# A Multi-Scale Feature Selection Framework for WiFi Access Points Line-of-sight Identification

Xu Feng
*Computing & Maths Division*
*University of Brighton*
Brighton BN2 4GJ, United Kingdom
X.Feng@brighton.ac.uk

Khuong An Nguyen
*Computing & Maths Division*
*University of Brighton*
Brighton BN2 4GJ, United Kingdom
K.A.Nguyen@brighton.ac.uk

Zhiyuan Luo
*Department of Computer Science*
*Royal Holloway University of London*
Surrey TW20 0EX, United Kingdom
Zhiyuan.Luo@rhul.ac.uk

***Abstract*—Despite its high accuracy in the ideal condition where there is a direct line-of-sight between the Access Points and the user, most WiFi indoor positioning systems struggle under the non-line-of-sight scenario. Thus, we propose a novel feature selection algorithm leveraging Machine Learning based weighting methods and multi-scale selection, with WiFi RTT and RSS as the input signals. We evaluate the algorithm performance on a campus building floor. The results indicated an accuracy of 93% line-of-sight detection success with 13 Access Points, using only 3 seconds of test samples at any moment; and an accuracy of 98% for individual AP line-of-sight detection.**

***Index Terms*—WiFi Round-Trip Time, indoor positioning, feature selection.**

## I. Introduction

Identifying line-of-sight (LOS) and non-line-of-sight (NLOS) condition is crucial for WiFi-based indoor positioning systems, as the WiFi signals attenuate greatly indoors [1]. In such environment, popular WiFi signal measures such as Round-Trip Time (RTT) and Received Signal Strength (RSS) become unstable and unpredictable, resulting in large positioning errors [2], [3].

Most existing approaches for LOS Access Points (APs) detection manually select several sets of signal statistical features. Such manual selection cannot guarantee an optimal feature set since it is highly time-consuming to enumerate all possible combinations of different features. Furthermore, the manual approaches extract the same features for all APs, although only certain features of some particular APs are informative to the LOS detection task [4]. As for the generalisation and transferability, manual selection lacks the efficiency for the implementation in large-scale datasets.

To this end, we propose a framework to identify the LOS conditions of individual WiFi AP or all of them simultaneously. Our framework contains a novel feature selection algorithm leveraging Machine Learning and weighting methods. We implemented several feature selection models for the evaluation of feature importance. A multi-scale selection (MSS) method was proposed for optimising the selected features. This algorithm helps the indoor positioning system choose a small set of the most meaningful RTT and RSS features. For implementation on heterogeneous devices (e.g., smartphones, laptops), RTT and RSS measures are leveraged as the input features. The performance of our framework was verified through a real-world dataset of a campus floor.

We summarise our contributions as follows:

- We propose a novel framework to identify the LOS conditions of individual WiFi AP or all of them at the same time.
- Within our framework, we include a novel feature selection algorithm using only WiFi RTT and RSS signal as inputs.
- We shed light on the empirical performance of the proposed framework, using a complex dataset in a real-world indoor environment.

The rest of the paper is organised as follows. Section II introduces the related work in WiFi LOS identification. Section III provides a detailed description of the framework architecture. Then the data preprocessing and the proposed feature selection method are investigated in Section IV. The experimental setup and empirical performances are presented in Section V. Finally, Section VI concludes the paper.

## II. Related work

One of the most popular LOS detection solutions was done by leveraging the Channel State Information (CSI) [5], which describes the WiFi channel properties between the APs and the receiver in a communication link [3], [6]. Extracted features included phase information, amplitude information, Time-of-flight (ToF), and Time Difference of Arrival (TDoA) [7]. Power-delay profile and power-angle spectrum (PAS) were used as features in the systems proposed by [8]. PhaseU proposed by [9] studied the phase information of each sub-carrier extracted from Channel Impulse Response (CIR). The statistics of CIR signals were used as the input to the identification system in [10]. However, CSI could only be implemented with a modified WiFi driver of the Intel 5300 NIC on a PC, making it challenging to be used on any smartphone-based systems.

Due to the limitation of obtaining CSI information, some researchers used WiFi RSS and RTT for LOS identifications [11]–[13]. For example, RSS measure was used as input to a Gaussian model [14]. The system proposed by [15] leveraged RTT measure with pedestrian dead reckoning to make identifications. In [16], the standard deviations of both RSS and RTT were used for LOS detection. Furthermore,

in [17], the statistics of mean and kurtosis were considered when making identifications based on RTT and RSS measures. Skewness, hyper-skewness and peak probability were added in an RSS-based system for channel LOS identification [18].

