Feature-Based Generalized Gaussian Distribution Method for NLoS Detection in Ultra-Wideband (UWB) Indoor Positioning System

Fuhu Che*, Qasim Zeeshan Ahmed*, Jaron Fontaine†, Ben Van Herbruggen†,
Adnan Shahid†, Eli De Poorter†, Pavlos I. Lazaridis *.
* School of Computing and Engineering, University of Huddersfield, Huddersfield, UK
† IDLab, Department of Information Technology at Ghent University - imec, Belgium
Email: fuhu.che@hud.ac.uk, q.ahmed@hud.ac.uk.

Abstract

Non-Line-of-Sight (NLoS) propagation condition is a crucial factor affecting the precision of the localization in Ultra-Wideband (UWB) Indoor Positioning System (IPS). Numerous supervised Machine Learning (ML) approaches have been applied for NLoS identification to improve the accuracy of the IPS. However, it is difficult for existing ML approaches to maintain a high classification accuracy when the database contains a small number of NLoS signals and a large number of line of sight (LoS) signals. The inaccurate localization of the target node caused by these small number of NLoS signals can still be problematic. To solve this issue, we propose features-based Gaussian Distribution (GD) and Generalized Gaussian Distribution (GGD) NLoS detection algorithm for imbalanced LoS and NLoS signals. By employing our detection algorithm for the imbalanced dataset, the NLoS classification accuracy can achieve 96.7% for GD and 98.0% for GGD. We also compared the proposed algorithm with the existing cutting-edge such as Support-Vector-Machine (SVM), Decision Tree (DT), Naive Bayes (NB) and Neural Network (NN), which can achieve an accuracy of 92.6%,92.8%,93.2% and 95.5%, respectively. The results demonstrate that the GGD algorithm can achieve high classification accuracy with the imbalanced dataset and also achieve a high classification accuracy for different ratios of LoS and NLoS signals which proves the robustness and effectiveness of the algorithm.

Index Terms

Ultra-wideband, Indoor Positioning System, Machine Learning , Non-Line-of-Sight Identification, Naïve Bayes,Gaussian mixed model,Generalized Gaussian Distribution.

I. Introduction

With the rapid development of the Internet of Things (IoT)s, the requirement of a precise indoor positioning system (IPS) has attracted considerable attention in the research community and industry [1]–[6]. Several examples aforesaid, pedestrian tracking systems [4], [7], [8], autonomous flying drones in warehouses [8]–[10], and social distancing requirements caused by pandemic such as COVID-19 [12], [13], etc., require accurate IPS. The Global Navigation Satellite System (GNSS) provides tremendous convenience to human life as they provide real-time localization in open space. Unfortunately, the GNSS signals are attenuated severely by the wall and fail to achieve the accurate positioning in indoor environments [2], [14]. Among various indoor positioning technologies, Ultra-wideband (UWB) can achieve high

Part of the paper is also presented in 26th IEEE International Conference on Automation and Computing (ICAC), 2021. This work is supported in part by the MOTOR5G Marie Sklodowska-Curie (MSCA) Innovative Training Networks (ITN)s-2019- and in part by RECOMBINE-MSCA-Research and Innovation Staff Exchange (RISE)-2019 under Grant ID: 861219 and ID: 872857. This work is also supported by the Capacity Building for Digital Health Monitoring and Care Systems in Asia EAC-A02-2019-CBHE under Grant ID: 619193.

accuracy due to its characteristics of extremely short pulse that provides good time resolution [8], [15]. However, the accuracy of UWB IPS could be significantly affected when the NLoS signal occurs [11], [16], [17]. The NLoS condition exists when the signals between the transceivers are reflected or blocked by the obstacles. In this case, signal propagation delay occurs, resulting in longer Time-of-Flight (ToF) and an positive bias estimated distance error between the transmitter and receiver [16], [17]. Thus, significantly reducing the accuracy of IPS.

The current literature includes several research works that enhance the accuracy of the UWB IPS by identifying whether the signal is LoS or NLoS component [8], [15], [16], [18]. The methods of NLoS identification can be coarsely summarised into two types (i) Non-features based NLoS signal classification that uses context information and (ii) feature-based NLoS identification relying on the UWB waveforms. The *non-feature based approach* from [18] uses a modified Kalman filter to classify LoS/NLoS conditions based on the Bayesian sequential of range measurements. In [19], the authors present a real-time NLoS identification approach based on the received signal strength without the training phase and prior knowledge of the environment. In [20] models the NLoS as a deterministic additive term and identifies NLoS based on the statistical features of range measurements. In [21], a fusion technique such as an Inertial Navigation System (INS) is combined with UWB for pedestrian tracking. In [8], the authors apply the floor map injunction with INS and UWB to predict the state and then determine and recognise the NLoS signals.

In contrast, in the *feature-based NLoS identification*, the UWB waveform signals under the LoS are different from those under NLoS conditions. These features can be extracted from the UWB signal to identify NLoS conditions by employing Machine Learning (ML) algorithms. One of the early ML-based NLoS identification approaches in UWB was proposing the Support Vector Machine (SVM) algorithm as a classifier in [22], [23]. In these papers [22]–[24], the identification of LoS and NLoS signals was considered as a binary classification problem. The results proved that the ML approaches could improve the accuracy of UWB IPS by identifying the NLoS signals. Different ML techniques like Naive Bayes (NB) [25], Boosted Decision Tree (BDT) [26], etc., were also investigated. Deep-learning based approach such as Convolutional Neural Network (CNN) was developed in [27]. Furthermore in [28], the authors propose a semi-supervised based ML approach using autoencoders which achieves 29% higher accuracy than state-of-the-art deep neural network algorithm. However, the above-mentioned feature-based methods have drawbacks, especially when the data is imbalanced and a small number of NLoS data samples are present. In such cases, it is hard for such algorithms to train a robust classifier for NLoS identification. To address this shortcoming, we propose a Gaussian Distribution (GD) and Generalized Gaussian Distribution (GGD) algorithm for NLoS signal detection in the presence of imbalance datasets. Our proposed algorithm is an unsupervised learning algorithm which start with training the LoS data to build a model, then determine a threshold through density estimation according to the GD and GGD of each feature, and test the classification of the new data according to this calculated threshold.

Therefore, the main contributions of this paper are as follows.

- We proposed an unsupervised learning algorithm based on Gaussian Distribution (GD) and Generalized Gaussian Distribution (GGD) algorithms to discriminate between LoS and NLoS conditions in presence of imbalance datasets with limited NLoS training data for IPS.

- We Compare the proposed algorithm with the existing supervised ML algorithm (SVM, DT,NB and NN) in terms of the confusion matrix, receiver operating characteristics (ROC) curve and the area under the curve (AUC) to show the superior performance of the proposed algorithms.

The remainder of this paper is organized as follows. Section II describes the overall UWB system model and the principle of UWB localization. In Section III, our proposed classification algorithms are discussed. In Section IV, the principle of our proposed algorithms is presented followed by the features used for NLoS signal classification. In addition, this section also discusses the environment in which the data was collected, and the hardware used for this data collection. Section V presents the performance evaluation of the proposed algorithms and compare the results with the state-of-art Ml algorithms in detail. The summary of the accomplishment is given in Section VI.

II. UWB Positioning System Model

A. Transmitted UWB Signal by the Anchor

We consider an UWB signal waveform $s(t)$ transmitted by the help of K pulses p with a period of T_p that consists of transmitted frames [29]. As the transmitted UWB signal location is known the transmitted signal is modelled as

$$s(t) = \sqrt{E_s} \sum_{k=1}^{K-1} p(t - kT_p), \tag{1}$$

where E_s is the energy of the UWB signal $s(t)$.

B. Received UWB Signal by the Mobile Node

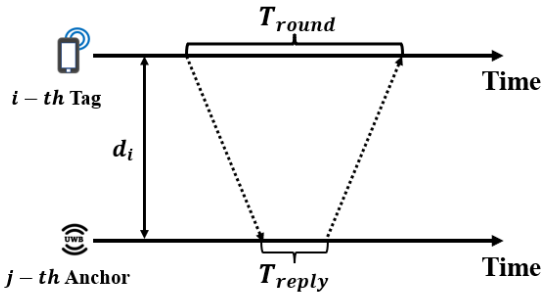
The transmitted signal $s(t)$ experiences multipath channel effects and the received signal at the i-th mobile node can be expressed as (2) [30], [31].

$$r_i(t) = \sum_{v_i=1}^{V_i} h_{v_i} s(t - \tau_{v_i}) + n(t), \quad i = 1, 2, \cdots, N. \tag{2}$$

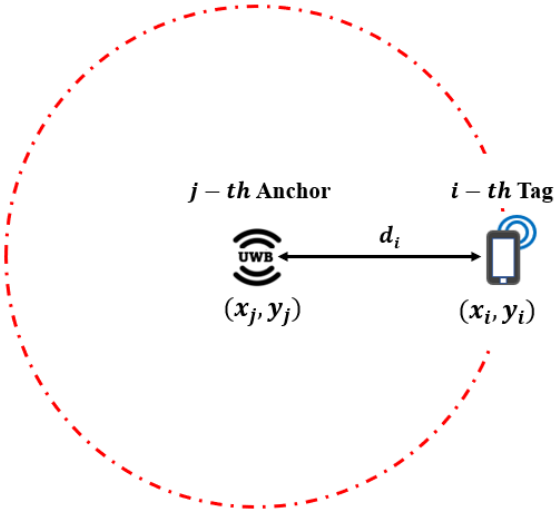
where V_i is the maximum number of multipath experienced by the i-th mobile node. h_{v_i} and τ_{v_i} represent the amplitude and delay of the v-th path respectively at the i-th mobile node, and $n(t)$ is the Additive White Gaussian Noise (AWGN) with zero mean and two-sided power spectral density $N_0/2$. However, as this i-th mobile node or referred to a tag in literature will be moving, therefore, our main interest will be to calculate the distance between anchor node and the i-th mobile node. The distance calculation is discussed in the upcoming subsection.

