Localization in cellular network using machine learning and fingerprint methods

Abstract: Location-based services (LBS) are widely utilized in various daily life applications. Research into localizing user or mobile devices has spurred the development of numerous techniques aimed at ensuring precise and efficient positioning capabilities. Although the Global Positioning System (GPS) is widely relied upon for location information, its integration into mobile devices incurs additional costs and drains battery life. Moreover, weak or jammed GPS signals make positioning unfeasible in certain scenarios. In contrast, cellular network localization presents a viable alternative, bypassing such challenges. Cellular networks, being the dominant mobile technology, facilitate communication among user devices and network operators while also enabling device localization using network information. In this study, we propose a methodology for collecting and processing cell information datasets, focusing on Received Signal Strength (RSS) via mobile devices, followed by an exploration of various localization methods including Centroid, Weighted Centroid, Linear Regression, Support Vector Regression, Multilayer Perceptron, and Fingerprinting. Additionally, we introduce an innovative approach called CNN-CellImage by using a CNN model with image data that depict the geographical relationship and RSS of cells. The experimental findings show the effectiveness of our proposed CNN-CellImage method for localization compared to the other techniques examined.

Keywords: Location-based services, Localization, Cellular Network, Signal Fingerprint, Received Signal Strength.

Reference to this paper should be made as follows: ...

Biographical notes:

1 Introduction

Location data is widely used in various applications of daily living, with the number of Location-Based Services (LBS) increasing steadily across different devices and applications. There is considerable research focus on advancing mobile device positioning technologies to offer swift, precise, and highly efficient location capabilities for users. While satellite-based systems such as the Global Positioning System (GPS) provide excellent accuracy and are widely utilized, their integration into mobile devices adds to costs and battery drain. Weak or disrupted GPS signals can pose challenges, rendering accurate positioning impossible in certain situations. In addition, GPS operates primarily in a one-way communication mode, which requires a supplementary communication infrastructure for data transfer when needed. In response to these challenges, localization using cellular networks is being explored due to the widespread adoption of the cellular network and its ability to facilitate communication between users and network operators, as well as peer-to-peer communication among users themselves.

Localization within the GSM network involves determining the physical location of a mobile device within the network. The widespread adoption and growth of related technologies within the GSM network enable users to access the necessary information for localization. For network providers, understanding the location of their users is crucial for delivering location-based services effectively. With seamless mobility and roaming capabilities between neighboring base stations, mobile devices can maintain connections and communicate across the network infrastructure. A straightforward method for positioning a device within the mobile network is by estimating its location based on the coordinates of the base station it is connected to, known as Cell Identification (Cell ID) Jose A. del Peral-Rosado et al. (2003). However, the accuracy of this method may not be sufficient for scenarios requiring precise positioning due to the wide effective range of base stations, which can span from hundreds of meters to several kilometers depending on the network infrastructure. Various methods have been explored that take advantage of information from the mobile network to improve positioning accuracy. The Received Signal Strength (RSS), a critical signal parameter within the network, helps to estimate the distance between the mobile device and the base stations Martinka, J. (2019). Additionally, mobile devices can collect information not only from the base station to which they are connected, but also from neighboring base stations, providing additional geographic data for localization methods.

In this study, we propose the methodology for gathering data from the cellular network and assess various localization techniques, ranging from basic approaches as the Cell ID method and centroid method to more intricate methods involving machine learning and radio fingerprinting. We also propose an original approach so-called CNN-CellImage, using CNN model

with image data which presents the geographical relationship and RSS of cells. The remainder of this paper is structured as follows: Section 2 provides a brief overview of localization techniques in cellular networks. The following section outlines our data collection process from the cellular network using a mobile application. Section 4 outlines the localization methods we explored and assessed. We present the results of our experiments in Section 5 and draw conclusions in Section 6.

2 Related Works

Localization methods within cellular networks rely primarily on the observation of various signal parameters. As outlined in Jose A. del Peral-Rosado et al. (2003), these methods encompass distinct fundamental positioning techniques, which can be categorized as proximity, trilateration, triangulation, and scene analysis techniques. The explanation of these fundamental positioning techniques is illustrated in figure 1

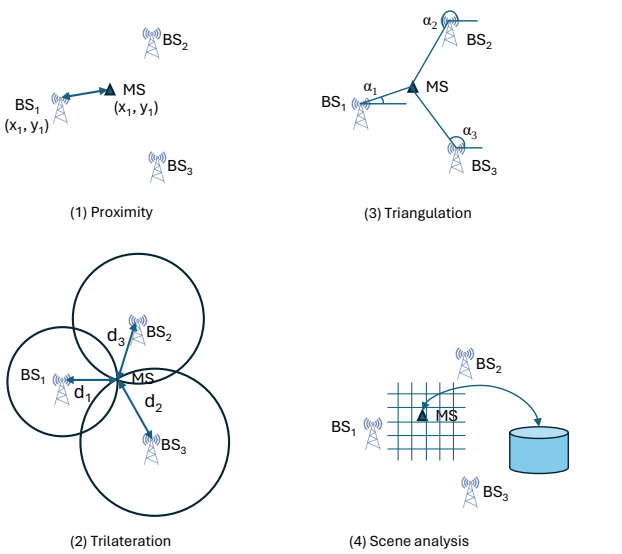


Figure 1: Fundamental positioning techniques using radio signals. Jose A. del Peral-Rosado et al. (2003)

Proximity techniques determine the position of the mobile device by associating it with the known location of the transmitting stations. For instance, in the Cell ID method Laoudias et al. (2018), the mobile device's position is inferred from the serving base station's position. Similarly, in the centroid method, the mobile device's position is estimated by calculating the centroid coordinate of a group of base stations, including the serving base station and neighboring stations. These methods are widely utilized in GSM networks.

Trilateration techniques ascertain the mobile device's position by calculating the intersection of geometric circles formed by distance measurements between the reference transmitting station and the mobile device Jose A. del Peral-Rosado et al. (2003). Various distance measurement techniques can be employed, such as Time of Arrival (ToA) Wang, X. et al. (2003), Time Difference of Arrival (TDoA) Bridal, P. et al. (2009) Deligiannis, N. et al. (2007), or Received Signal Strength (RSS) Brida, P. et al. (2006), Hata, M. et al. (1980), Laitinen, H. et al. (2000), Le, M. H. et al. (2022). The core principle involves estimating the distance based on parameters like RSS or ToA and subsequently employing trilateration to determine the mobile device's position.

Triangulation techniques estimate the position of the mobile device using the intersection of at least two known directions of incoming signals based on the direction or angle of arrival (AoA) of the received signals Jose A. del Peral-Rosado et al. (2003), Le, M. H. et al. (2022).

Scene analysis techniques, also referred to as fingerprinting or radio frequency pattern matching. The main idea of the fingerprint approach is to collect radio signal features (fingerprints) at every location (reference points) in the area of interest and then build a fingerprint database. The location of an user device is then estimated by mapping the fingerprints collected against the database, relying on identifying the best match for a specific signal measurement, such as Received Signal Strength. Each fingerprint is associated with a specific location Bartoletti, S. et al. (2018), Brída, P. et al. (2010), Jose A. del Peral-Rosado et al. (2003), Takenga, C. M. et al. (2006), Vin, I. et al. (2015).

In the fingerprinting method, a database called radio map is created at the offline phase, in which fingerprints are stored together with the corresponding location. To build this database, the mobile device makes signal measurements and records the data. Each signal measurement is bound to GPS coordinates where the measurements were taken called reference position. The fingerprint stores the data of the base stations visible at that reference position together with their signal strengths. Reference positions may be grouped to form small areas with an average signal strength for each base station. After forming the database, in the online phase, the location of the mobile device can be obtained by comparing a new signal measurement with the fingerprints in the database Bartoletti, S. et al. (2018). This comparison may be implemented in many ways, but the purpose is to find the best match to the measurement. The accuracy of the fingerprinting method depends on the number of fingerprints.