## III. System architecture and problem formulation

This section provides a comprehensive introduction to the architecture of our proposed framework, including our feature selection algorithm process.

### A. System architecture

Our proposed framework consists of 4 steps (see Fig. 1).

- Step 1: A feature preprocessing algorithm is applied to the raw WiFi RTT and RSS training data to generate different sets of filtered features.
- Step 2: Based on the performance evaluation, these sets of features are given the initial weights, which are then fed into a feature selection algorithm (i.e., the feature selector).
- Step 3: A multi-scale selection (MSS) method is leveraged to reduce the weight of less important features. The new weights will be used as inputs to the feature selector.
- Step 4: When the users present their test samples, our framework will preprocess them so that only features selected by the feature selector remain. Finally, the LOS conditions of all APs are identified by the Random Forest Classifier (RFC).

### B. Problem formulation

To perform LOS identification, a set of raw WiFi data is collected at each reference point (RP) $R_j (j = 1, 2, \ldots, J)$ in the testbed, where $J$ is the total number of RPs. The data at the point $R_j$ contain $K$ consecutive scans of WiFi RTT measure $X_{RTTj} = \{x^{(1)}_{RTTjk}, x^{(2)}_{RTTjk}, \ldots, x^{(T)}_{RTTjk}\}^K_{k=1}$ and WiFi RSS measure $X_{RSSj} = \{x^{(1)}_{RSSjk}, x^{(2)}_{RSSjk}, \ldots, x^{(T)}_{RSSjk}\}^K_{k=1}$ from $T$ number of APs.

The ground-truth label to indicate the LOS condition of the APs at the reference point $R_j$ is defined as a vector $Y_j = [y^{(1)}_{jk}, y^{(2)}_{jk}, \ldots, y^{(T)}_{jk}]^K_{k=1}$, as follows:

$$y^{(t)}_{jk} = \begin{cases} 1 & \text{if the } t^{th} \text{ AP is LOS at } R_j \\ 0 & \text{if the } t^{th} \text{ AP is NLOS at } R_j \end{cases} \tag{1}$$

where $t = 1, 2, \ldots, T$. The raw training data are defined as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{Y} = \{Y_j\}^J_{j=1}$ and $\mathcal{X} = \{X_{RTTj}, X_{RSSj}\}^J_{j=1}$.

When new sample $\mathcal{X}_{Test}$ is collected at $R_{Test}$ by the user, it is preprocessed by the feature selection algorithm. Then, the LOS condition $y^{(t)}_{Test}$ of the $t^{th}$ AP is detected by RFC. The output $\mathcal{Y}_{Test}$ of RFC is defined as:

$$\mathcal{Y}_{Test} = [y^{(1)}_{Test}, y^{(2)}_{Test}, \ldots, y^{(T)}_{Test}] \tag{2}$$

## IV. Feature preprocessing and feature selection algorithms

This section explains the steps of our proposed framework including feature preprocessing, initial weights assignment, feature selector and testing data validation.

### A. Feature preprocessing

The feature preprocessing extracts statistics from the input WiFi measures and filters out those with low importance or weak correlation to the label using several selection models.

*1) Statistical feature extraction:* A feature extraction method is used to generate statistical features (i.e., mean ($\mu$), median ($Med$), standard deviation ($\sigma$), Kurtosis ($\mathcal{K}$) and Skewness ($\mathcal{S}$)) from the WiFi input data, as follows:

$$\mu = \frac{1}{K}\sum_{k=1}^{K} x_k \tag{3}$$

$$\sigma = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu)^2} \tag{4}$$

$$\mathcal{K} = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu)^4}{\sigma^4} \tag{5}$$

$$\mathcal{S} = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu)^3}{\sigma^3} \tag{6}$$

where $x_k$ indicates either RTT or RSS data collected at a specific reference point.

These statistics were reported as the most effective features for LOS detection [3]. After the process, the features in the original data $\mathcal{X}$ are replaced by a new feature vector $\mathbb{X} = \{\mu_{RTT}, \mu_{RSS}, Med_{RTT}, Med_{RSS}, \sigma_{RTT}, \sigma_{RSS}, \mathcal{K}_{RTT}, \mathcal{K}_{RSS}, \mathcal{S}_{RTT}, \mathcal{S}_{RSS}\}$.