C. Localization Theory

UWB-based IPS consists of two different kinds of nodes. Nodes with a known position are called anchors whereas the nodes with unknown position are tags and the position is to be determined. Firstly, time of arrival (TOA) technique can be used to measure the distance between the anchor and tag. Secondly,



(a) Propagation time calculations



(b) Range scheme

Fig. 1: UWB localization theory in a 2-dimensional ($2D$) environment

the triangulation technique helps to determine the position of the tag in a 2-dimensional ($2D$) environment with three or more than three anchors.

1) Time of Arrival (TOA) Measurements

The first process of TOA will require both the anchor and the tag to have synchronized clocks. As shown in Figure 1, a timestamp will be sent by the i-th tag to the j-th anchor. The timestamp will be processed by the j-th anchor in T_{reply} seconds and send back to the i-th tag. The time taken for the i-th tag is T_{round} and the propagation time $\tau_{\mathrm{i,j}}$ can be expressed as

$$\tau_{i,j} = \frac{T_{round} - T_{reply}}{2}, \quad j = 1, 2, 3. \tag{3}$$

The estimate distance $d_{i,j}$ between the i-th anchor and j-th tag can be calculated as

$$d_{i,j} = c \times \tau_{\mathrm{i,j}}, \tag{4}$$

where c is the speed of light in meter per seconds (m/s).

Trilateration Approach for UWB Localization

Figure 1b shows the position of the $i-th$ tag with respect to the j-th anchor. The coordinates of the $j-th$ anchor is (x_j, y_j) which are already known. The coordinates of the i-th tag are represented as $(\hat{x}_i, \hat{y}_i)$, where $(\hat{\cdot})$ indicates the estimate of the position. The distance between each anchor and the tag is calculated as

$$\begin{aligned} d_{i,j} &= \sqrt{(\hat{x}_i - x_j)^2 + (\hat{y}_i - y_j)^2}, \\ & i = 1, 2, \cdots, N, \quad j = 1, 2, 3. \end{aligned} \tag{5}$$

The position of tag $(\hat{x}_i, \hat{y}_i)$ can be determined by employing least-squares solution.

III. Proposed Algorithms

Given a UWB dataset of size N such that $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\}^T$. If n represents the index of the dataset $\mathcal{S}$, then n will consist of M features and will be represented as $\boldsymbol{s}_n = \{s_{n,1}, s_{n,2}, \cdots, s_{n,M}\}$. The collection of these features of datasets are discussed in detail in the Section IV. Upon collection of these features, we need to design ML algorithms such that we can classify the test data $\boldsymbol{u}$ as LoS or NLoS signals.

A. Gaussian Distribution (GD)

Assuming the feature s_i to be gaussian distribution with mean μ_i, and variance σ_i^2 can be written as

$$\begin{aligned} P(s_i, \mu_i, \sigma_i^2) &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(s_i - \mu_i)^2}{2\sigma_i^2}\right), \\ & \quad i = 1, 2, \cdots, M. \end{aligned} \tag{6}$$

However, we will still require to calculate the mean μ_i and the variance σ_i^2 of the i-th feature. As the exact mean and variance of the features of the dataset is unknown, we can incorporate the provided training data N to calculate their estimates. The estimate of the mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ for the i-th feature can be calculated as

$$\hat{\mu}_i = \frac{1}{N}\sum_{n=1}^{N} s_{n,i}, \quad i = 1, 2, \cdots, M, \tag{7}$$

$$\hat{\sigma}_i^2 = \frac{1}{N}\sum_{n=1}^{N} (s_{n,i} - \hat{\mu}_i)^2, \quad i = 1, 2, \cdots, M. \tag{8}$$

Once we have the estimates of the mean and variance of each feature on training data, given a test data $\boldsymbol{u}$ we can calculate the probability as

$$P(\boldsymbol{u}) = \prod_{i=1}^{M} P(u_i, \hat{\mu}_i, \hat{\sigma}_i^2), \tag{9}$$

and classify the testing data as:

$$\begin{cases} P(\boldsymbol{u}) > \epsilon, & \text{LoS} \\ P(\boldsymbol{u}) \leq \epsilon, & \text{NLoS} \end{cases} \tag{10}$$

where ϵ is judgment boundary and will be discussed in detail in section IV.

B. Generalized Gaussian Distribution, (GGD)

By using GD algorithm, some abnormal features of Los data may have a difficulty to classify, as a result the model could wrongly classify it as a NLoS component. Furthermore, the GD algorithm require two key parameters to be modelled: a) mean and b) variance of the data as mentioned III-A. In such cases, the GD model may not be able to high accuracy identify the NLoS dataset. Therefore, instead of

GD, Generalized Gaussian Distribution (GGD) can be adopted [29]. The GGD of the i-th feature can be written as

$$P(s_i, \mu_i, \alpha_i, \beta_i) = \frac{\beta_i}{2\alpha_i \Gamma(1/\beta_i)} \exp\left(-\frac{|s_i - \mu_i|}{\alpha_i}\right)^{\beta_i} \tag{11}$$

where μ_i is the mean, β_i determines the shape of the PDF, α_i is the scale parameter of the GGD and $\Gamma(\cdot)$ is the gamma function. The variance σ_i^2 and the kurtosis κ_i if the GGD is given as

$$\begin{aligned} \sigma_i^2 &= \frac{\alpha_i^2 \Gamma(3/\beta_i)}{\Gamma(1/\beta_i)} \\ \kappa_i &= \frac{\Gamma(5/\beta_i)\Gamma(1/\beta_i)}{\Gamma(3/\beta_i)^2} - 3. \end{aligned} \tag{12}$$

Given a dataset $\mathcal{S}$, for GGD algorithm we need to estimate the mean $\hat{\mu}_i$, variance $\hat{\sigma}_i^2$ and kurtosis $\hat{\kappa}_i$ which are calculated as

$$\hat{\mu}_i = \frac{1}{N} \sum_{i=1}^{N} s_{n,i}, \quad i = 1, 2, \cdots, M, \tag{13}$$

$$\hat{\sigma}_i^2 = \frac{1}{N} \sum_{i=1}^{N} (s_{n,i} - \hat{\mu}_i)^2, \quad i = 1, 2, \cdots, M, \tag{14}$$

$$\hat{\kappa}_i = \frac{\frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}_i)^4}{\left[\frac{1}{N}\sum_{i=1}^{N} (s_i - \hat{\mu}_i)^2\right]^2} - 3, i = 1, 2, \cdots, M, \tag{15}$$

where the estimate of kurtosis $\hat{\kappa}_i$ can be used to measure the shape parameter β_i and estimate of variance $\hat{\sigma}_i^2$ can help to determine the scale parameter α_i of the GGD. Now the test data $\boldsymbol{u}$ can be employed to calculate the probability as

$$P(\boldsymbol{u}) = \prod_{i=1}^{M} P(u_i, \hat{\mu}_i, \hat{\alpha}_i, \hat{\beta}_i), \tag{16}$$

where the classification can be done with the help of equation(10). For the sake of clarity, the GGD algorithm is summarised in the next subsection.

C. Classification Algorithm

For the proposed classification technique, we first extract the LoS and NLoS signal features from the received dataset $\mathcal{S}$, then use distribution of each features to establish a model $P(\boldsymbol{s})$. After the model is built and the threshold is selected, we fit this model with the test dataset $\boldsymbol{u}$ and classify whether the signal experiences a LoS or NLoS signal. The steps of the training and testing stage of the proposed GGD algorithm are shown in Algorithm 1 and 2, respectively. As GGD is the more generalised algorithm we have only summarised it. For GD algorithm step 5 will not be required in Algorithm 1. For Algorithm 2 in step-1 we will replace equation (16) with equation (9).

IV. Experiment Setup

Algorithm 1 : Training Stage

Input: Collected dataset of UWB consisting of LoS and NLoS signal features.
Output: Create model $P(\boldsymbol{s}_n)$ for the n-th index of the dataset $\mathcal{S}$.
Algorithm

