
[53] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annu. Rev. Statist. Appl.*, vol. 6, pp. 355–378, Jan. 2019.
[54] H. Bozdogan and S. L. Sclove, "Multi-sample cluster analysis using Akaike's information criterion," *Ann. Inst. Stat. Math.*, vol. 36, no. 1, pp. 163–180, 1984.
[55] G. J. McLachlan and S. Rathnayake, "On the number of components in a Gaussian mixture model," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 4, no. 5, pp. 341–355, 2014.
[56] V. Renaudin and C. Combettes, "Magnetic, acceleration fields and gyroscope quaternion (MAGYQ)-based attitude estimation with smartphone sensors for indoor pedestrian navigation," *Sensors*, vol. 14, no. 12, pp. 22864–22890, 2014.
[57] H. Verma and S. Kumar, "An accurate missing data prediction method using LSTM based deep learning for health care," in *Proc. 20th Int. Conf. Distrib. Comput. Netw.*, 2019, pp. 371–376.
[58] C. Feng, W. S. A. Au, S. Valaee, and Z. Tan, "Compressive sensing based positioning using RSS of WLAN access points," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.
[59] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.
[60] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. Int. Conf. Database Theory*. Berlin, Germany: Springer, 2001, pp. 420–434.

**ANKUR PANDEY** (S'18) received the M.Tech. degree in communication systems engineering from the Indian Institute of Technology Patna, India, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. He was an IT Analyst with TCS, India. His broad research interests include signal processing for the IoT networks with a particular focus on the node localization issues in the IoT networks. He was a recipient of National Award from L&T - ISTE for the Best M.Tech. Thesis (2nd Prize), Best M.Tech. Thesis award from the institute, Institute Silver Medal for scoring highest CPI in the branch, and IEEE ANTS student travel grant.

**RAGHU VAMSI** is currently pursuing the B.Tech. degree from the Indian Institute of Technology Patna. He is in his junior year and has interest in the broad area of wireless sensor networks and signal processing.

**SUDHIR KUMAR** (S'13–M'16) received the B.Tech. degree in ECE from the West Bengal University of Technology, Kolkata, in 2010, and the Ph.D. degree from the EE Department, Indian Institute of Technology Kanpur, in 2015. He was an Assistant Professor with the Department of ECE, Visvesvaraya National Institute of Technology Nagpur, India. He was a Scientist with TCS Research Kolkata, India, and an Erasmus Mundus Fellow with the Department of Computer Science, University of Oxford, U.K. He is currently an Assistant Professor with the Department of Electrical Engineering, Indian Institute of Technology Patna, India. He has published more than 35 research articles in prestigious journals and conference proceedings. His broad research interests include wireless sensor networks and the internet-of-things (IoT). He was a recipient of several awards and fellowships, such as National Award from L&T - ISTE for having guided the Best M.Tech. Thesis (2nd Prize), SERB Indo-US postdoctoral fellowship, India-EU Namaste fellowship, TCS research scholarship, MHRD scholarship, IEEE ICC student travel grant award, and COMSNETS travel grant.