*2) Importance filter:* After feature preprocessing, an importance filter is adopted. In this filter, five feature selection models (Chi-squared, Recursive Feature Elimination, Mean Decrease in Impurity, Permutation Importance, and Hierarchical Clustering) are used to remove the features in $\mathbb{X}$ with low importance or weak correlations to the label $\mathcal{Y}$ and generate five corresponding feature sets for the next step. These models rank the features based on their importance measures, and select the top $N_{initial}$ features defined by the user. A brief introduction of each feature selection model is described below.

**Chi-squared:** The Chi-squared model calculates the score $\chi^2_{score}$ of each feature $\mathtt{x}_n$ $(n = 1, 2, \ldots, N)$, where $N$ is the total number of statistical features in $\mathbb{X}$, to evaluate its correlation to the identification result $\mathcal{Y}$, as follows:

$$\chi^2_{score} = \sum \frac{(observed - expected)^2}{expected} \tag{7}$$

where $observed$ represents the observation in LOS identification based on existing $\mathcal{Y}$ and features in $\mathcal{X}$, $expected$ represents the expected observation output when $\mathcal{X}$ and $\mathcal{Y}$ are completely independent. The higher the $\chi^2_{score}$ is, the more relevant the feature is to the result.

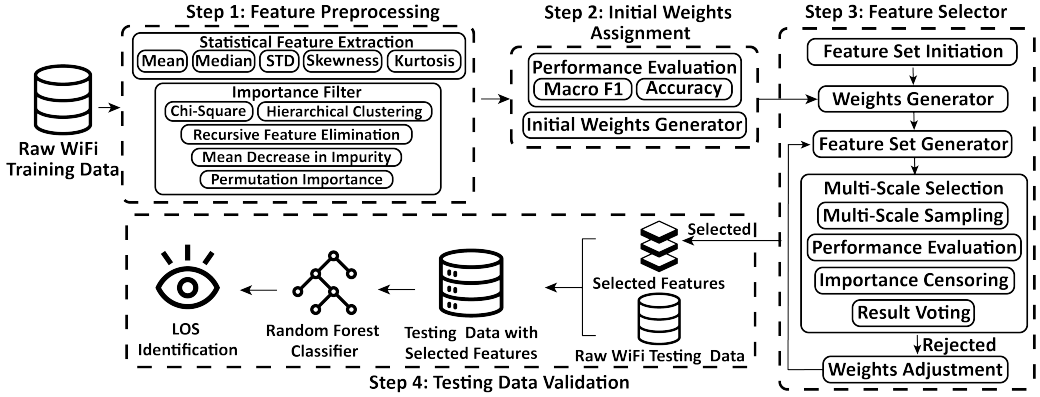


Fig. 1. The architecture of our proposed framework.

**Recursive Feature Elimination (RFE):** The RFE model recursively removes the least important feature until a feature set of a predefined size is found. At its core, RFE employs Support Vector Machine (SVM) to assign the weights. During the training phase, when a feature is removed, the cost function $J_{Cost}$ changes accordingly. Therefore, the importances of the features are assigned according to the changes of the cost function $\Delta J_{Cost}(n)$ indicating its importance to LOS identification as follows [19]:

$$\Delta J_{Cost}(n) = \frac{1}{2}\frac{\partial^2 J_{Cost}}{\partial w_n^2}(\Delta w_n^2) \tag{8}$$

where $\Delta w_n^2$ is the change in the weight of $\mathtt{x}_n$.

**Mean Decrease in Impurity (MDI):** One of the most popular methods to calculate the feature importance is via Gini impurity [20], which represents the probability of a false classification for a randomly chosen variable, as follows:

$$Gini = 1 - \sum_{l=1}^{L}(p_l)^2 \tag{9}$$

$$\Delta Gini(t) = Gini(t) - \frac{\mathcal{X}_{tL}}{\mathcal{X}_t}Gini(t_L) - \frac{\mathcal{X}_{tR}}{\mathcal{X}_t}Gini(t_R) \tag{10}$$

where $L$ is the number of classes in the label $\mathcal{Y}$, $p_l$ is the probability of the data to be identified as class $l \in \mathcal{Y}$, $t$ is a specific node in Random Forest, $t_L$ and $t_R$ are the child nodes of $t$, $\mathcal{X}_t$ is the input to the $t$, $\mathcal{X}_{tL}$ and $\mathcal{X}_{tR}$ are data divided into $t_L$ and $t_R$ respectively. $I_{MDI}$ is the weighted average of all $\Delta Gini(t)$ where the feature is used by $t$ [21].