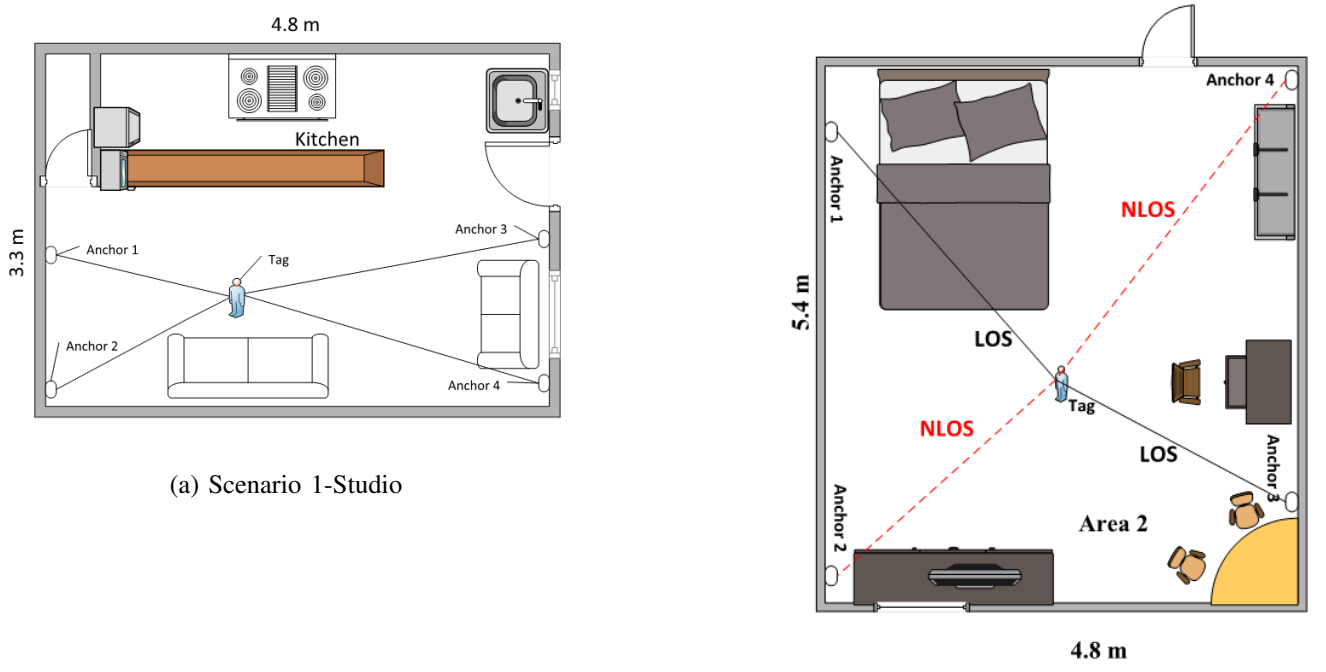
1) Initialize and pre-process the dataset.
2) Select part of LoS as training data.
3) Estimate the mean $\hat{\mu}_i$,
4) Estimate the variance $\hat{\sigma}_i^2$, and
5) Estimate the kurtosis κ_i,
6) Construct the model $P(s_i)$ for the i-th feature.
7) Select the threshold ϵ by calculating F_1-score value.
8) Construct the model $P(\boldsymbol{s}_n)$.

Algorithm 2 : Testing Stage

Input: Test dataset with a mixture of LoS and NLoS signals.
Output: Determine whether it is LoS or NLoS and then determine the exact location of the tag or the moving node.
Algorithm:

1) Fit the model by calculating the probability $P(\boldsymbol{u})$ of the testing data as mentioned in equation (16).
2) **If** $P(\boldsymbol{u}) \leq \epsilon \leftarrow$ NLoS signal.
3) **Else** $P(\boldsymbol{u}) > \epsilon \leftarrow$ LoS signal.
4) Determine the distance between the tag and the respective anchor using (5).

A. Data Extraction and Key Feature Selection Process



(a) Scenario 1-Studio

(b) Scenario 2-Room

Fig. 2: The scenarios where training and test data were collected for evaluation:(a) Scenario 1-Studio (b) Scenario 2-Room

TABLE I: Configurations of the MDEK-1001 UWB kit.

Properties	**Values**
Chip	DW 1000
Transceiver	DWM1000
Pulse shape	Gaussian pulse
Number of Anchors	4
Data Rate	6.8 Mbps
Frequency	3993.6 MHz
Bandwidth	499.2 MHz
Channel	2
Pulse Repetition Frequency (PRF)	16 MHz

In this paper, MDEK-1001 UWB kits from DECAWAVE company are employed for data preparation. The configuration of the UWB kit is shown in Table I. Two separately independent datasets for evaluation were collected in a small studio environment of dimensions (3.3×4.8) m as shown in Figure 2a and a room environment of size (4.8×5.4) m as shown in Figure 2b. We used the studio environment to train the model and evaluate the performance of the proposed algorithm. However, in order to evaluate the proposed algorithms, we also used the model built in the studio environment to test the room environment as a new environment. In both environment, during the LoS data collection, there was absolute clear environment between the anchors and tag. For NLoS data collection, there was an iron sheet placed between the two anchors and the tag to make the partial LoS and NLoS links, so that no direct path of signal can be transmitted or received by it. Due to placement of this iron piece, the main path of the signal was attenuated causing propagation delays. All the experiment is carried out using MATLAB (R2020b). The tag was connected to a PC and the data was logged via the Teraterm software into a comma-separated value (CSV) file. As mentioned, in this paper our focus is to classify the NLoS data in a unbalanced data. Therefore, 100 NLoS signals and 1000 LoS signals are randomly selected from the collected data. This selection results in a ratio of $1 : 0.1$ for the LoS and NLoS signal. Then, we randomly select different proportions of LoS and NLoS data to test the robustness of proposed algorithm. In total, 7 signal components are extracted that are:

1) The first path amplitude (F_1) of the UWB signal.
2) The second path amplitude (F_2) of the UWB signal.
3) The third path amplitude (F_3) of the UWB signal.
4) The preamble accumulation count value.
5) The amplitude of the channel impulse response (CIR).
6) The standard noise variance reported in the DW-1000 chipset.
7) The estimated distance.

In brief, among the mentioned NLoS identification methods in the literature,the threshold difference between the first-path power and received power have been widely used in different ML algorithms [16], [32], [33]. In our analysis we will use the above 7 signal components to calculate our 4 key features which are the estimated distance by using equation(5), first path power level by equation (17), received

power level by equation (18)and the power difference between the first and the received power level by equation (19). The first-path power level is calculated as [16], [34]

$$\text{FP Level} = 10 \times \log_{10}\left(\frac{F_1^2 + F_2^2 + F_3^2}{N^2}\right) - A \quad \text{dBm}, \tag{17}$$

where F_1, F_2, and, F_3 represents first, second, and third harmonics of the first-path signal amplitudes. A is a constant equivalent to 113.77 when a PRF is 16 MHz as mentioned in [34] (page-46), and N is the Preamble Accumulation Count value. The received power level of the signal can be computed as [34]

$$\text{RX Level} = 10 \times \log_{10}\left(\frac{CIR \times 2^{17}}{N^2}\right) - A \quad dBm, \tag{18}$$

where CIR is the Channel Impulse Response Power value. The power dissipation in NLoS environment is higher than the LoS environment due to multi-path effects, resulting in the first path of the LoS signal to have more power than the the first path of the NLoS signal. This knowledge can now be employed to improve the detection capability of the algorithm, therefore, the difference between the received and first-path power can also be employed. The formula is as shown in

$$\text{Threshold Power = RX Level - FP Level} \tag{19}$$

Figure 3 and 4 show the probability density functions (pdf)s of the selected four features which are the estimated distance, first-path power, received path power, and the power difference for LoS and NLoS respectively. Histogram function is employed to generate the pdfs of these features and are represented by the blue bars. The GD and GGD distribution is plotted by using the equation (6) and (16). For GD distribution, only mean and variance of the data is required. However, for GGD distribution, mean, variance, and kurtosis is required. The distribution of the features can be more closely approximated with the GGD as it has three parameters to update as compared to GD which only has two parameters. From these figures, it can be seen that the selected features follow GD and GGD distribution, therefore, the proposed algorithms can be used for the data classification. Let us now look into designing the selection threshold for these pdfs.

B. Threshold Selection, ϵ

In order to classify the LoS and NLoS signals by proposed algorithm, we need to select an appropriate threshold ϵ as mentioned in III. For threshold ϵ selection, we start by constructing a training set based on the LoS data, then apply the cross-validation to the remaining LoS and NLoS datasets. According to the training data set, we estimate the mean, variance and kurtosis of each features to build function $P(\boldsymbol{s}_n)$ as mentioned in algorithm 1. The threshold ϵ was chosen based on the F-score. F-Score is defined as the weighted average of precision and recall and calculated as