Another approach for localization in cellular network is using machine learning algorithms with collected data. Basic machine learning algorithms, such as support vector regression and multilayer perception, to evaluate localization results Calábek, A. (2020), Phruksahiran, N. (2023). In Abdallah, A. A. et al (2018) a machine learning method is developed that integrates a weighted neighbor K-nearest and a multilayer neural network for localization based on received signal strength (RSS) from cellular towers. This method only assumes knowledge of RSS fingerprints of the environment, and does not

require knowledge of the cellular base transceiver station (BTS) locations, nor uses any RSS mathematical model.

3 Data Collection

We developed a scanning application compatible with Android OS, designed to collect GSM network data using a mobile phone. The application performs measurements at one-second intervals, scanning and reporting a list of cell information observed by the mobile phone in its vicinity, including cell identification and various radio signal properties. We specifically utilize four properties to identify the cell: Mobile Country Code (MCC), Mobile Network Code (MNC), Local Area Code (LAC), and Cell Identifier (CID), along with the Received Signal Strength Indicator (RSSI) value, to estimate the device's location. With each measurement, the application also requests the GPS coordinates (latitude and longitude) of the device to serve as the reference or ground-truth location for that measurement. Data records are transmitted to a server to compile a database of measurements and are also stored in a CSV log file. Each measurement is recorded as a line in the log file, comprising the measurement index, device's reference location, measurement timestamp, the number of cells detected, followed by a list of cell identification information including LAC, CID, and RSSI for each cell in the list. In the context of collecting data from a specific network provider in Vietnam, the MCC and MNC are already known and do not need to be saved in the measurement data record. The format of a data record in the CSV log file follows this structure.

$$(id, lat, lon, timestamp, n, LAC_1, CID_1, RSSI_1, ..., LAC_n, CID_n, RSSI_n) \quad (1)$$

where:
id: index of measurement
lat: GPS latitude of the device that the measurement was taken
lon: GPS longitude of the device on which the measurement was taken
timestamp: time stamp of the measurement
n: number of cell that the device scans its surroundings for the measurement.

$LAC_i, CID_i, RSSI_i$ with i from 1 to n corresponding to LAC, CID, and RSSI of the ith cell in the list of cells that the device scanned. For identifying the cell, since the experiment was conducted for collect the data in Hanoi City of Vietnam with a particular network provider for example Viettel, the parameters MCC, MNC are known parameters for the experiment, therefore these parameters can be used by default and do not need to describe in data record of the measurements.

In order to gather the measurements, we employed various methods, including driving a car and motorbike through the streets of the Hanoi City area while running the scanning application on the device. Given the variability in the characteristics of signal propagation in different environments, our focus was primarily on the urban area of Hanoi City, characterized by high vehicular and pedestrian traffic and surrounded by buildings. Through this process, we accumulated a database of measurements within the urban area of Hanoi. Figure 2 shows the visualization of the reference location of some data samples collected in an urban area of the city of Hanoi.

By utilizing cells identification details comprising MCC, MNC, LAC, and CID, the location of the base station (BTS) can be retrieved by querying various services that offer cell location data, such as https://www.opencellid.org/, https://findcellid.com/, or Google's Geolocation API services. In this study, we developed a script to retrieve cell locations from the API service provided by findcellid.com. The latitude and longitude coordinates for each cell, along with its identifier, were then appended to the collected data.

Now the collected data adding with cell location as the following format:

$$(id, n, LAC_1, CID_1, Lat_1, Lon_1, RSSI_1, ..., LAC_n, CID_n, Lat_n, Lon_n, RSSI_n) => (lat, lon)_{ref} \quad (2)$$

4 Localization Methods

We compile a data set from the information collected and apply various localization methods, including the centroid method, machine learning techniques, and the fingerprinting method, to predict the location of the mobile device using this data set. The processed data consists of measurement data records detailed in Section 3. Each data record includes the GPS coordinates of the mobile device, serving as the reference or ground truth location for assessing the effectiveness of the localization methods. To gauge the accuracy of the predicted locations, we computed the distance error between the predicted location and the reference location of the mobile device for each data record. This distance is calculated using the Haversine distance formula between two points (lat_1, lon_1) and (lat_2, lon_2) as described in the equation 3.

$$\begin{aligned} d &= R.c \\ c &= 2 \times atan2(\sqrt{a}, \sqrt{1-a}) \\ a &= \sin^2(\frac{lat_2 - lat_1}{2}) + \cos(lat_1) \times \cos(lat_2) \\ &\quad \times \sin^2(\frac{lon_2 - lon_1}{2}) \end{aligned} \quad (3)$$

where R is the radius of the Earth in meters, the coordinates lat, lon are converted to radian.

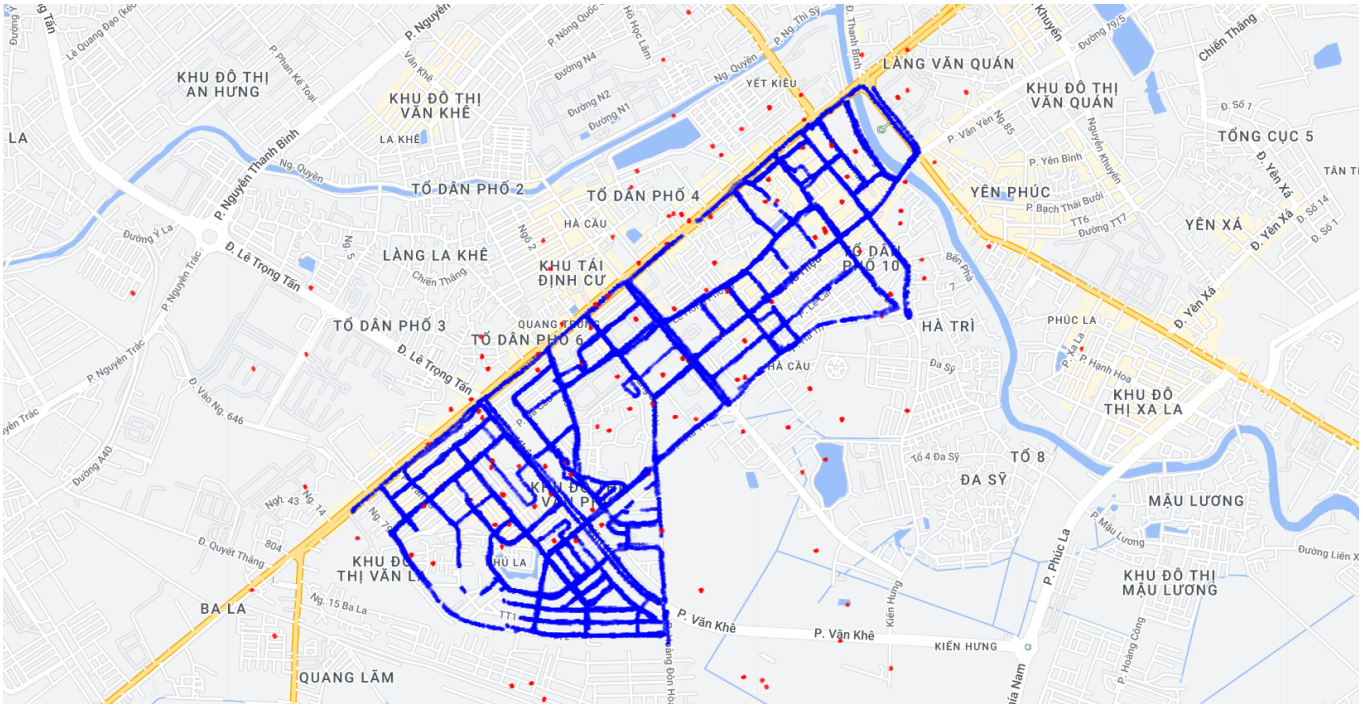


Figure 2: Visualization of some data samples collected in an urban area of Hanoi City.

4.1 Centroid and Weighted Centroid Methods

The most straightforward approach to localize a mobile device within the cellular network is relying on the base station it is currently connected to. This method, known as the Cell-ID method, establishes the device's location as identical to that of the serving cell it is connected to. By identifying the cell, one can retrieve the base station's location from a database. However, the accuracy of this method is heavily contingent upon the physical layout of the network infrastructure, as base stations can have an effective radial range spanning from 10 meters to 30 kilometers. Despite its limitations, this method may serve as a simple alternative when other techniques fail or are not applicable.