**Permutation Importance (PI):** PI calculates the mean decrease in accuracy of a feature to investigate its importance to a classifier. First, the accuracy $Acc$ of the classifier in the original dataset is computed. During the iteration $r$ in $1, 2, \ldots, R$, feature $\mathtt{x}_n$ is shuffled. Then, the performance of the classifier in this shuffled dataset is evaluated via the mean accuracy decrease (MAD). The importance $I_{PIn}$ of $\mathtt{x}_n$ is defined as:

$$I_{PIn} = Acc - MAD \tag{11}$$

$$MAD = \frac{1}{R}\sum_{r=1}^{R} Acc_{rn} \tag{12}$$

By randomly shuffling each feature, PI evaluates the relevance between the feature and the output, which is adequate for selecting non-linear features [22]. The MAD of each feature on our dataset is illustrated in Fig. 2.

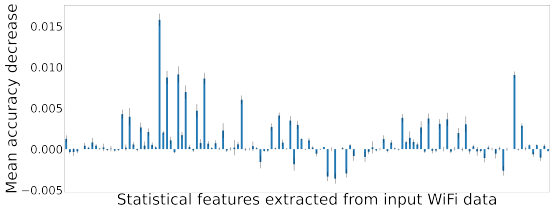


Fig. 2. The mean decrease in accuracy from Permutation Importance. The X-axis indicates different statistical features listed as in their original order from left to right. Negative permutation importance indicates that the feature is less necessary.

**Hierarchical Clustering (HC):** When the feature is correlated to others, its permutation importance is also likely to be smaller. To address this challenge, HC is introduced to investigate correlated features. This algorithm groups similar features to form clusters. The result is a set of clusters that are different from each other, but the features in the cluster are similar. HC treats each input feature as a separate cluster, then follows two steps: identifying the two most similar clusters and merging them together as a new cluster. The process is repeated until all clusters are unchanged.

### *B. Initial weights assignment*

Once the importance filter has been applied, each of the five selection models generates its feature sets $\hat{\mathbb{X}}_m$ based on the statistical features $\mathbb{X}$, where $m$ indicates different feature selection models. Next, RFC is used to identify the LOS conditions. Finally, the initial weights will be assigned to each feature in $\mathbb{X}$.

The results from RFC is evaluated by macro F1 score and accuracy. Macro F1 score is the average F1 score of all classes. Then, an evaluation vector $E_m = \{F1_m, Accuracy_m\}$ for each feature set $\hat{\mathbb{X}}_m$ is generated. The reason for using macro F1 score is that we value all APs equally. The predictions made by RFC are the final results from cross-validation.

An initial weight $w_m$ is given to each filtered feature set by the weights generator as $w_m = E_m \Big/ \sum_{m=1}^{5} E_m$.

When $\hat{\mathbb{X}}_m$ is given the weight $w_m$, every included feature $\hat{\mathtt{x}}_{nm}$ in $\hat{\mathbb{X}}_m$ gets the same weight. In addition, the removed features of this set have the weight of 0. If one feature is selected by more than one filtered feature set, then the weight $W_n$ of this specific feature $\mathtt{x}_n$ is defined as:

$$W_n = \sum_{m=1}^{5} w_{nm} \begin{cases} w_{nm} = 0 & \text{if } \mathtt{x}_{nm} \notin \hat{\mathbb{X}}_m \\ w_{nm} = w_m & \text{if } \mathtt{x}_{nm} \in \hat{\mathbb{X}}_m \end{cases} \tag{13}$$

The weight vector of $\mathbb{X}$ is defined as $\mathbb{W} = \{W_n\}_{n=1}^{N}$. A initial set of selected features $\hat{\mathbb{X}}$ is generated based on $\mathbb{W}$.