$$\text{F-Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}, \tag{20}$$

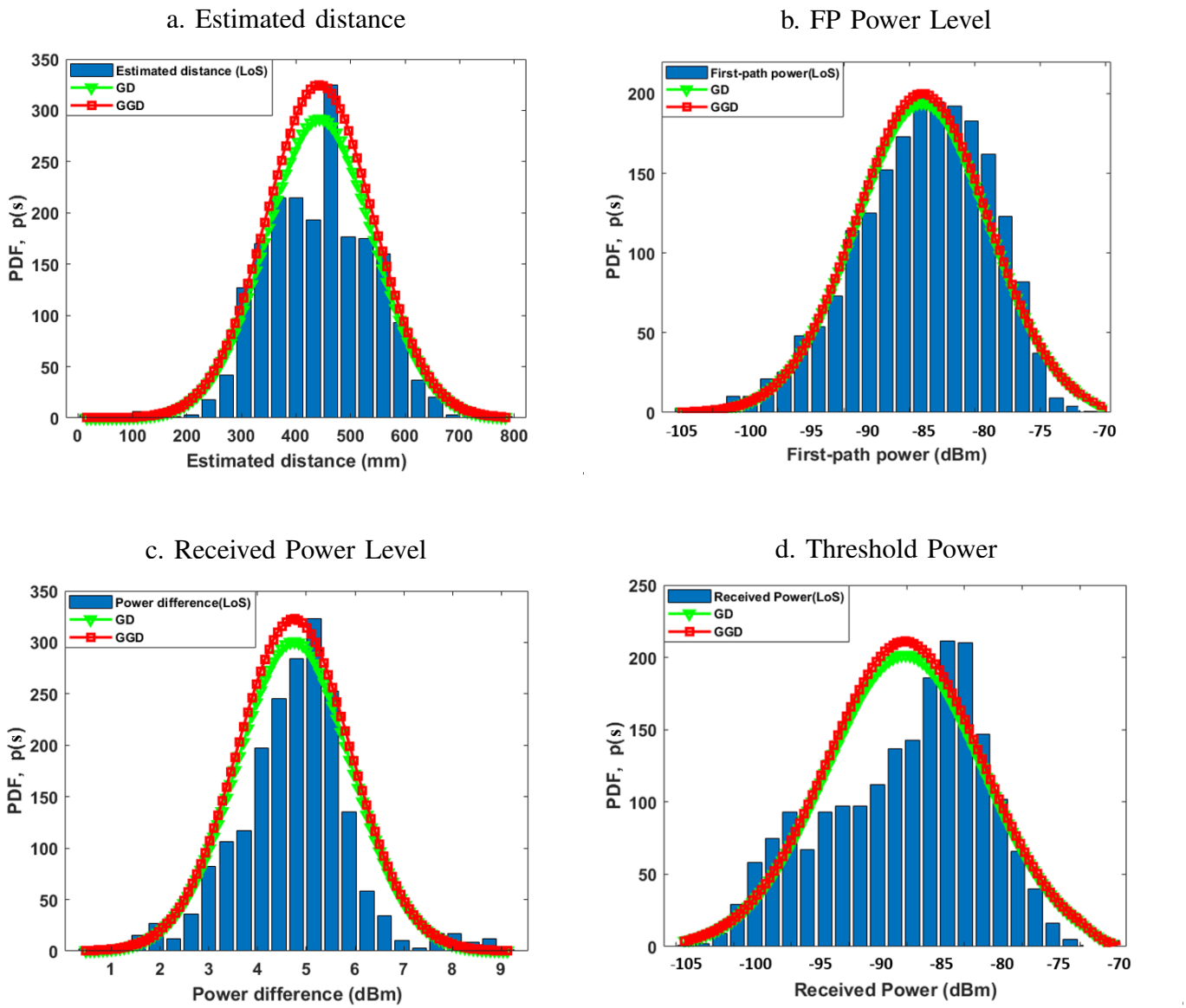


Fig. 3: Histogram distribution of our 4 key features in LoS environment a) Estimated distance, b) first path power level, c) received power level, and d) the threshold power.

where Precision and Recall are defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{21}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{22}$$

where TP is the True Positive, FP is the False Positive, FN is the False Negative, and TN is the True Negative, respectively. TP means that the instances are classified as positive when they are actually positive, TN illustrates the instances are classified as negative when they are in negative condition. FP shows that the instances are classified as positive when they are negative. Similarly, FN represents the instances classified as negative when they are actually positive. However, the environment will not remain the stable situation, therefore, We added a forgetting factor α for the calculation of threshold ϵ. In such case, the threshold can be updated after training and therefore could not incorporate the moving objects

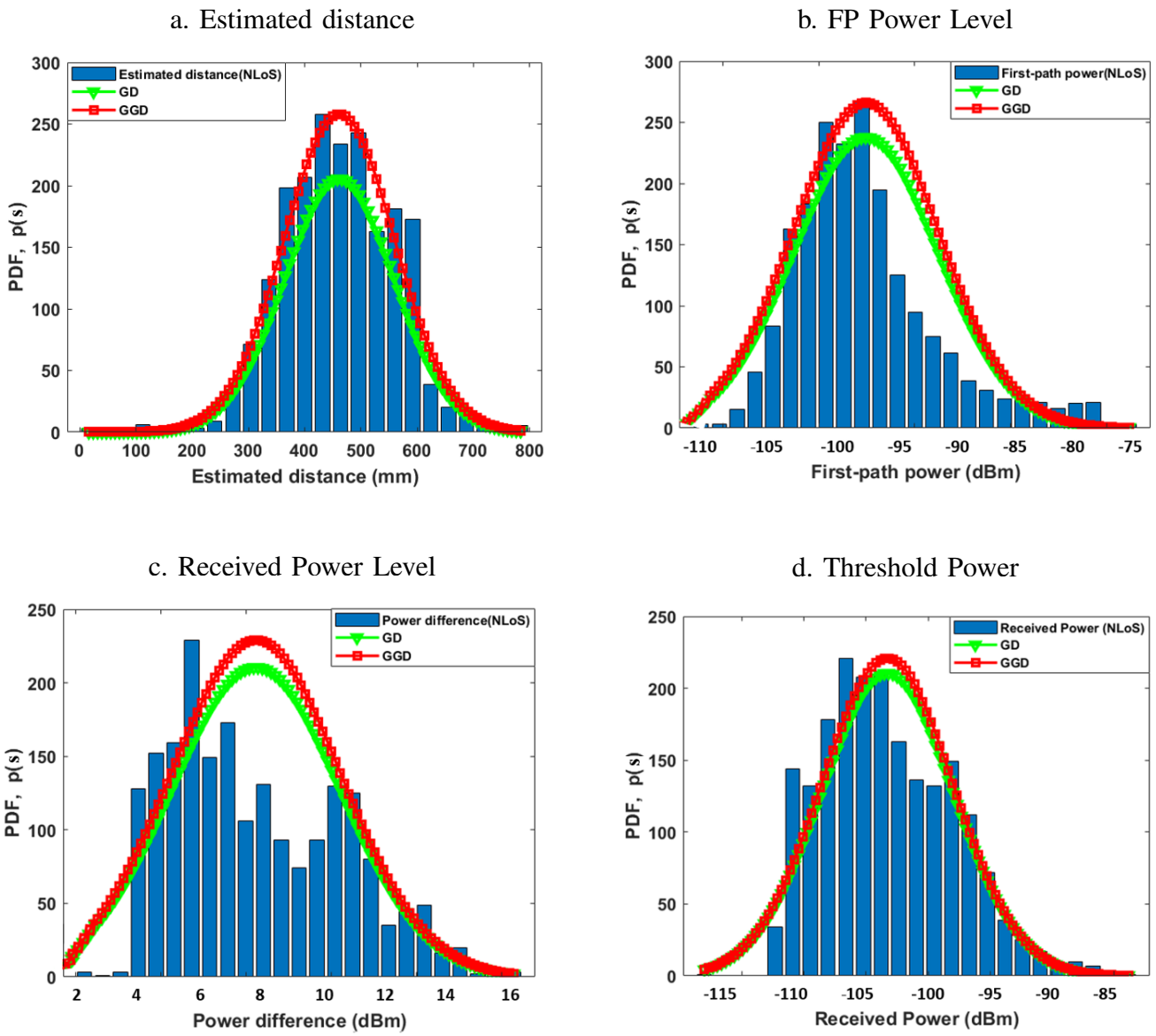


Fig. 4: Histogram distribution of our 4 key features in NLoS environment a) Estimated distance, b) first path power level, c) received power level, and d) the threshold power.

in the field. The threshold can be calculated by the following equation

$$\epsilon_{t+1} = (1-\alpha)\epsilon_t + \alpha \times \epsilon_t, \tag{23}$$

where t is the time index. By this method, threshold get updated in the training phase.

V. Performance Evaluation

In this section, we examine the performance of proposed algorithm. We first compare the results of the classifier with selected threshold values. Second, we compare the results with the state-of-the-art machine learning algorithm based on SVM, DT, NB and NN, then we start by computing two quantitative metrics: (i) confusion matrix that shows the classification results of each individual classifier, and (ii) Receiver Operating Characteristics (ROC) curve and corresponding Area-Under-the-ROC curve (AUC) value. Third,

we calculate the performance of these algorithms in terms of precision, recall and accuracy. Finally, the effect of different ratios of LoS and NLoS dataset is studied for these algorithms.

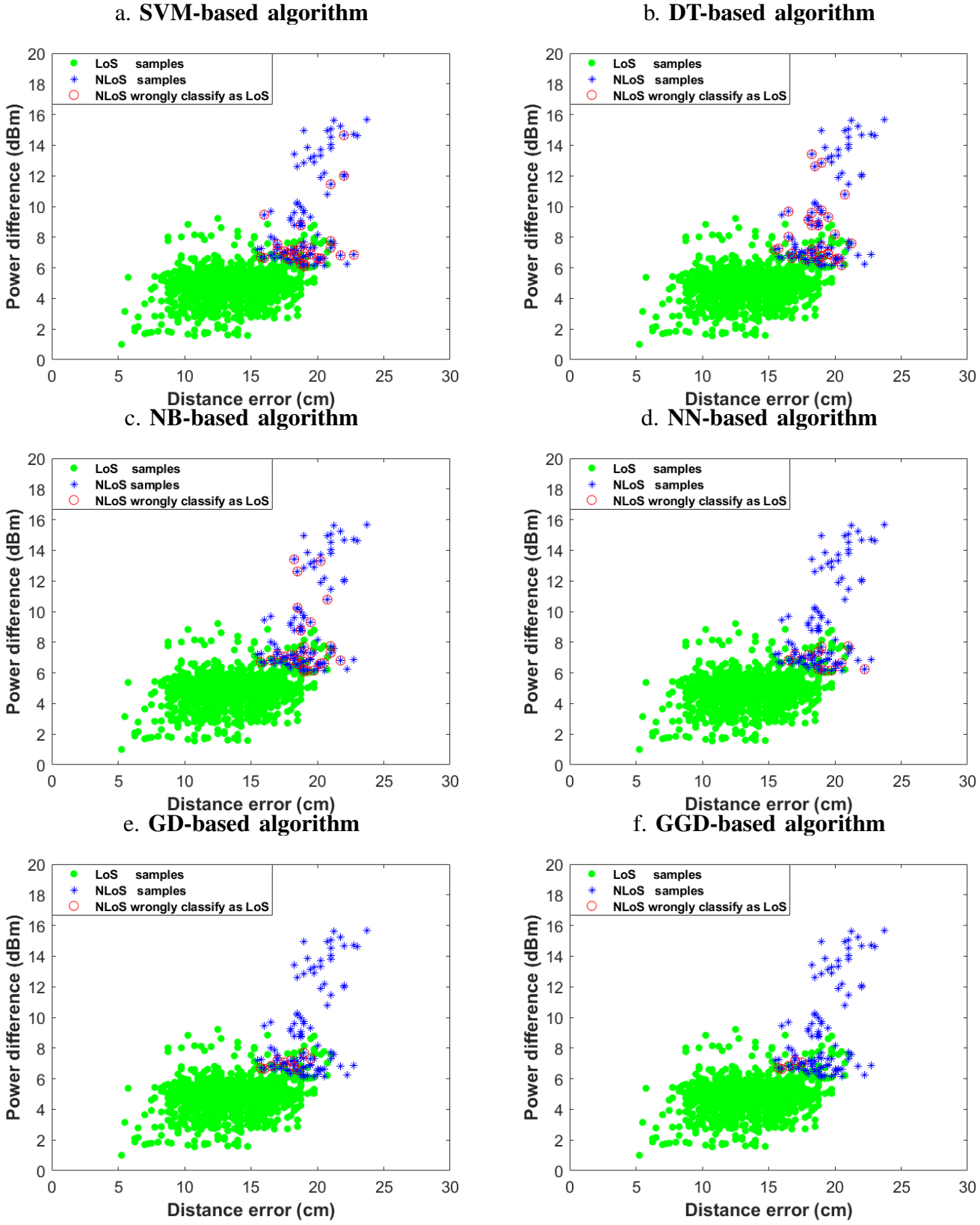


Fig. 5: Visualization of samples a) SVM-, b) DT-, c) NB-, d) NN-, e) GD-, and f) GGD-based algorithms.