In the Cell-ID method, we rely solely on the registered cell to localize the mobile device. However, the mobile device also possesses information about the neighboring cells of the serving cell. An enhancement to this method is the centroid method, which considers not only the position of the serving cell but also that of the neighboring cells. A group of cells visible to the mobile device is grouped together, and the mobile device's position is estimated as the centroid of this cluster. The centroid of the cluster is determined by calculating the arithmetic mean of the coordinates of the cell locations.

$$Loc_{MS} \approx Loc_{Centroid} = \sum_{i=1}^{n} Loc_i \tag{4}$$

where
Loc_{MS} refers to the estimated location of a mobile station (or mobile device) ,
$Loc_{Centroid}$ is the estimated location of the centroid of the cell cluster visible to the mobile device, encompassing both the serving cell and neighboring cells,
Loc_i is the location (latitude, longitude) of base station of *ith* cell in the cluster.

In the centroid method, each cell within the cluster contributes equally to determining the estimated location of the mobile device. However, the weighted centroid method allows for varying contributions from each cell within the clusters by assigning a weight to each cell coordinate. Figure 3 shows the illustration of the weighted centroid method. These weights are determined on the basis of the signal strength measured by the RSSI (Received Signal Strength Indicator) value of each visible cell reported by the mobile device. When the signal strength of a base station is higher, the centroid tends to be closer to that base station. However, this assumption is not always accurate, as there are scenarios in which two base stations exhibit equal measured signal strength on the mobile device. In such cases, one base station may be farther away but transmitting a stronger signal, while the other may be closer but transmitting a weaker signal.

$$Loc_{MS} \approx Loc_{WeigthedCentroid} = \frac{\sum_{i=1}^{n} Loc_i \times w_i}{\sum_{i=1}^{n} w_i} \tag{5}$$

where
Loc_{MS} refers to the estimated location of a mobile station (or mobile device) ,
Loc_i is the location (latitude, longitude) of *ith* cell in the cluster visible by the mobile device including a serving cell and other neighboring cells,
w_i is the signal strength weight of ith cell.

The weight w_i of the signal strength of a cell is determined based on the RSSI value measured on the

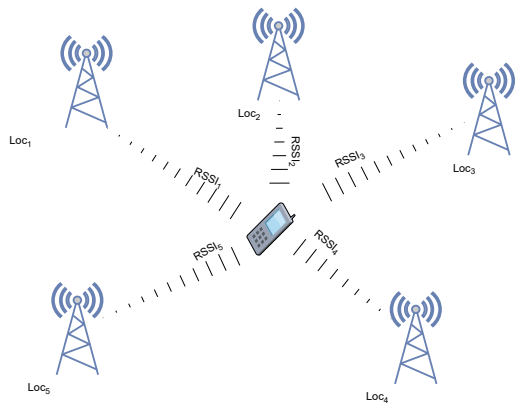


Figure 3: Weighted centroid method

mobile device for that cell. As the RSSI value is measured in decibels, a value closer to 0 indicates a stronger signal. To calculate the weight w_i, the RSSI value is first converted to an RSSI' value. Then, the weight w_i is computed as the ratio between the RSSI' value of the ith cell and the sum of the RSSI' values of all visible cells, as shown in Equation 6. In our experiment, the value α is assigned to 113 due to the RSSI value measured in the range of 0 to -113 dBm.

$$\begin{aligned} RSSI'_i &= \alpha - |RSSI_i| \\ w_i &= \frac{RSSI'_i}{\sum_{i=1}^{n} RSSI'_i} \end{aligned} \tag{6}$$

In practice, the serving cell may not necessarily have the strongest signal. A significant factor that affects the accuracy of the weighted centroid technique is the one-sided shadowing issue. This problem arises due to obstacles such as buildings, terrain, or people, causing the mobile device to receive weak signals from one or more cells that originate in approximately the same direction. Such occurrences can greatly influence the calculation of the centroid's coordinates.

4.2 Linear Regression and Support Vector Regression Methods

Since the number of cells visible by the mobile device in each data record may differ, the data records are aligned to ensure uniformity in the number of cells, facilitating consistency across records for machine learning models by the following procedure:

Procedure 1. Processing the dataset for Linear Regression and Support Vector Regression methods

1. Define k as the number of visible cells in each data record. In this study, we opt for k = 7 as it typically reflects the number of cells visible to the mobile device in most instances.
2. For each record in the data set:
   (a) If the record contains fewer than k cells, replicate the data from the serving cell $(LAC, CID, Lat, Lon, RSSI)$ to ensure the record has the requisite k cells.
   (b) If the record contains more than k cells, remove the data from cells $(LAC, CID, Lat, Lon, RSSI)$ with the lowest RSSI value until k cells remain.
   (c) Generate the vector $(Lat_1, Lon_1, RSSI_1, ..., Lat_k, Lon_k, RSSI_k)$ from the record. This vector is used as input sample for machine learning models Linear Regression and Support Vector Regression.

Following this processing method, every record in the dataset possesses an identical count of cells and is tagged with GPS coordinates serving as the ground truth. The dataset is divided into training and testing subsets in an 80:20 ratio, respectively. We trained Linear Regression and Support Vector Regression models using the training subset, subsequently assessing the accuracy of prediction outcomes on the testing subset. The performance of the models is gauged by the distance error between the predicted location and the reference location (or ground truth), calculated according to equation 3.

4.3 Multilayer Perceptron Method

We also utilize a neural network model known as multilayer perceptron (MLP) to predict the location of the mobile device using the data set. Prior to employing the MLP model, the dataset undergoes pre-processing to transform it according to the subsequent procedure:

Procedure 2. Processing of the data set for the multilayer perceptron method

1. Define k as the number of visible cells in each data record. For this study, we opt for k = 7 since, in most cases, there are 7 cells visible to the mobile device.
2. For each data record:
   (a) If the record contains fewer than k cells, replicate the data from the serving cell $(LAC, CID, Lat, Lon, RSSI)$ to ensure it has the required k cells.
   (b) If the record contains more than k cells, remove the data from cells $(LAC, CID, Lat, Lon, RSSI)$ with the smallest RSSI value until k cells remain.
   (c) Generate the vector $(Lat1, Lon1, RSSI_1, ..., Lat_k, Lon_k, RSSI_k)$ from the record.
3. For each vector:
   (a) Calculate the centroid coordinates $(Lat_{centroid}, Lon_{centroid})$ of the cells list in the vector.

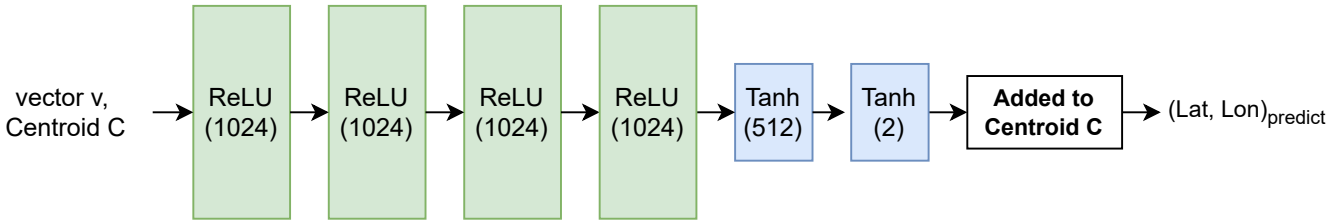


Figure 4: Multilayer Perceptron method.