**Algorithm 1** Feature Selector with Multi-Scale Selection

**Input:** $\mathcal{X}^{(s)}, \mathcal{Y}^{(s)}$: input data and label of $s$ sample size, $\mathbb{X}, \hat{\mathbb{X}}, \mathbb{W}$, $F1, ACC$: macro F1 score and accuracy, $N_{min}$: minimum number of features $SampleSizes$: a set of different sample sizes
**Output:** $\hat{\mathbb{X}}^{\star}$: final set of selected features

$M \leftarrow |FeatureSelectionModels|$
$\mathbb{X}_{new} \leftarrow \hat{\mathbb{X}}$
$\mathbb{W}_{new} \leftarrow \mathbb{W}$
**while** $\mathbb{X}_{new} \neq \mathbb{X}_{old}$ or $|\mathbb{X}_{new}| \leq N_{min}$ **do**
  $\mathbb{X}_{old} \leftarrow \mathbb{X}_{new}$
  $\mathbb{W}_{old} \leftarrow \mathbb{W}_{new}$
  **for** $m$ = 1, 2, ..., $M$ **do**
    $model \leftarrow m^{th}$ model in $Models$
    **for** $s$ in $SampleSizes$ **do**
      $\mathbb{X}_{old}^{(s)} \leftarrow GenerateFeature(\mathcal{X}^{(s)}, \mathbb{X}_{old})$
      Predict $\leftarrow RFC(\mathbb{X}_{old}^{(s)}, \mathcal{Y}^{(s)})$
      $v_m^{(s)} \leftarrow \{F1$,$Acc$(Predict, $\mathcal{Y}^{(s)}$), $model(\mathbb{X}_{old}^{(s)}, \mathcal{Y}^{(s)})\}$
        ▷ the output of $model$ is the feature importances
    **end for**
  **end for**
  $V \leftarrow ResultVoting(\sum v_m^{(s)})$
  $\mathbb{W}_{new} \leftarrow WeightsAdjust(V, \mathbb{W}_{old})$
  $\mathbb{X}_{new} \leftarrow GenerateFeatureSet(\hat{\mathbb{X}}, \mathbb{W}_{new})$
**end while**
$\hat{\mathbb{X}}^{\star} \leftarrow \mathbb{X}_{new}$
**return** $\hat{\mathbb{X}}^{\star}$

### C. Feature selector and testing data validation

The feature selector is used to decide the final set of features. We select an initial set of features using the above weight vector $\mathbb{W}$ and validate it with the Multi-Scale Selection.

*1) Multi-Scale Selection (MSS):* MSS process aims to validate and update the initial set of selected features by evaluating them on multi-scale sampled datasets as shown in Algorithm 1.

**Multi-Scale sampling:** It is used to extract datasets with different sampling sizes from the original dataset. As introduced in Section V, the maximum sampling size is 120 scans per reference point. Different sampling sizes (i.e., 5, 10, 15, 30, 60, 120) were also tested (see Fig. 4).

**Performance evaluation:** As explained in Section IV-B, we use macro F1 score and accuracy to validate the features using RFC. The overall LOS detection performance on multi-scale datasets represents the features' importance.

**Importance censoring:** Since datasets of different scales contain information from different perspectives, this process re-visits the features' importance in the multi-scale datasets by adopting the importance filter.

**Result voting:** This is to censor each feature based on the RFC performances and its importance assigned by the feature selection models. Only features that perform well in all multi-scale datasets are seen as informative and indicative to LOS identification and given increase in their weights.

*2) Final feature set and testing data validation:* Rejected features containing fewer hidden patterns have their weights reduced through weight adjustment. After updating $\mathbb{W}$, the weight generator and feature set generator re-select a new set of features for the MSS. The above process repeats until a final set of features $\hat{\mathbb{X}}^{\star}$ is decided where all selected features are included.

When the user reports test samples $\mathcal{X}_{Test}$, statistical features $\mathbb{X}_{Test}$ are extracted. Then a set of the most meaningful features $\hat{\mathbb{X}}_{Test}^{\star}$ are selected based on the final set of features $\hat{\mathbb{X}}^{\star}$ generated by the feature selector. Finally, RFC makes LOS identification.

## V. Experimental setup and empirical results

This section details the test bed and evaluates the proposed framework.

### A. Test bed

We evaluated the proposed framework's performance on an entire campus floor with offices and long corridors for better generalisation of the empirical results (see Fig. 3). The area was of more than 92 × 15 m$^2$ and was divided into 642 square-size grids, to be used as reference points (RPs). Each grid unit-size was 0.6 × 0.6 m$^2$. 13 RTT-enabled Google APs were set up in the same locations as the university APs'.