The visualization of the samples are shown in Figure 5. The virtualization of the samples is plotted with the help of the power difference calculated using equaiton 19 and the distance error that is calculated in centimeters. There are 1000 LoS and 100 NLoS signals respectively. In this figure, the green samples represents the LoS signals and the blue coloured represents the NLoS signals. The blue samples with red circle are the NLoS samples which have been falsely classify as LoS. It can be observed from these figures that the proposed GGD algorithm can provide a higher classification by setting an appropriate threshold by training the features as mentioned in section II as compared to the GD and the classical ML algorithms. For the SVM-, DT-, NB- and NN-based algorithms the performance is relatively poor as compared to GD and the GGD algorithm this is due to the limited number of NLoS signals in the dataset failed to train a robust model for classification. Finally, it can be concluded that for unbalanced dataset the GD and GGD performs exceptionally better in classifying the LoS and NLoS signals as compared to conventional ML algorithms such as SVM-, DT-, NB-, and NN- based algorithms.

TABLE II: Dataset of the LoS and NLoS signals

Signal	**Distance m**	**RX level dBm**	**FP level dBm**	**PD dBm**	**Classified**
LoS	7.92	-92.7187	-89.6319	3.027	LoS
LoS	9.98	-94.155	-89.553	4.603	NLoS
NLoS	7.22	-95.273	-88.401	6.87	NLoS
NLoS	7.38	-93.96	-88.5263	5.434	LoS

The Table II shows an example of LoS and NLoS signal for the proposed UWB system. From the table it can be observed that we have distance and three features that are RX power level, FP power level, and PD power levels followed by how it is classified by the algorithm. From the table, it can be observed that a single feature cannot be simply employed for classification of a signal as LoS or NLoS.

Figure 6 plots the confusion matrix of the four supervised ML algorithms (SVM, DT, ND, and NN) followed by the two proposed algorithms GD and GGD. All these algorithms are based on 1000 LoS and 100 NLoS signals. From these confusion matrices, it can be concluded that the worst performance in terms of True Positive Rate (TPR) for LoS components is achieved by SVM which is 94.9%. NB algorithm performs 0.01% better than DT in terms of TPR, however, NN algorithm achieves the best performance of 96.1% as compared to traditional ML algorithms. We can observe that the proposed GD and GGD performance is better than the existing ML algorithms and in terms of TPR in LoS components is 97.3% and 98.3%, respectively. Similarly, for NLoS components the performance of NN is much superior as compared to the SVM, DT, and NB. Both SVM and DT can correctly classify only 69% of NLoS components. NB classifies only 72% of TN. NN can classify 89% while GD and GGD can classify more than 90% NLoS components accurately. Therefore, from figure 6it can be observed that for both the LoS and NLoS components classification GD and GGD algorithms perform much superior to the classical ML algorithms such as SVM, DT, NB, and NN.

Figure 7 plots the receiver operating characteristics (ROC) curve. With these ROC curves, area under the curve (AUC) can be also studied for the proposed and ML based positioning algorithms. The ROC curve is plotted with respect to the true positive rate (TPR) versus the false positive rate (FPR). Generally, in a ROC curve the best classifier is closer to the upper left corner, resulting in a larger AUC. From

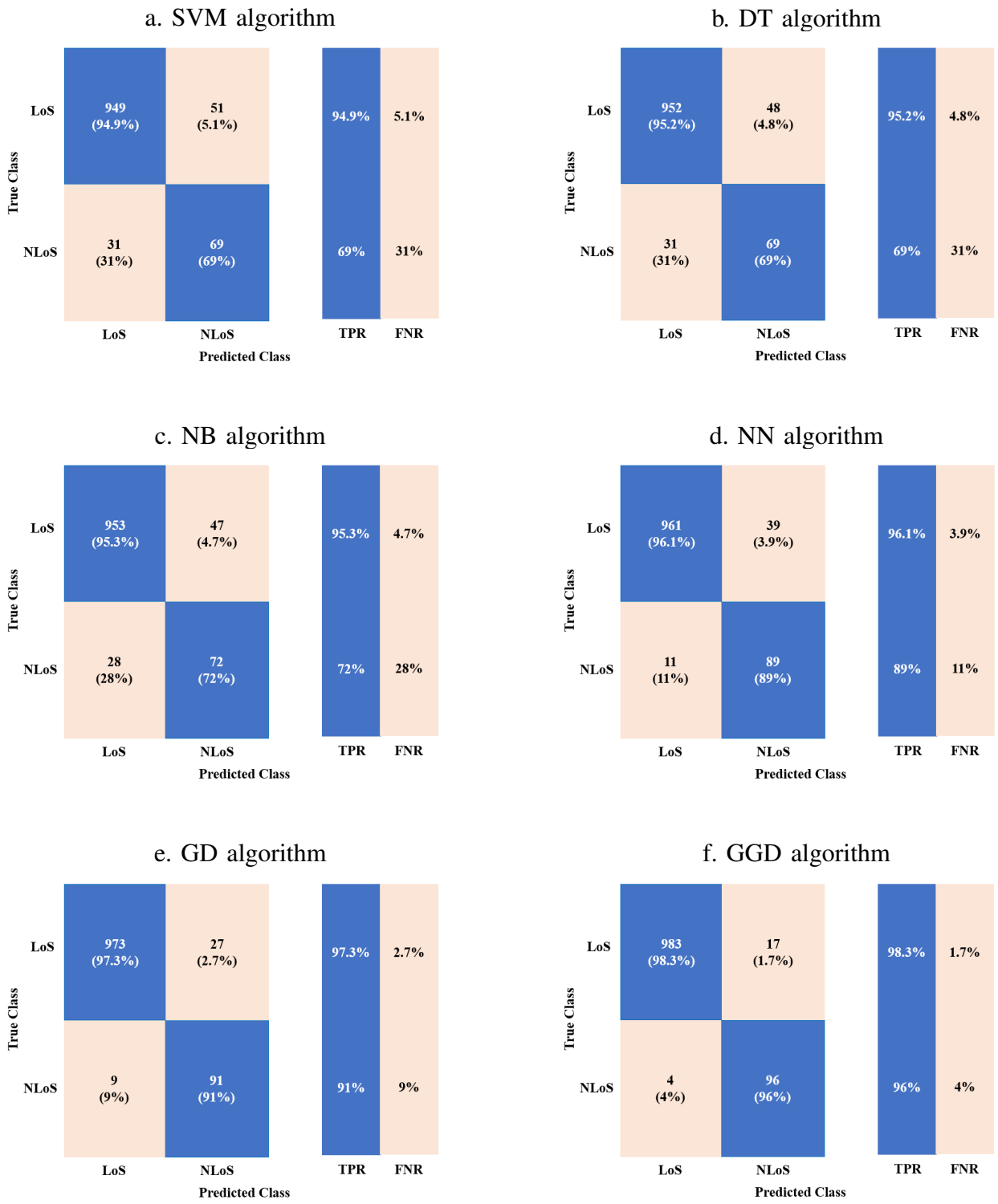


Fig. 6: The confusion matrix of the validation data set on six algorithms a) SVM, b) DT, c) NB, d) NN, e) GD, and f) GGD algorithms. The proposed algorithms improve the classification significantly.

figure 7 it can be observed that the GGD algorithm is closer to the upper left corner as compared to other algorithms. Furthermore, the AUC of GGD algorithm is 0.982 which is more than any compared algorithm. As a result, the overall GGD algorithm will be superior in terms of classification accuracy as compared to other algorithms. The second best performance was achieved by the GD algorithm that was