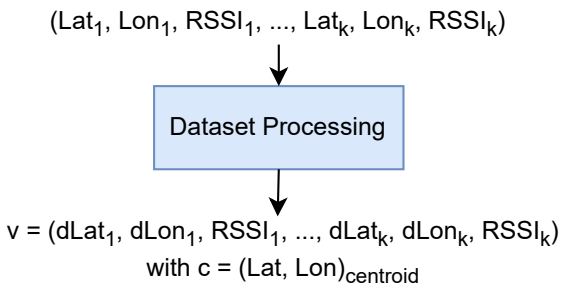


Figure 5: Processing of dataset for MLP method

(b) Calculate the latitude and longitude distance offsets between each cell location and the centroid location using the following formula:

$$\begin{aligned} dLat_i &= Lat_i - Lat_{centroid} \\ dLon_i &= Lon_i - Lon_{centroid} \end{aligned} \tag{7}$$

(c) Obtain a vector $\vec{v}$ for each data record along with the centroid point.

$$\begin{aligned} \vec{v} = (&dLat_1, dLon_1, RSSI_1, dLat_2, dLon_2, \\ &RSSI_2, ..., dLat_k, dLon_k, RSSI_k) \end{aligned} \tag{8}$$

Figure 5 illustrates the dataset processing procedure for the MLP method. Consequently, we acquire the transformed dataset $DS = \{\vec{v}\}$. Following this, the data set is split into training and testing subsets in an 80:20 ratio, respectively. The training subset is utilized to train the MLP model. The vector v serves as an input sample for the MLP neural network to predict the distance offset (dLat, dLon) from the centroid location. The mobile device's absolute location is determined by adding the predicted distance offset to the centroid coordinate for each sample. The architecture of the MLP neural network in our experiment is depicted in Figure 4. The accuracy of this MLP approach is assessed using the same distance error metric as in previous methods, computed using Equation 3.

4.4 Fingerprint Method

In this research, we also implement the fingerprint technique to predict the location of a mobile device. The fingerprint method process involves two phases: constructing the fingerprint database (offline phase) and localizing new records using the fingerprint database (online phase). During the construction of the fingerprint database phase, a user traverses an area employing a mobile device with a scanning application to gather information about cells with measurements taken at regular intervals. Each measurement logs the unique identifiers of the Base Transceiver Stations (BTS) along with their signal strengths, linked to the GPS coordinates of the mobile device as a reference position. This procedure results in the creation of a database, essentially a map featuring reference positions where measurements were captured, detailing visible cells (or corresponding BTS) at these positions alongside their signal strengths. These reference positions can be grouped to delineate small areas with averaged signal strengths for each BTS. Subsequently, each grouped measurement forms a data record comprising a list of cell identifiers and their signal strengths, attached to a reference position. These data records, termed fingerprints, encapsulate the radio signal characteristics of BTS at the locations where the measurements were taken. This process is illustrated in Figure 6. Once the fingerprint database is constructed, any new measurement can be compared with all fingerprints in the database to identify candidate fingerprints closely resembling the measurement. The outcome of this search yields one or more fingerprints whose BTS signal strengths closely match those of the given measurement. The accuracy of the fingerprint method depends on the number of fingerprints in the database, necessitating continuous scanning of the areas of interest.

During the localization phase, we introduced a method to identify candidate fingerprints within the database that closely resemble the fingerprint of a new measurement. Subsequently, we utilize these identified fingerprints to estimate the position of the mobile device.

Procedure 3. Predict the location using the fingerprint method

Input: A new input record and the fingerprints database

Output: The predicted location corresponding to the input record and the distance error between the predicted location and the reference location of the input record

1. Generate the characteristic vector v_{input} of the new input record by extracting the list of RSSI values $v_{input} = (RSSI_1, RSSI_2, \ldots RSSI_n)$ where $RSSI_i$ represents the signal strength indicator received corresponding to the $i-th$ cell in the input record measurement.

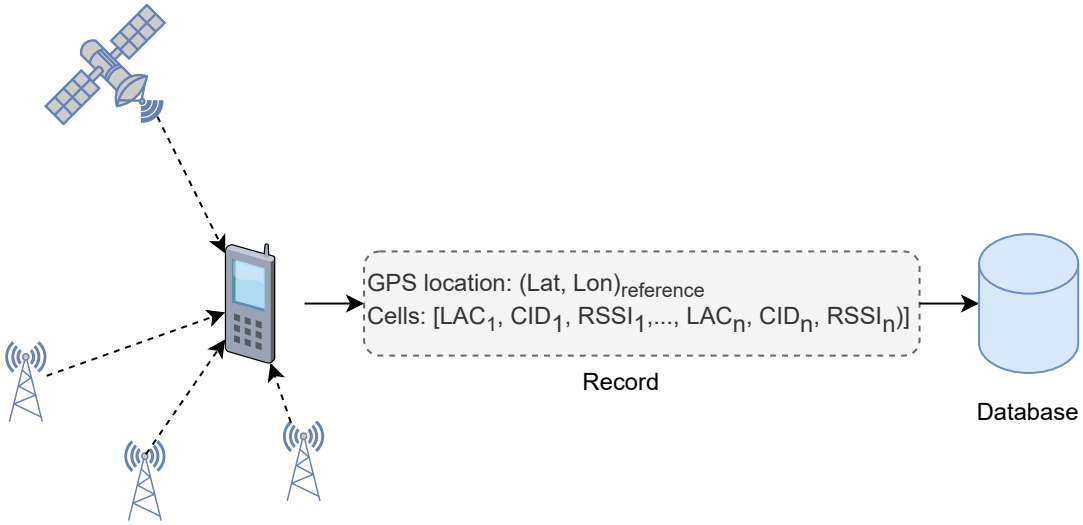


Figure 6: Constructing fingerprints database in offline phase

2. Query the fingerprint database to find the records (or fingerprints) containing at least one cell present in the input record. Obtain a set of records $S_1 = \{record_k\}$
3. For each $record_k$ in the set S_1:
   (a) Generate a candidate $candidates_k = \{RSSI_{ki}\}$ with $i = 1, ..., n$ where $RSSI_{ki}$ is equal to $RSSI_i$ of the ith cell in the $record_k$ if this ith cell appears in the input record respectively, or equal -200 (as a punishment) if this ith cell does not appear in the input record.
   (b) Add $candidates_k$ to the set S_2
4. Compute the similarity d_k of the input vector v_{input} with each vector $candidates_k$ in the set S_2 using the Euclidean distance
5. Sort the values of d_k, then select the feature vector $candidates_k$ corresponding to the smallest d_k value.
6. Predict the location of the input record based on the location of the record corresponding to the vector $candidates_k$.

Figure 7 depicts the procedure for estimating the location of input record using similarity matching in fingerprint method. In our experiment with this method, we partitioned the data set into a fingerprint database and a test set in an 80:20 ratio, respectively. Subsequently, we assessed the accuracy of the predicted location results on the test set using the distance error metric, as done in previous methods.

4.5 CNN-CellImage Method

The methodologies discussed in earlier sections, namely Linear Regression, Support Vector Regression, and Multilayer Perceptron, utilize feature vectors that contain information about cells, including their location (latitude, longitude) and RSSI values. However, these feature vectors lack spatial relationships between cells. Furthermore, to ensure uniformity in the size of the feature vectors (the number of cells in each record), the serving cell may be duplicated or redundant cells removed, potentially resulting in the loss of cell information or duplication of serving cells within records. To address this issue, we propose a method for location prediction by converting the cell information in each record into an image format that includes geographic location and signal strength RSSI of cells. Subsequently, we apply a Convolutional Neural Network (CNN) model for location prediction. The procedure for generating an image from the data record is outlined below.