The WiFi RTT and RSS data were collected with an LG G8X ThinQ smartphone. The ground-truth positioning labels were manually collected and verified by two human surveyors using measuring tapes and ground markers. The label of each data sample included the coordinates of the RP and the LOS APs. The dataset is publicly available[1] (see Table I for a summary). Note that the training RPs and the testing RPs do not overlap. The default sample size of the dataset was 120 scans per RP (i.e. about 40 seconds).

### B. The importance of feature selection

To investigate the impact of different features, we performed LOS identification of individual AP and all APs simultaneously using different sets of features. For individual AP LOS

[1]https://github.com/Fx386483710/WiFi-RTT-RSS-dataset

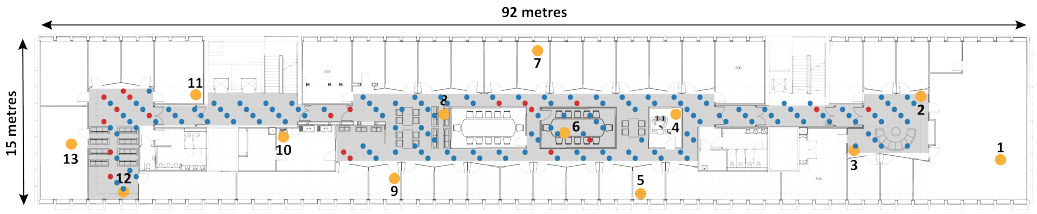


Fig. 3. The RPs where the system makes misidentifications. The blue dots indicate the RPs of testing data while the red ones indicate the misidentifications. The orange dots show the locations of the RTT-enabled APs. All measurements were taken in the grey area. Most misclassifications took place in areas surrounded by complicated interior changes.

TABLE I
SUMMARY OF THE DATASET.

| Data features | Details |
|---|---|
| Test bed area | 92 × 15 $m^2$ |
| Grid size | 0.6 × 0.6 $m^2$ |
| Number of reference points | 642 |
| Samples per reference point | 120 |
| Total samples | 77,040 |
| Training samples | 57,960 |
| Testing samples | 19,080 |
| Signal measure | RTT, RSS |
| Collection period | 3 days |

detection, AP6, AP8 and AP12 (as shown in Fig. 3), which have the most LOS paths to the RPs, were chosen.

As shown in Table II, statistical features extracted from APs were insufficient in predicting AP LOS. Because of the complex indoor environment, features from the APs often do not contain enough information. Most importantly, using all APs' features is likely to have a negative effect on the prediction. Therefore, our proposal of selecting only informative features is of great significance for LOS identification.

TABLE II
THE MACRO F1 SCORE PERFORMANCE OF ALL APS AND INDIVIDUAL AP LOS IDENTIFICATION USING DIFFERENT SETS OF FEATURES* .

| | AP6 | AP8 | AP12 | All APs |
|---|---|---|---|---|
| Features of this AP | 0.845 | 0.89 | 0.609 | N/A |
| All features | 0.609 | **0.933** | 0.851 | 0.689 |
| Features selected by MSS | **0.897** | **0.933** | **0.865** | **0.780** |

* Features included are all statistical features introduced in Section IV-A1.

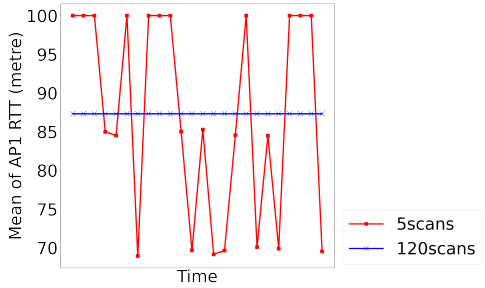


Fig. 4. The mean value of AP1's RTT measures with different sampling sizes at a RP. Value of 100 indicates that there is no WiFi signal at this RP. Datasets of different sample sizes contain signal patterns of different time period.

### C. The impact of sampling size

To assess our MSS algorithm, multiple sample sizes (i.e., 5, 10, 15, 30, 60, 120) were used to generate the multi-scale datasets. As such, the hidden information in the features were evaluated from both macroscopical and microscopical perspectives. In addition, small-size test data help us understand the system performance when the user moves quickly (with limited number of real-time samples).

The hidden information within the features vary on different time scales. As shown in Fig. 4, when using a 120-scan dataset, every RP only had one data sample. During the features extraction phase, sudden changes of WiFi measures were masked out when calculating the statistics of all 120 scans. On the contrary, if a 5-scan dataset was utilised, the system would only focus on the signal patterns for a short period (see Fig. 5).