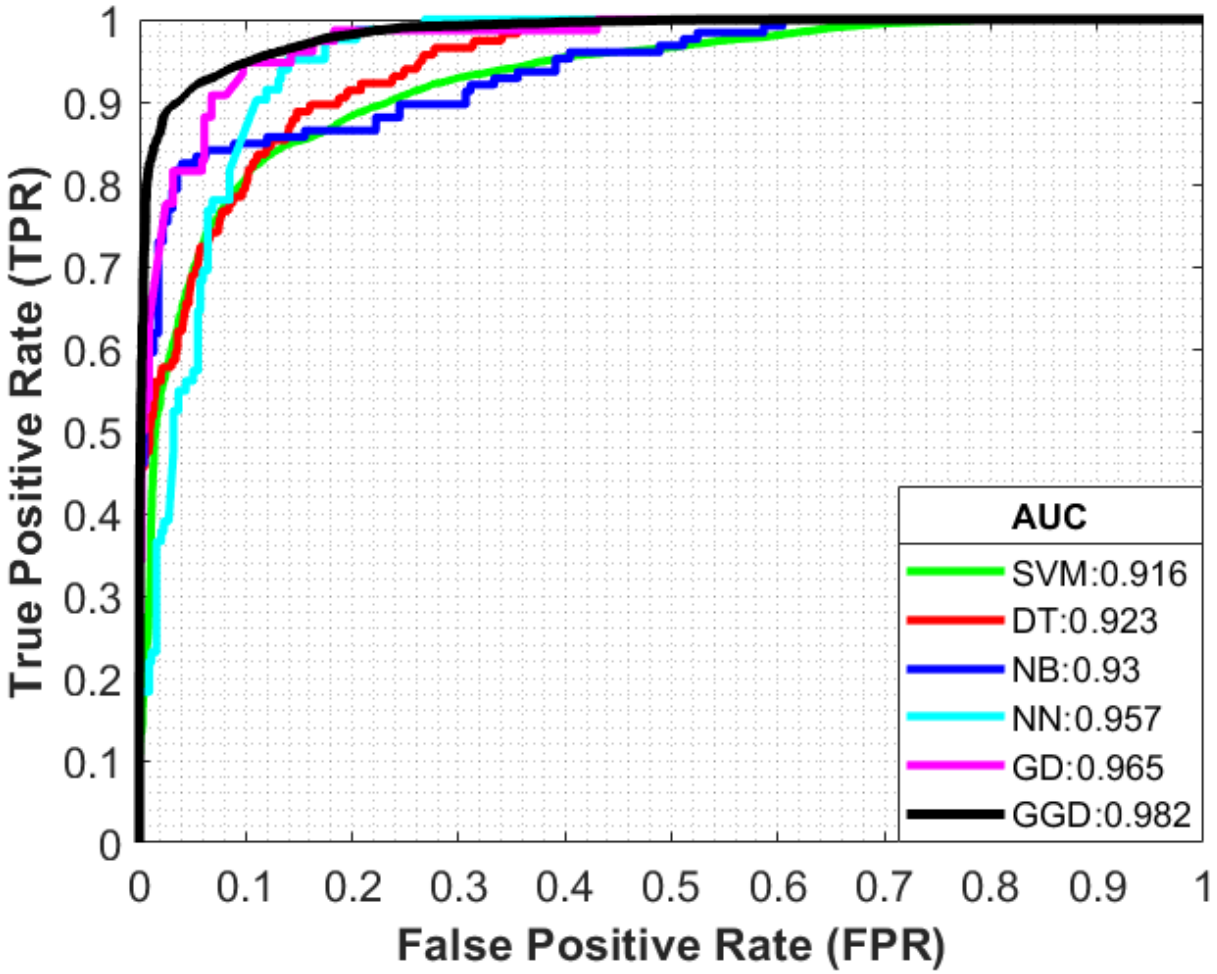


Fig. 7: Receiver Operating Characteristics (ROC) and Area under the curve (AUC) comparison of the algorithms

around 0.965 and the SVM performance was the worst despite achieving a value of 0.916. Out of all the ML algorithms NN performed superior as the area under the curve was equivalent to 0.957. Finally, it can be observed that the proposed algorithms, therefore GD and GGD, perform superior to the other ML algorithms as they can classify the data more accurately.

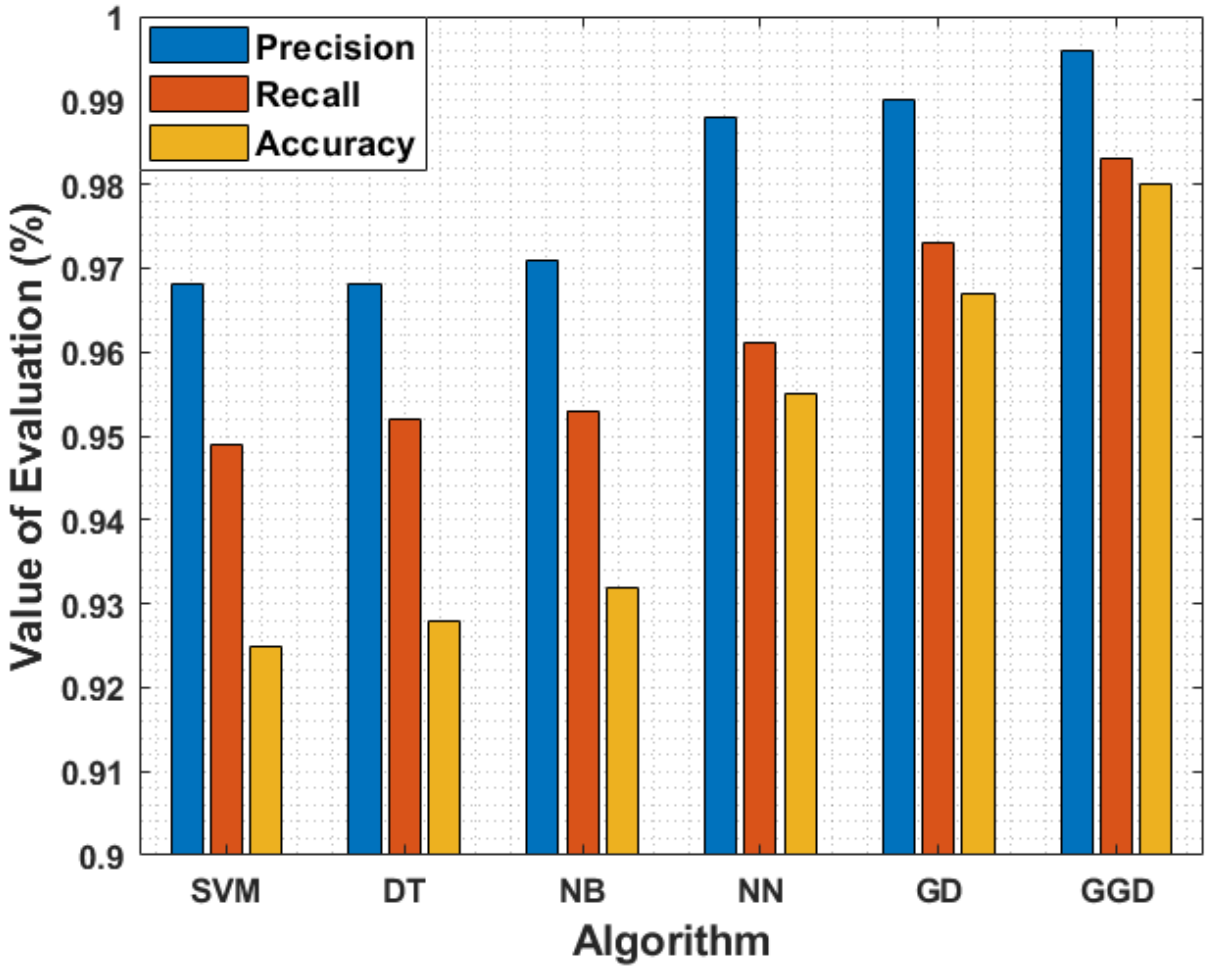


Fig. 8: Performance comparison in terms of Precision, Recall, and Accuracy.

Figure 8 shows the evaluation results of these algorithms in terms of precision, recall, and accuracy. From the figure, it can be observed that the accuracy of the SVM-based algorithm is equivalent to 92.6%. The accuracy of the DT-based, NB-based, and NN-based algorithms is 92.8%, 93.2%, and 95.5%, respectively. The GD-based algorithm results in more than 96%. The overall accuracy achieved with the GGD-based algorithm is around 98.0% which shows the GGD-based algorithm is superior to other algorithms. For precision and recall, it can be seen that the proposed algorithm also performance better than the other algorithms which shows that the proposed can provide the high accuracy classification accuracy of NLoS and LoS.

Figure 9 graphically shows the classification accuracy of different ratios of LoS and NLoS samples for the classical ML, GD, and GGD algorithms. The reason for plotting this curve is to see the effect of the imbalanced data set. From the figure, it can be observed that we have five different ratios $0.1, 0.2, 0.5, 0.8$, and 1.0, respectively. With a ratio of 0.1, we will have 1000 LoS and 100 NLoS signals. However, as the ratio increases the number of NLoS signals increases. Therefore, for a ratio of 1, we have 1000 LoS and 1000 NLoS signals. From the figure, it can be observed that the classification accuracy significantly improves as the ratio in the imbalanced datasets increase. For the SVM algorithm, the improvement in