Procedure 4. Generate image presenting geolocation and signal strength of cells for each record

Input: a data record

$$(LAC_1, CID_1, Lat_1, Lon_1, RSSI_1, ..., LAC_n, CID_n, Lat_n, Lon_n, RSSI_n) \quad (9)$$

attached to the reference location $(lat, lon)_{ref}$

Output: Image attached with the reference location

1. Compute the centroid coordinate of a cluster of cells for each record: $c = (Lat, Lon)_{centroid}$
2. For each cell c_i in a cluster of cells, computer the distance offset of latitude, longitude between the location of the cell c_i and the centroid c:

$$dLat_{c_i} = Lat_{c_i} - Lat_{centroid}$$
$$dLon_{c_i} = Lon_{c_i} - Lon_{centroid} \quad (10)$$

3. Find $dmax = max(|dLat_{c_i}|, |dLon_{c_i}|)$ with $i = 1, \ldots, k$

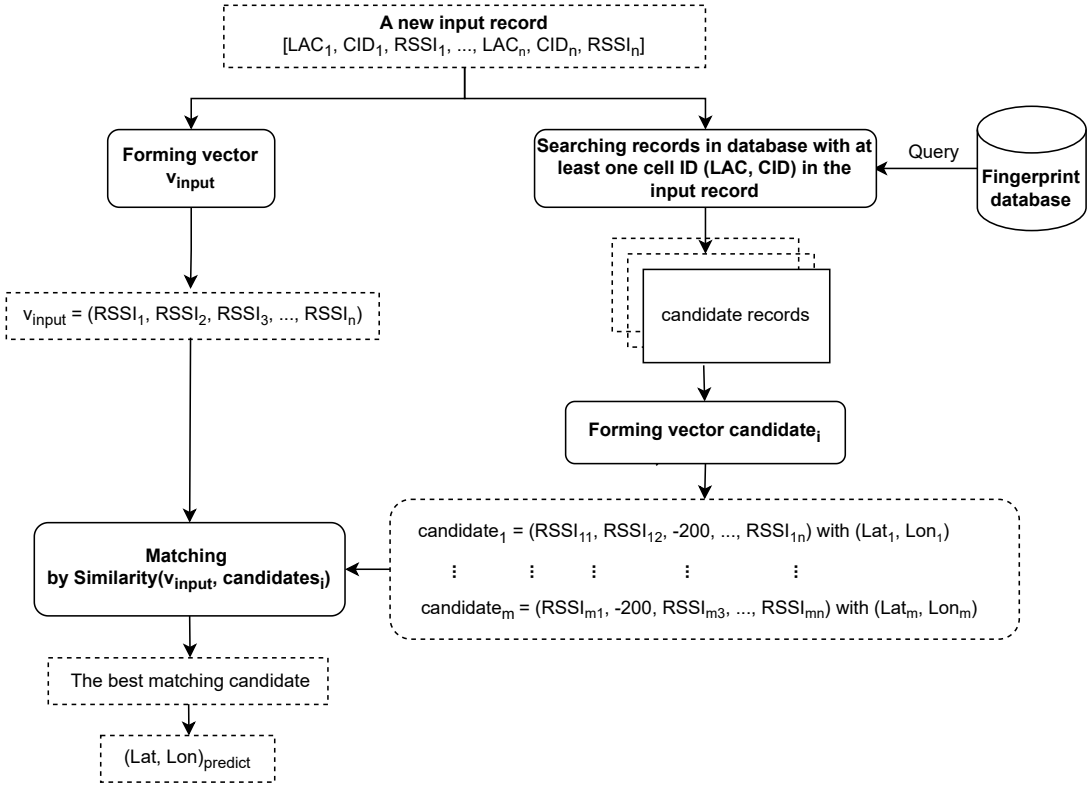


Figure 7: The process of localization phase in fingerprint method using similarity matching

4. Create a white image with size: width x height = 100 x 100 pixels

5. For each cell c_i in the cluster of cells:

   (a) Identify the coordinate (x, y) of pixel for c_i in the image by the formula:
   if $d_{max}! = 0$ then

$$x = \frac{dLon_{c_i}}{2 \times d_{max}} \times (width - 5) + \frac{width}{2}$$
$$y = \frac{dLat_{c_i}}{2 \times d_{max}} \times (height - 5) + \frac{height}{2} \quad (11)$$

   else

$$(x, y) = (\frac{weight}{2}, \frac{height}{2}) \quad (12)$$

   (b) Calculate the pixel value p of the point (x, y) in the image of the cell c_i based on the RSSI value by the formula:

$$p = \lfloor \frac{RSSI_i + 120}{120} \times 255 \rfloor \quad (13)$$

   (c) Draw the square blob of pixels representing cell c_i on the image with the center at the point (x, y), the size 5x5 and the value of the pixels p. If there is an overlap area of the blobs, the pixel value of that area is calculated by the sum of p of the corresponding blobs.

6. Convert the reference (or ground truth) position for each image based on the centroid by:

$$dLat_{ref} = Lat_{ref} - Lat_{centroid}$$
$$dLon_{ref} = Lon_{ref} - Lon_{centroid} \quad (14)$$

Finally, we obtain an image representing the cluster of cells for each data record. A sample image is shown as in Figure 8.

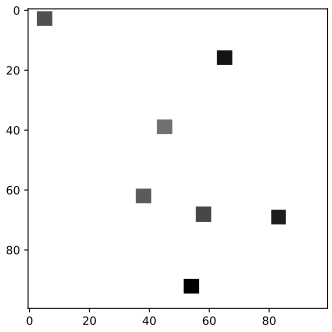


Figure 8: A sample image contains the geographical relationship and RSSI representation of cells

Once the image data set is constructed, we use a CNN model on this data set to predict the output values of *dLat* and *dLon* corresponding to the input image. Figure 9 shows the architecture of CNN model using in this experiment. Subsequently, the estimated position

is calculated on the basis of the output values and the centroid value of the input image.

5 Experiment Results

Experiments were conducted for the methods described in the previous section using a dataset comprising 21155 records derived from measurements taken along the streets in various urban areas of Hanoi city, characterized by numerous houses, people, and vehicles. We evaluated two experiment scenarios with two ways of splitting the data set. In Scenario 1, the data set is divided into training set and test set in an 80:20 ration, respectively, in which, with each five records of five continuous measurements, there are four records used for training set and one record used for test set. This process results in 16924 records for the training set and 4231 records for the test set. In Scenario 2, the data set is also divided into a training set and a test set in approximately 80:20 ration, respectively. However, the training set and the test set is built from measurements taken in different days. This helps the training set and the test set contain records from measurements independently of environment conditions and time.

For each scenario, the training set is used for methods that include linear regression, support vector regression, multilayer perceptron, and CNN-CellImage. The training set is also used as database for fingerprint method. Then, all methods are evaluated on the same test set to compare the results. The accuracy of location prediction for each method was assessed on the test set using the distance error metric as described by Equation 3 and visualized using corresponding histogram distributions of these distance errors. For Scenario 1, the results of the distance error for each method were considered and presented in Table 1.

Figure 10 illustrates the histogram distributions of distance errors in the test set for the centroid method (a) and the weighted centroid method (b) in Scenario 1. Both methods exhibit similar performance in terms of the distance error metric, with mean errors of 160.7 and 155.3 meters, respectively, as indicated in Table 1. The weighted centroid method slightly outperforms the centroid method.

Figure 11 presents the histogram distributions of distance errors on the test set for the Linear Regression method (a) and the Support Vector Regression method (b) in Scenario 1. In particular, support vector regression demonstrates significantly superior performance compared to Linear Regression, with mean errors of 32.2 and 137.3 meters, respectively, as shown in Table 1. These machine learning methods also outperform centroid-based methods.

Figure 12 displays the histogram distributions of distance errors on the test set for the Multilayer Perceptron method (a) and the Fingerprint method (b) in Scenario 1. The experimental results indicate that the multilayer perceptron method achieves an accuracy similar to that of support vector regression. However, the Fingerprint method outperforms all investigated methods, boasting a mean error of 26.1 meters and a minimum error of 0.2 meters, as shown in Table 1. However, the fingerprint method also exhibits a notable maximum distance error of 610.8 meters.

Figure 13 depicts the histogram distribution of distance errors on the test set for the CNN-CellImage method in Scenario 1. It is evident that the CNN-CellImage method yields superior results compared to other methods, excluding the fingerprint method. The CNN-CellImage method achieves a mean error of 33.1 meters and a minimum error of 0.3 meters, as detailed in Table 1.

For Scenario 2, the experiment results of the distance error for each method were considered and presented in Table 2.

Figure 14 illustrates the histogram distributions of distance errors in the test set for the centroid method (a) and the weighted centroid method (b) in Scenario 2. Both methods exhibit similar performance in terms of the distance error metric, with mean errors of 159.0 and 153.1 meters, respectively, as indicated in Table 2. The weighted centroid method slightly outperforms the centroid method. Compared to Scenario 1, the centroid and weighted centroid methods procedures perform similarly to that in Scenario 2. These methods are not dependent on the training stage.