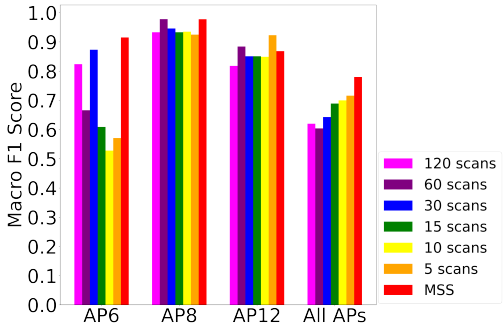


Fig. 5. The performance of all APs and individual AP LOS identification using datasets of different sampling sizes, compared to MSS. Statistical features of all APs were used except for MSS where only selected features were used. Features selected by MSS are more informative than features from single sample size dataset.

### D. The performance of the proposed framework

We evaluate the performance of the proposed framework in identifying the LOS conditions of all APs simultaneously. For comparison, the features chosen by several state-of-the-art models were used. Furthermore, we compare our work with the selected features from previous works $C_7^{SEL}$ [17] and SVM-FTM ($S$-$F$) [23]. $C_7^{SEL}$ leveraged mean values and Kurtosis of RTT, and mean RSS. $S$-$F$ utilised mean, standard deviation, Skewness and Kurtosis of both RTT and RSS (see Table III). We employed Macro F1 score as the primary

evaluation metric, because we value all APs and their LOS conditions equally in this large indoor environment. Whereas, weighted F1 score was used to mitigate the effect of class imbalance and to investigate the performance in individual AP identification. We observed that the proposed framework had an improvement of up to 15.3% in macro F1 score and up to 28.8% in weighted F1 score. Hence, our proposed framework was robust in both large and small scale LOS identifications.

TABLE III
THE PERFORMANCE OF USING DIFFERENT SETS OF FEATURES IN ALL APS LOS IDENTIFICATION.

| | **Weighted F1** | **Macro F1** | **Accuracy** (%) |
|---|---|---|---|
| Chi | 0.88 | 0.70 | 89.89 |
| RFE | 0.72 | 0.68 | 91.36 |
| MDI | 0.88 | 0.68 | 91.62 |
| PI | 0.90 | 0.68 | 91.62 |
| HC | 0.88 | 0.71 | 91.21 |
| Proposed method | **0.93** | **0.78** | **93.55** |
| $C_7^{SEL}$ | 0.89 | 0.69 | 91.92 |
| $S$-$F$ | 0.88 | 0.68 | 89.38 |

The RPs where misclassification happened are demonstrated in Fig. 3. Most false identifications took place in the cornered areas or those surrounded by complex interiors. This was consistent with the intuition that the more steady the signals were, the more accurate the identification would be.

Next, we evaluated the performance for individual AP LOS identification (see Table IV). We observed that our system performed reliably in identifying LOS of AP with few LOS paths. Our proposed framework only used 34 features in total with testing samples recorded within 3 seconds.

TABLE IV
THE PERFORMANCE OF INDIVIDUAL AP LOS IDENTIFICATION.

| | **AP2** | **AP3** | **AP6** | **AP7** | **AP8** | **AP10** | **AP12** |
|---|---|---|---|---|---|---|---|
| Accuracy(%) | 99.34 | 99.54 | 98.98 | 98.98 | 99.39 | 98.78 | 98.12 |
| Macro F1 | 0.97 | 0.97 | 0.92 | 0.62 | 0.98 | 0.50 | 0.87 |

* APs which do not have LOS path to any RP are not included.
** Only 8 and 4 RPs have LOS path to AP7 and AP10, respectively.

## VI. CONCLUSIONS

In this paper, we have proposed a novel feature selection framework to address the research challenges and limitations of Access Points LOS identification for WiFi-based indoor positioning. Our framework was proposed with its unique feature selection methods using popular WiFi RTT and RSS signal as inputs. We evaluated our system on a real-world dataset, which is also publicly availabe. The results demonstrated an accuracy of 93% and 98% when identifying all APs and individual AP, respectively, using only 3 seconds of the WiFi data.

## REFERENCES


[1] K. A. Nguyen, Z. Luo, G. Li, and C. Watkins, "A review of smartphones-based indoor positioning: Challenges and applications," *IET Cyber-Systems and Robotics*, vol. 3, no. 1, pp. 1–30, 2021.