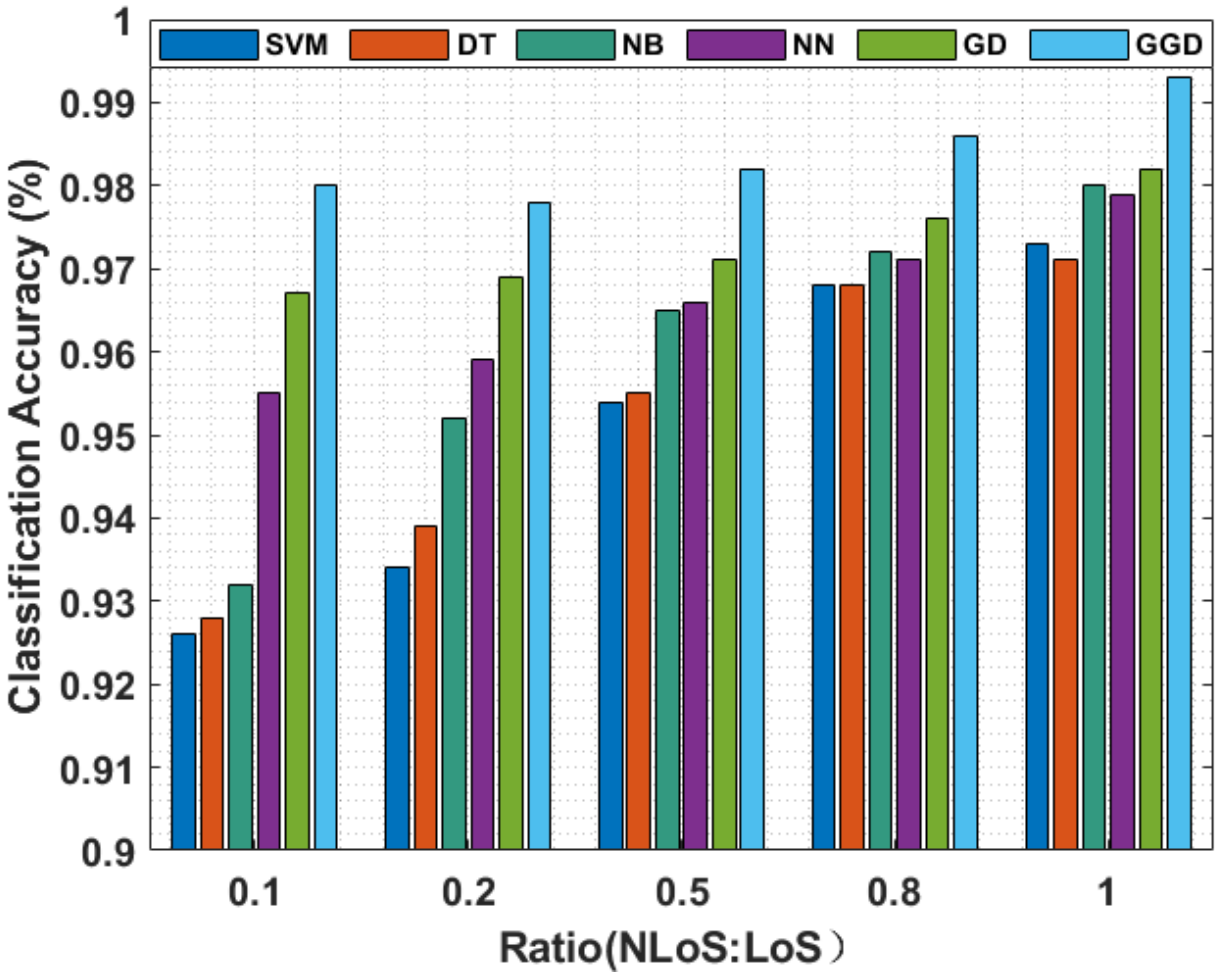


Fig. 9: Classification accuracy with different ratios of LoS and NLoS for different algorithms.

classification accuracy performance is significantly more as compared to the other algorithms. For a ratio of 0.1, the accuracy is around 92.6%, While the maximum accuracy is achieved when we have a balanced dataset with a ratio of 1, therefore, around 97.3%. The DT,NB and NN algorithms also achieved higher accuracy when the dataset tend to balanced. For the GD algorithm, the classification accuracy improves from 96.7% to 98.2% when the imbalance in the dataset is improved. Finally, for the GGD algorithm, the worst classification is around 98.0% and the best is around 99.3% for the balanced dataset. The total improvement in classification accuracy is around 3.7% for the SVM algorithm, 1.5% for the GD algorithm, and 1.3% for the GGD algorithm. This indicates that the GGD algorithm is more robust as compared to conventional algorithms for both imbalanced or balanced dataset. Finally, this simulation result proves that the GD and GGD can guarantee better results for NLoS identification under different situations compared to the conventional algorithms especially when the dataset is imbalanced.

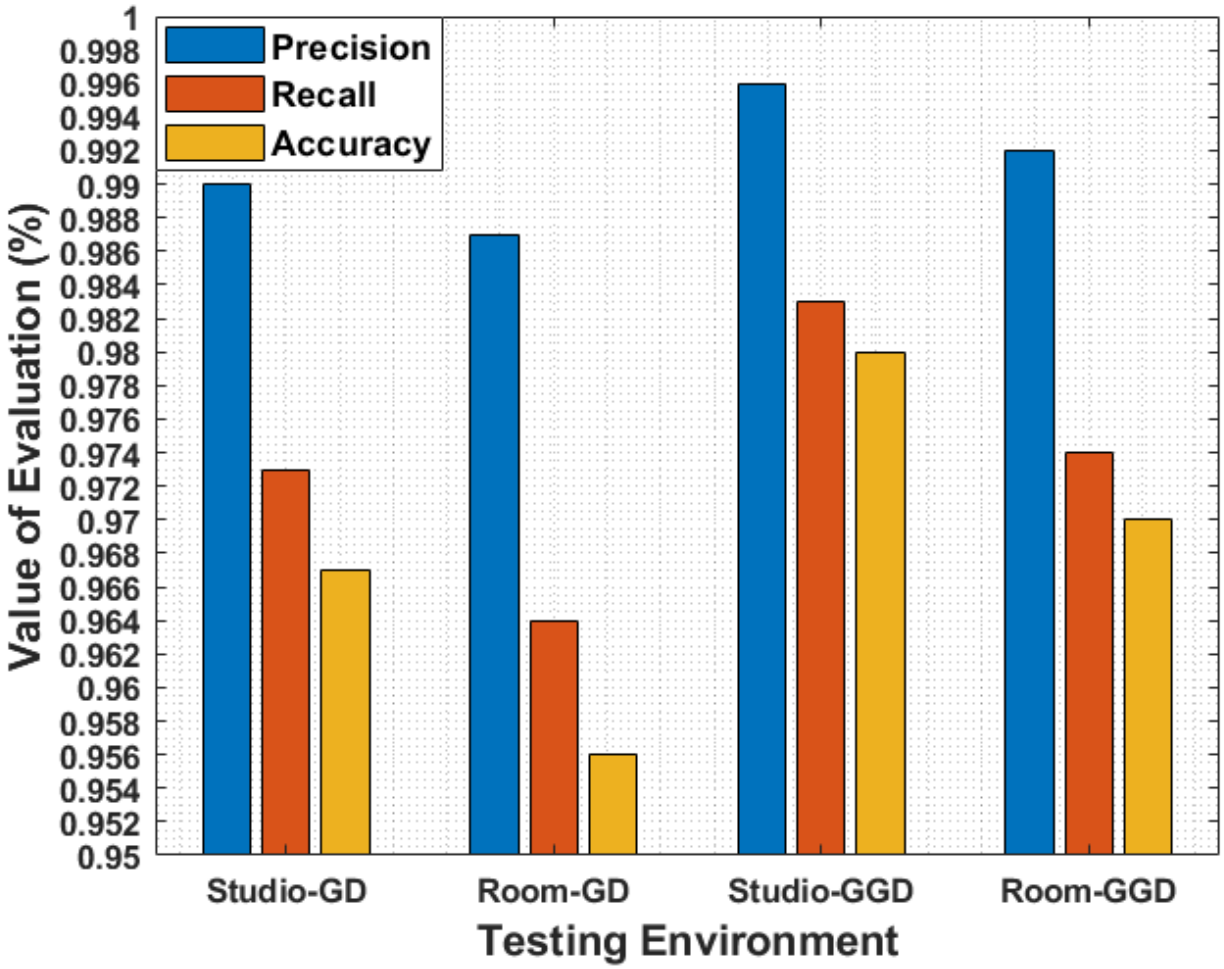


Fig. 10: Accuracy results of two different environments

TABLE III: Performance comparison of the two different environments.

Algorithm	**Training Scenario**	**Testing Scenario**	**TP**	**FP**	**FN**	**TN**	**AUC**
GD	Studio	Studio	973	27	9	91	0.965
GD	Studio	Room	964	36	11	89	0.96
GGD	Studio	Studio	983	17	4	96	0.982
GGD	Studio	Room	974	26	7	93	0.975

As mentioned, in order to examine the proposed algorithm, we carried out experiments in two different scenarios. The configuration of the UWB devices remained the same during the measurements. We also applied the trained model in the studio and test it in the room environment. The results are shown in Table III and plotted in Figure 10. From Table III and Figure 10 it can be observed from the figure and the data that the accuracy reduced by 0.07 and 0.08 when employing the GD- and GGD- based algorithm, respectively. Therefore, the performance of GD and GGD-based algorithms is not impacted much by changing the environment. It can be conclude that the proposed algorithms can remain the high accuracy when applied to a new environment.

VI. Conclusions

In this paper, a featured-based method for the UWB localisation is proposed. The main aim is to improve the classification of the UWB IPS especially when the dataset is imbalanced, therefore, having a large number of LoS signals as compared to NLoS signals. Initially, in this work seven UWB signal components were collected, and based on these seven signal components, four key features were selected which are estimated distance, first path power level, received power level, and threshold power. With these key features, the joint probability densities for Gaussian and generalised Gaussian are calculated. The confusion matrix, ROC, and AUC area for these ML algorithms were compared to evaluate the results. From the simulation and experimental results, it can be observed that the performance of the NLoS classification was significantly improved by designing the GD and the GGD algorithms as compared to the existing state-of-the-art work in the imbalanced LoS and NLoS situation. In addition, the proposed algorithm is effective for NLoS signal identification with balanced Los-NLoS mixed data and remain highly accurate even if we have unbalanced data. For future work, the data can be extended to a large dataset with more signal features to evaluate the classification accuracy. Furthermore, the testing environment could be setted in the lab, office and warehouse environments.