Figure 15 presents the histogram distributions of distance errors on the test set for the Linear Regression method (a) and the Support Vector Regression method (b), and Figure 16a show the result of the Multilayer Perceptron method in Scenario 2. These three machine learning methods produce similar performance with mean errors of 137.6, 131.7, and 137.2 meters, respectively, as shown in Table 2, better than centroid-based methods. However, in Scenario 2, the Support Vector Regression and Multilayer Perceptron methods produce the distance errors much higher than that in Scenario 1. This observation shows that the change of environment condition, which is caused by measuring the test set on different days, greatly impacts the accuracy of these machine learning methods.

Figure 16b displays the histogram distributions of distance errors on the test set for the Fingerprint method (b) in Scenario 2. The fingerprint method gives a mean error of 69.5 meters, as shown in Table 2. Compared to Scenario 1, the distance errors result of the Fingerprint method in Scenario 2 is also higher than that in Scenario 1. The Fingerprint method is also greatly affected by the change of environment condition when collecting data set. To reduce this, we need to make more measurements at different times with different environment conditions to enrich the fingerprint database.

Figure 17 shows the histogram distribution of distance errors in the test set for the CNN-CellImage method in Scenario 2. Our proposed CNN-CellImage method produces the result of distance errors higher than in Scenario 1, as detailed in Table 2. This method is

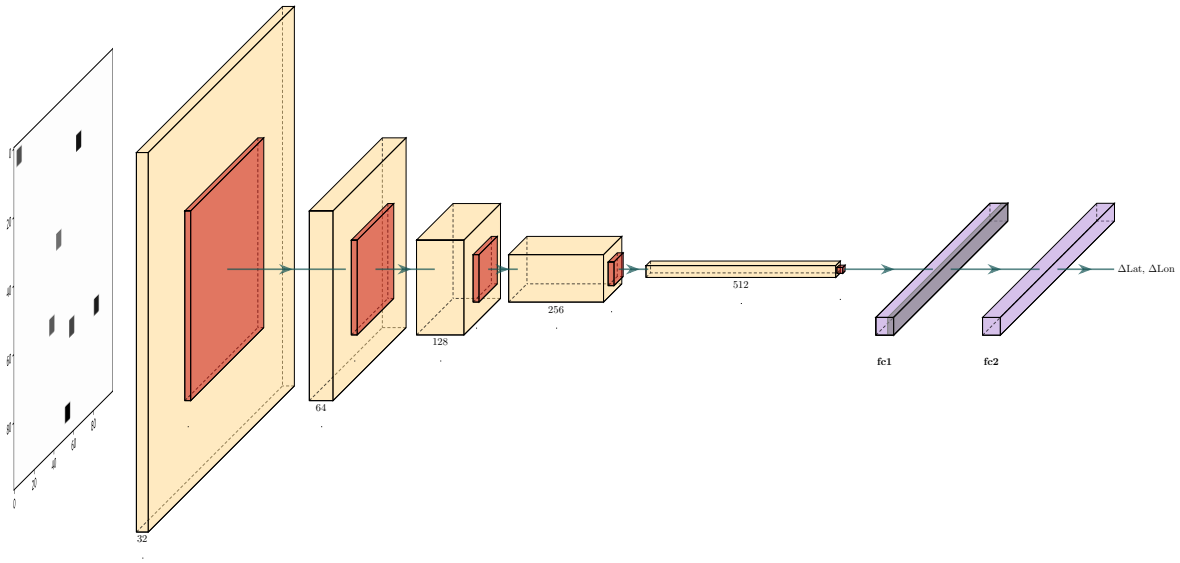


Figure 9: The illustration of CNN-CellImage method with the architecture of CNN model for predicting the output dLat, dLon from one input image

Table 1 Distance error of localization methods in Scenario 1

Method	Mean (m)	Median (m)	Min (m)	Max (m)
Cell ID	242.6	199.6	4.8	1522.8
Centroid	160.7	150.6	1.0	471.0
Weighted Centroid	155.3	144.6	2.8	403.4
Linear Regression	137.3	126.6	1.5	496.1
Support Vector Regression	32.2	25.5	0.2	325.4
Multilayer Perceptron	29.7	21.3	0.3	400.3
Fingerprint	26.1	13.5	0.2	610.8
CNN-CellImage	33.1	19.8	0.3	642.1

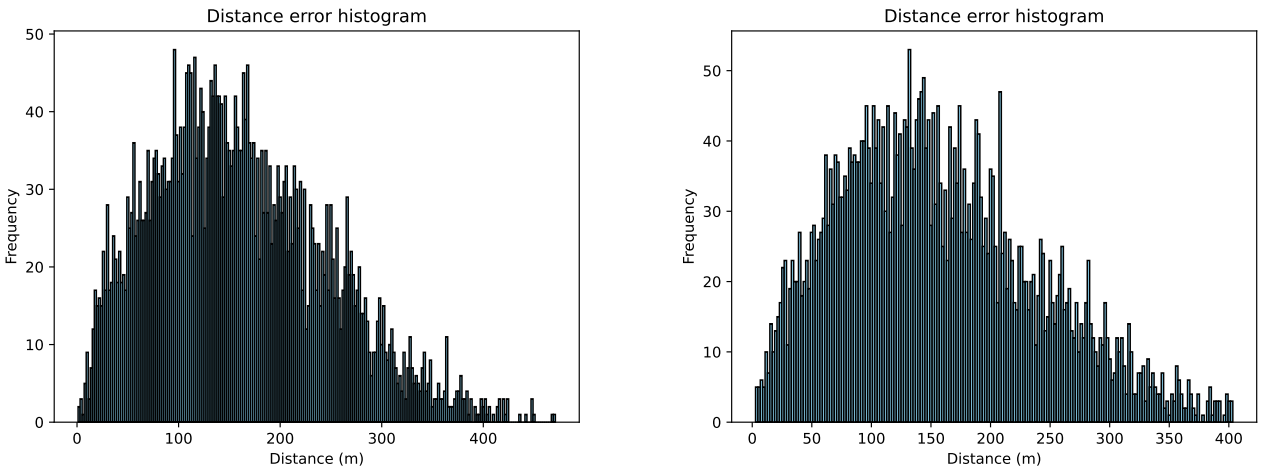
Table 2 Distance error of localization methods in Scenario 2

Method	Mean (m)	Median (m)	Min (m)	Max (m)
Cell ID	231.5	195.3	3.2	1431.0
Centroid	159.0	152.0	1.0	449.5
Weighted Centroid	153.1	146.3	1.7	401.1
Linear Regression	137.6	128.3	2.4	507.3
Support Vector Regression	131.7	121.5	1.3	409.6
Multilayer Perceptron	137.2	125.4	2.7	530.3
Fingerprint	69.5	50.1	0.7	441.0
CNN-CellImage	119.7	108.4	1.9	440.4

also influenced by the change of environment condition when building the data set.

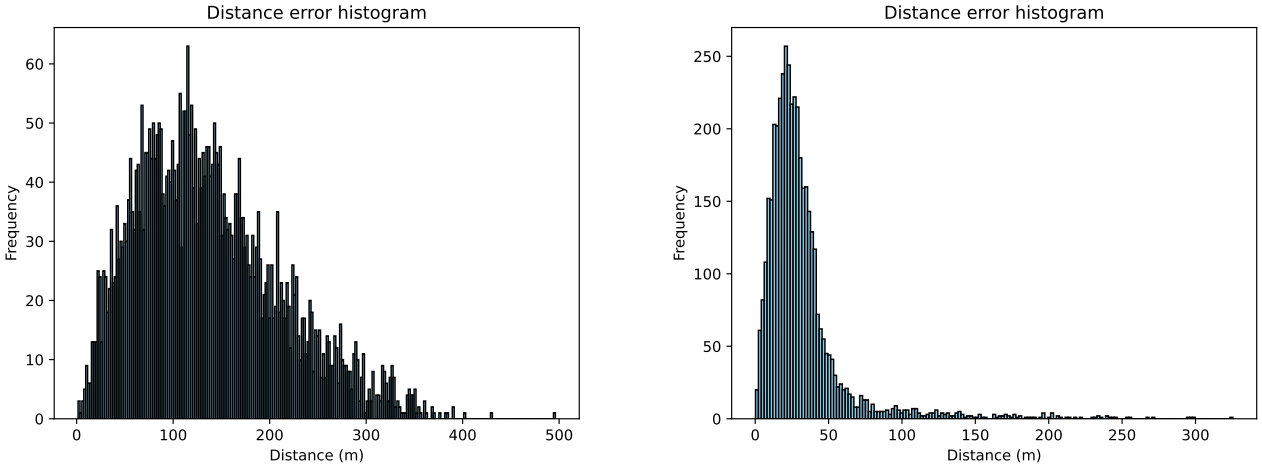
6 Conclusion

In this study, we present the approaches for the location of mobile devices within a cellular network. Our research makes several key contributions. Firstly, we propose a method for gathering and processing data to construct a dataset containing cell information, including location coordinates and signal strength RSSI values. Secondly, we evaluate various localization methods such as Cell-ID, centroid, weighted centroid, linear regression, support vector regression, multilayer perceptron, and fingerprinting. Third, we introduce an original localization approach, the so-called CNN-CellImage method, which utilizes a CNN model