[2] K. A. Nguyen and Z. Luo, "On assessing the positioning accuracy of google tango in challenging indoor environments," in *2017 international conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2017, pp. 1–8.

[3] C. Huang, R. He, B. Ai, A. F. Molisch, B. K. Lau, K. Haneda, B. Liu, C.-X. Wang, M. Yang, C. Oestges *et al.*, "Artificial intelligence enabled radio propagation for communications—part ii: Scenario identification and channel modeling," *IEEE Transactions on Antennas and Propagation*, 2022.

[4] X. Feng, K. A. Nguyen, and Z. Luo, "Wifi access points line-of-sight detection for indoor positioning using the signal round trip time," *Remote Sensing*, vol. 14, no. 23, p. 6052, 2022.

[5] T. Chang, S. Jiang, Y. Sun, A. Jia, and W. Wang, "Multi-bandwidth nlos identification based on deep learning method," in *2021 15th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2021, pp. 1–5.

[6] X. Feng, K. A. Nguyen, and Z. Luo, "A survey of deep learning approaches for wifi-based indoor positioning," *Journal of Information and Telecommunication*, pp. 1–54, 2021.

[7] Z. Hajiakhondi-Meybodi, A. Mohammadi, M. Hou, and K. N. Plataniotis, "Dqlel: Deep q-learning for energy-optimized los/nlos uwb node selection," *IEEE Transactions on Signal Processing*, vol. 70, pp. 2532–2547, 2022.

[8] K. Jiokeng, G. Jakllari, A. Tchana, and A.-L. Beylot, "When ftm discovered music: Accurate wifi-based ranging in the presence of multipath," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1857–1866.

[9] C. Wu, Z. Yang, Z. Zhou, K. Qian, Y. Liu, and M. Liu, "Phaseu: Real-time los identification with wifi," in *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 2015, pp. 2038–2046.

[10] M. Ramadan, V. Sark, J. Gutierrez, and E. Grass, "Nlos identification for indoor localization using random forest algorithm," in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*. VDE, 2018, pp. 1–5.

[11] X. Feng, K. A. Nguyen, and Z. Luo, "An analysis of the properties and the performance of wifi rtt for indoor positioning in non-line-of-sight environments," in *17th International Conference on Location Based Services*, 2022.

[12] D. Schepers, M. Singh, and A. Ranganathan, "Here, there, and everywhere: security analysis of wi-fi fine timing measurement," in *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 78–89.

[13] K. A. Nguyen, "A performance guaranteed indoor positioning system using conformal prediction and the wifi signal strength," *Journal of Information and Telecommunication*, vol. 1, no. 1, pp. 41–65, 2017.

[14] M. Si, Y. Wang, S. Xu, M. Sun, and H. Cao, "A wi-fi ftm-based indoor positioning method with los/nlos identification," *Applied Sciences*, vol. 10, no. 3, p. 956, 2020.

[15] S. Xu, R. Chen, Y. Yu, G. Guo, and L. Huang, "Locating smartphones indoors using built-in sensors and wi-fi ranging with an enhanced particle filter," *IEEE Access*, vol. 7, pp. 95 140–95 153, 2019.

[16] M. Sun, Y. Wang, S. Xu, H. Qi, and X. Hu, "Indoor positioning tightly coupled wi-fi ftm ranging and pdr based on the extended kalman filter for smartphones," *Ieee Access*, vol. 8, pp. 49 671–49 684, 2020.

[17] Y. Dong, T. Arslan, and Y. Yang, "Real-time nlos/los identification for smartphone-based indoor positioning systems using wifi rtt and rss," *IEEE Sensors Journal*, 2021.

[18] F. Carpi, L. Davoli, M. Martalò, A. Cilfone, Y. Yu, Y. Wang, and G. Ferrari, "Rssi-based methods for los/nlos channel identification in indoor scenarios," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2019, pp. 171–175.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[20] B. Leo, "Manual on setting up, using, and understanding random forests v3.1," *Statistics Department University of California Berkeley*, vol. 1, p. 58, 2002.

[21] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Advances in neural information processing systems*, vol. 26, 2013.

[22] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[23] K. Han, S. M. Yu, and S.-L. Kim, "Smartphone-based indoor localization using wi-fi fine timing measurement," in *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2019, pp. 1–5.