NOMENCLATURE

Angle-of-Arrival	- AOA
Area under the curve	- AUC
Additive White Gaussian Noise	- AWGN
Boosted Decision Tree	- BDT
Channel Impulse Response	- CIR
Conventional Neural Network	- CNN
Comma-Separated Value	- CSV
Decision Tree	- DT
False Negative	- FN
False Positive	- FP
False Positive Rate	- FPR
Gaussian Distribution	- GD
Generalized Gaussian Distribution	- GGD
Global Navigation Satellite System	- GNSS
Inertial Navigation System	- INS
Internet of Things	- IoTs
Indoor Positioning System	- IPS
Line-of-Sight	- LoS
Import Vector Machine	- IVM
K-Nearest Neighbor	- KNN
Multi-Layer Perceptron	- MLP
Naïve Bayes	- NB
Neural Network	- NN
Non-Line-of-Sight	- NLoS
Probability Distribution Function	- PDF
Receiver Operating Characteristics	- ROC
Support Vector Machine	- SVM
Time-of-Flight	- ToF
True Positive	- TP
True Positive Rate	- TPR
Ultra-wideband	- UWB

REFERENCES

[1] B. Pinto, R. Barreto, E. Souto and H. Oliveira, “Robust RSSI-Based Indoor Positioning System Using K-Means Clustering and Bayesian Estimation,” in IEEE Sensors Journal, vol. 21, no. 21, pp. 24462-24470, 1 Nov.1, 2021.

[2] W. Jiang, Z. Cao, B. Cai, B. Li and J. Wang, “Indoor and Outdoor Seamless Positioning Method Using UWB Enhanced Multi-Sensor Tightly-Coupled Integration,” in IEEE Transactions on Vehicular Technology, vol. 70, no. 10, pp. 10633-10645, Oct. 2021.

[3] Z. Wu, Y. Yue, M. Wen, J. Zhang, J. Yi and D. Wang, “Infrastructure-Free Hierarchical Mobile Robot Global Localization in Repetitive Environments,” in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-12, 2021.

[4] Y. Wu, R. Chen, W. Li, Y. Yu, H. Zhou and K. Yan, “Indoor Positioning Based on Walking-Surveyed Wi-Fi Fingerprint and Corner Reference Trajectory-Geomagnetic Database,” in IEEE Sensors Journal, vol. 21, no. 17, pp. 18964-18977, 1 Sept.1, 2021

[5] L. Alsmadi, X. Kong, K. Sandrasegaran and G. Fang, "An Improved Indoor Positioning Accuracy Using Filtered RSSI and Beacon Weight," in IEEE Sensors Journal, vol. 21, no. 16, pp. 18205-18213, 15 Aug.15, 2021

[6] I. M. Abou-Shehada, A. F. AlMuallim, A. K. AlFaqeh, A. H. Muqaibel, K. -H. Park and M. -S. Alouini, "Accurate Indoor Visible Light Positioning Using a Modified Pathloss Model With Sparse Fingerprints," in Journal of Lightwave Technology, vol. 39, no. 20, pp. 6487-6497, Oct.15, 2021

[7] Q. Pu, J. K. -Y. Ng and M. Zhou, "Fingerprint-Based Localization Performance Analysis: From the Perspectives of Signal Measurement and Positioning Algorithm," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-15, 2021.

[8] C. Wang, A. Xu, J. Kuang, X. Sui, Y. Hao and X. Niu, "A High-Accuracy Indoor Localization System and Applications Based on Tightly Coupled UWB/INS/Floor Map Integration," in IEEE Sensors Journal, vol. 21, no. 16, pp. 18166-18177, 15 Aug.15, 2021.

[9] S. Hayat, E. Yanmaz and R. Muzaffar, "Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint," in IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2624-2661, Fourthquarter 2016.

[10] P. Chhikara, R. Tekchandani, N. Kumar, V. Chamola and M. Guizani, "DCNN-GA: A Deep Neural Net Architecture for Navigation of UAV in Indoor Environment," in IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4448-4460, 15 March15, 2021.

[11] W. Wang, D. Marelli and M. Fu, "Multiple-Vehicle Localization Using Maximum Likelihood Kalman Filtering and Ultra-Wideband Signals," in IEEE Sensors Journal, vol. 21, no. 4, pp. 4949-4956, 15 Feb.15, 2021.

[12] C. T. Nguyen et al., "A Comprehensive Survey of Enabling and Emerging Technologies for Social Distancing—Part I: Fundamentals and Enabling Technologies," in IEEE Access, vol. 8, pp. 153479-153507, 2020.

[13] V. Chamola, V. Hassija, S. Gupta, A. Goyal, M. Guizani and B. Sikdar, "Disaster and Pandemic Management Using Machine Learning: A Survey," in IEEE Internet of Things Journal, vol. 8, no. 21, pp. 16047-16071, 1 Nov.1, 2021.

[14] N. Kbayer and M. Sahmoudi, "Performances Analysis of GNSS NLOS Bias Correction in Urban Environment Using a Three-Dimensional City Model and GNSS Simulator," in IEEE Transactions on Aerospace and Electronic Systems, vol. 54, no. 4, pp. 1799-1814, Aug. 2018.

[15] B. Hanssens et al., "An Indoor Variance-Based Localization Technique Utilizing the UWB Estimation of Geometrical Propagation Parameters," in IEEE Transactions on Antennas and Propagation, vol. 66, no. 5, pp. 2522-2533, May 2018.

[16] F. Che, A. Ahmed, Q. Z. Ahmed, S. A. R. Zaidi and M. Z. Shakir, "Machine Learning Based Approach for Indoor Localization Using Ultra-Wide Bandwidth (UWB) System for Industrial Internet of Things (IIoT)," 2020 International Conference on UK-China Emerging Technologies (UCET), 2020, pp. 1-4.

[17] W. B. Abbas, F. Che, Q. Z. Ahmed, F. A. Khan and T. Alade, "Device Free Detection in Impulse Radio Ultrawide Bandwidth Systems," Sensors, vol. 21, no. 9, pp. 1–19, Sept. 2021.

[18] L. Yan, Y. Lu and Y. Zhang, "An Improved NLOS Identification and Mitigation Approach for Target Tracking in Wireless Sensor Networks," in IEEE Access, vol. 5, pp. 2798-2807, 2017.

[19] K. Gururaj, A. K. Rajendra, Y. Song, C. L. Law and G. Cai, "Real-time identification of NLOS range measurements for enhanced UWB localization," 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 2017, pp. 1-7.

[20] A. Abolfathi Momtaz, F. Behnia, R. Amiri and F. Marvasti, "NLOS Identification in Range-Based Source Localization: Statistical Approach," in IEEE Sensors Journal, vol. 18, no. 9, pp. 3745-3751, 1 May1, 2018.

[21] Q. Tian, K. I. Wang and Z. Salcic, "An INS and UWB Fusion Approach With Adaptive Ranging Error Mitigation for Pedestrian Tracking," in IEEE Sensors Journal, vol. 20, no. 8, pp. 4372-4381, 15 April15, 2020.

[22] S. Maranò, W. M. Gifford, H. Wymeersch and M. Z. Win, "NLOS identification and mitigation for localization based on UWB experimental data," in IEEE Journal on Selected Areas in Communications, vol. 28, no. 7, pp. 1026-1035, September 2010.

[23] H. Wymeersch, S. Marano, W. M. Gifford and M. Z. Win, "A Machine Learning Approach to Ranging Error Mitigation for UWB Localization," in IEEE Transactions on Communications, vol. 60, no. 6, pp. 1719-1728, June 2012.

[24] B. Chitambira, S. Armour, S. Wales and M. Beach, "NLOS Identification and Mitigation for Geolocation Using Least-Squares Support Vector Machines," 2017 IEEE Wireless Communications and Networking Conference (WCNC), 2017, pp. 1-6.

[25] F.Che, A. Ahmed., Q. Z. Ahmed, and M. Z. Shakir," Artificial intelligence for localisation of ultra-wide bandwidth (UWB) sensor nodes, "In AI for Emerging Verticals: Human-Robot Computing, Sensing and Networking (2020 ed.). IET

[26] S. Krishnan, R. Xenia Mendoza Santos, E. Ranier Yap and M. Thu Zin, "Improving UWB Based Indoor Positioning in Industrial Environments Through Machine Learning," 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), 2018, pp. 1484-1488.

[27] Q. Zheng et al., "Channel Non-Line-of-Sight Identification Based on Convolutional Neural Networks," in IEEE Wireless Communications Letters, vol. 9, no. 9, pp. 1500-1504, Sept. 2020, doi: 10.1109/LWC.2020.2994945.

[28] J. Fontaine, M. Ridolfi, B. Van Herbruggen, A. Shahid and E. De Poorter, "Edge Inference for UWB Ranging Error Correction Using Autoencoders," in IEEE Access, vol. 8, pp. 139143-139155, 2020.

[29] Q. Z. Ahmed, K. Park and M. Alouini, “Ultrawide Bandwidth Receiver Based on a Multivariate Generalized Gaussian Distribution,” in IEEE Transactions on Wireless Communications, vol. 14, no. 4, pp. 1800-1810, April 2015.

[30] Q. Z. Ahmed, L. Yang and S. Chen, “Reduced-Rank Adaptive Least Bit-Error-Rate Detection in Hybrid Direct-Sequence Time-Hopping Ultrawide Bandwidth Systems,” in IEEE Transactions on Vehicular Technology, vol. 60, no. 3, pp. 849-857, March 2011.

[31] Q. Z. Ahmed and L. Yang, “Reduced-rank adaptive multiuser detection in hybrid direct-sequence time-hopping ultrawide bandwidth systems,” in IEEE Transactions on Wireless Communications, vol. 9, no. 1, pp. 156-167, January 2010.

[32] C.L. Sang, B. Steinhagen, J. D. Homburg, M. Adams, M. Hesse, U. Rückert,“ Identification of NLOS and Multi-Path Conditions in UWB Localization Using Machine Learning Methods, ” in Applied Sciences 2020, vol. 10, no. 11, pp. 3980-4005, June 2020.

[33] V. F. Miramá, L. E. Díez, A. Bahillo, and V. Quintero, “A Survey of Machine Learning in Pedestrian Localization Systems: Applications, Open Issues and Challenges,” in IEEE Access, vol. 9, pp. 120138-120157, 2021.

[34] DW1000 User Manual. Available online: https://www.decawave.com/sites/default/files/resources (accessed on 07/12/2021).