(a) (b)

Figure 10: Histogram of distance error of centroid method (a) and weighted centroid method (b) in Scenario 1

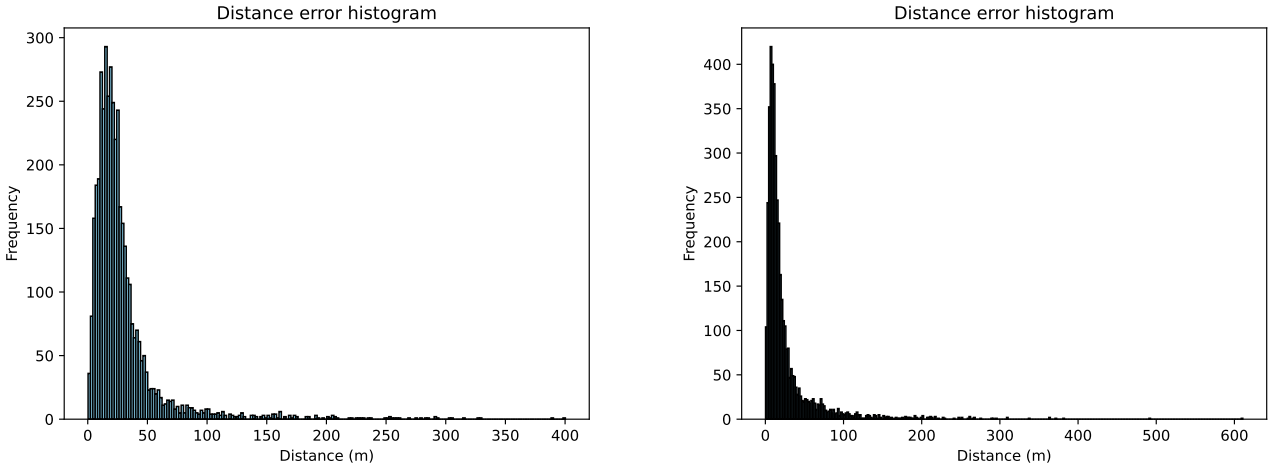


(a) (b)

Figure 11: Histogram of distance error of linear regression method (a) and support vector regression method (b) in Scenario 1

with image data encapsulating geographical location information and RSSI values of cells observed by a mobile device at a time. Our experimental findings indicate that the fingerprint method yields the most accurate results with minimal distance error metric. However, this method necessitates the creation of a fingerprint database, the size of the database significantly influences its accuracy. The CNN-CellImage method, leveraging geolocation images of cells, demonstrates consistent and satisfactory performance. Machine learning techniques such as Linear Regression, Support Vector Regression, and Multilayer Perceptron also yield favorable results, surpassing centroid-based methods. Nevertheless, centroid methods remain appealing because of their simplicity and cost-effectiveness. The results of two experiment scenarios also show that the change in environment condition has a huge impact on methods linear regression, support vector regression, multilayer perceptron, fingerprint, and CNN-CellImage.

For future research, it is essential to update the fingerprint database regularly to accommodate environmental changes and ensure high precision in new tests. A method of alleviating the burden of frequently updating the radio fingerprint database is the passive crowd-sourcing approach. Passive crowd-sourcing aims at performing location annotation without any user intervention. With the user's permission, a background application is installed on the user's smartphones with the task of autonomously collecting RSSI measurements and associating locations with these measurements. And these measurements data is sent to a server for updating the fingerprint database.



(a) (b)

Figure 12: Histogram of distance error of MLP method (a) and fingerprint method (b) in Scenario 1

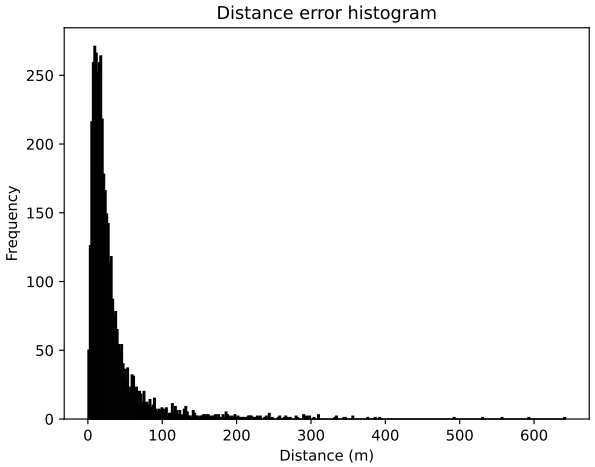


Figure 13: Histogram of distance error of CNN-CellImage method in Scenario 1

References

del Peral-Rosado, José A. and Raulefs, Ronald and López-Salcedo, José A and Seco-Granados, Gonzalo (2003). 'Survey of cellular mobile radio localization methods: From 1G to 5G', *IEEE Communications Surveys & Tutorials*, Vol. 20, No. 2, pp.1124–1148.

Martinka, J. (2019). 'Locating mobile phones using signal strength measurements', *Master's Thesis*, Masaryk University.

Laoudias, C., Moreira, A., Kim, S., Lee, S., Wirola, L., Fischione, C. (2018). 'A survey of enabling technologies for network localization, tracking, and navigation', *IEEE Communications Surveys & Tutorials*, 20(4), 3607–3644.

Wang, X., Wang, Z., O'Dea, B. (2003). 'A toa-based location algorithm reducing the errors due to non-line-of-sight (nlos) propagation', *IEEE Transactions on Vehicular Technology*, 52(1), 112–116.

Brida, P., Cepel, P., Duha, J. (2009). 'Mobile positioning in next generation networks', *In Handbook of research on heterogeneous next generation networking: Innovations and Platforms*, IGI Global, 223–252.

Deligiannis, N., Louvros, S., Kotsopoulos, S. (2007). 'Mobile positioning based on existing signalling messages in gsm networks', Proceedings of the 3rd MOBIMEDIA.

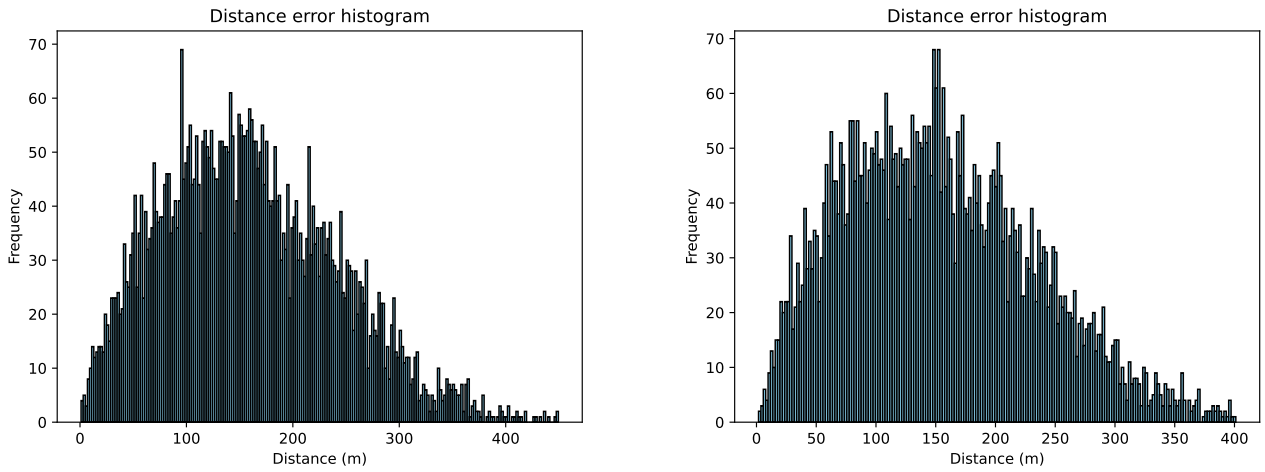
Brida, P., Cepel, P., Duha, J. (2006). 'A novel adaptive algorithm for rss positioning in gsm networks', *In Proceedings of the International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP'06)*. 748–751.

Hata, M., Nagatsu, T. (1980). 'Mobile location using signal strength measurements in a cellular system', *IEEE Transactions on Vehicular Technology*, 29(2), 245–252.

Laitinen, H., Nordström, T., Lähteenmäki, J. (2000). 'Location of gsm terminals using a database of signal strength measurements', *In URSI XXV National Convention on Radio Science*, 88–89.

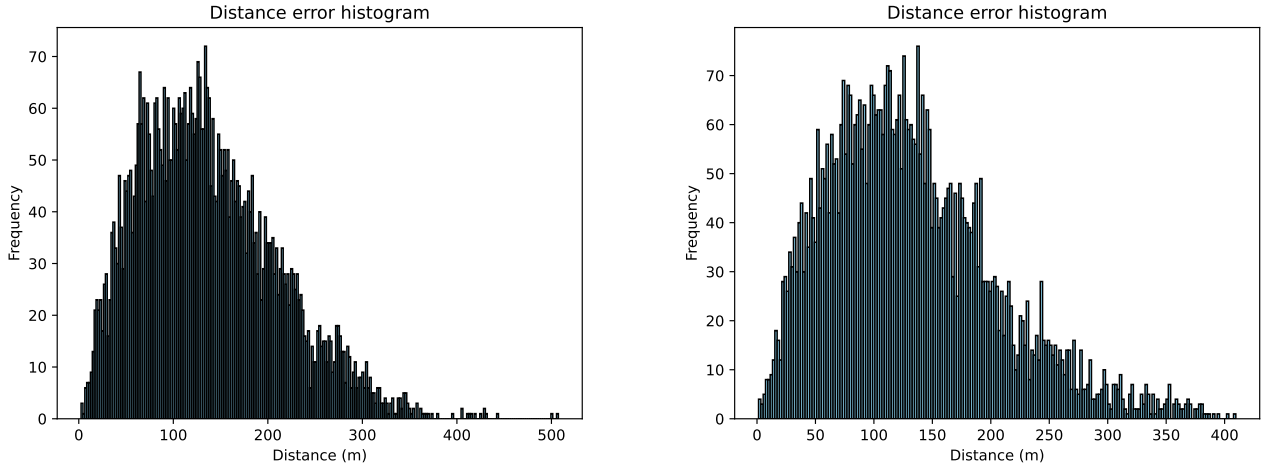
Le, M. H. (2022). 'Various positioning algorithms based on received signal strength and/or time/direction (difference) of arrival for 2d and 3d scenarios', *Ph.D. thesis*, Sorbonne université.

Bartoletti, S., Conti, A., Dardari, D., Giorgetti, A. (2018). '5g localization and context-awareness',*5G Italy White Book: From Research to Market*, 167–187.



(a) (b)

Figure 14: Histogram of distance error of centroid method (a) and weighted centroid method (b) in Scenario 2



(a) (b)

Figure 15: Histogram of distance error of linear regression method (a) and support vector regression method (b) in Scenario 2

Brída, P., Benikovský, J., Machaj, J. (2010). 'Localization in real gsm network with fingerprinting utilization', *In 2nd International ICST Conference on Mobile Lightweight Wireless Systems MOBILIGHT*

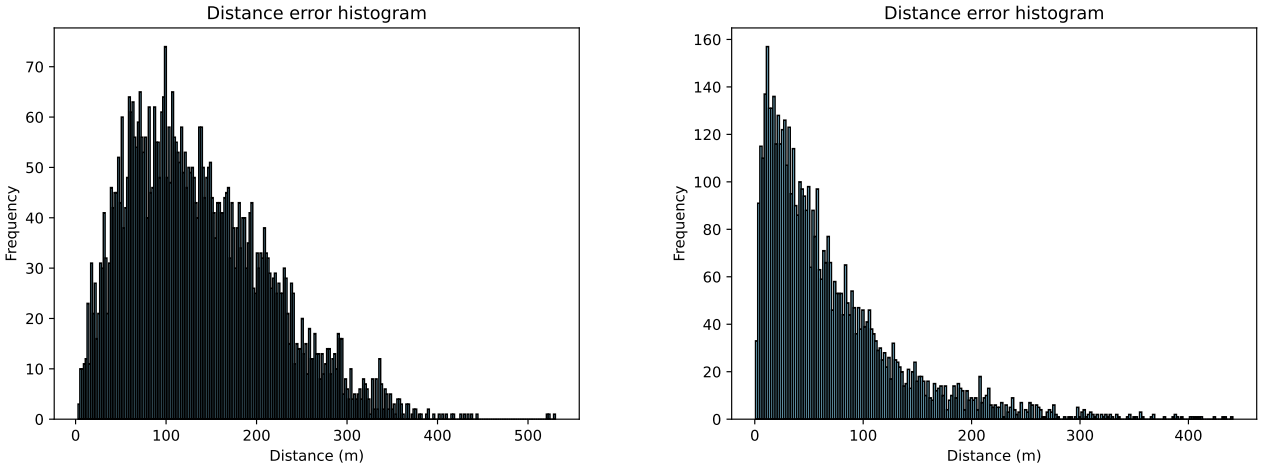
Takenga, C. M., Wen, Q., Kyamakya, K. (2006). 'On the accuracy improvement issues in gsm location fingerprinting', *In IEEE Vehicular Technology Conference*, 1–5.

Vin, I., Gaillot, D. P., Laly, P., Liénard, M., Degauque, P. (2015). 'Overview of mobile localization techniques and performances of a novel fingerprinting-based method', *Comptes Rendus Physique*, 16(9), 862–873.

Calábek, A. (2020), Localization of mobile devices using machine learning, *Bachelor's Thesis*, Masaryk University.

Phruksahiran, N. (2023). 'Improvement of source localization via cellular network using machine learning approach', *Telecommunication Systems*, 82(2), 291–299.

Abdallah, A. A., Saab, S. S., Kassas, Z. M. (2018). 'A machine learning approach for localization in cellular environments', *In 2018 IEEE/ION position, location and navigation symposium (PLANS)*, 1223–1227.



(a) (b)

Figure 16: Histogram of distance error of Multilayer Perceptron method (a) and Fingerprint method (b) in Scenario 2

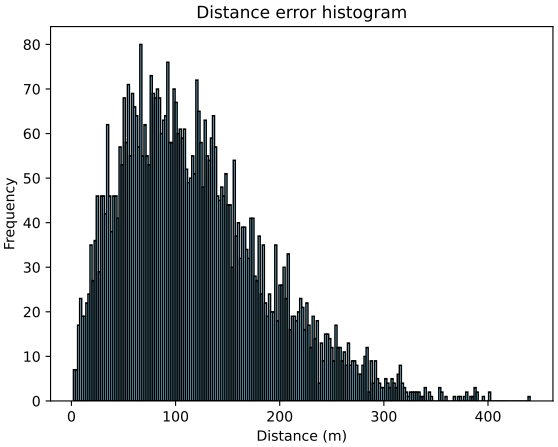


Figure 17: Histogram of distance error of CNN-CellImage method in Scenario 2