**Indoor localization and Wi-Fi Fingerprinting**

DSAI 3201 Machine Learning: Project Part 1

Indoor localization and Wi-Fi Fingerprinting Using Machine Learning and Deep Learning

Joy Anne P. Dela Cruz[[1]](#footnote-1)\*

Student, University of Doha for Science and Technology, 60301959@udst.edu.qa

Indoor localization is a widely used system for determining the position of devices indoors. As part of the University of Doha for Science and Technology's DSAI-3201 Machine Learning Course, this project will analyze the UJIIndoorLoc database to explore Wi-Fi fingerprinting techniques for indoor positioning. This first of two parts will overview the concepts behind indoor localization using Wi-Fi Fingerprinting, highlight key feature observation, formulating feature engineering strategies, and validate these strategies with related works.

CCS CONCEPTS • Wi-Fi Fingerprinting • Data Exploration • Machine Learning • Convolutional Neural Networks

**Additional Keywords and Phrases:** Wi-Fi Fingerprinting, Indoor localization

1. Introduction

Indoor localization, also known as indoor positioning system (IPS), is a technology that determines the position of objects or people within enclosed spaces [1]. There are many ways of applying indoor localization; this project will focus on using Wi-Fi Fingerprinting. There are three components to apply indoor localization in this method. (1) Wi-Fi access points that broadcast signals to be received by (2) mobile devices; these signals will be input in (3) an algorithm that processes the data to give an estimated location. Wi-Fi Fingerprinting technique relies on signal strength data which represents the distance of the Wi-Fi access point from a mobile device. It involves training the fingerprinting algorithm to learn the RSSI (Received Signal Strength Indicator) at various points, making markings in the indoor environment creating a fingerprint database [2]. Then, a mobile device can detect signals from the access points and the database matches this current information with the ones stored from the training phase to estimate the device's location. This project aims to create and implement the algorithm using Machine Learning and Deep Learning.

* 1. Dataset

The UJIIndoorLoc dataset is a Multi-Building, Multi-Floor indoor localization database designed to test Indoor Positioning Systems that rely on WLAN/Wi-Fi fingerprinting. It was created by Torres-Sospedra et al [3]. This dataset contains 528 attributes. Of these, 520 are RSSI values, which represent signal intensities from different Wi-Fi Access Points (WAPs). These values help determine the presence and strength of Wi-Fi signals at various locations. The dataset also includes positional attributes: Longitude and Latitude (provide geographic coordinates), Floor and Building ID (indicate where the data was recorded). The other features in this dataset will provide context but they will not be used in the model: Space ID, Relative Position, User ID, Phone ID, and Timestamp.

* 1. Objective

With my knowledge of Wi-Fi fingerprinting, my task is to explore, clean, and prepare this dataset for model development. I am assigned to develop two models using RSSI features:

• Classification Model – Predicts Building ID and Floor using Wi-Fi fingerprints.

• Enhanced Regression Model – Predicts Longitude and Latitude to determine precise location coordinates based on Wi-Fi fingerprints.

These models will help improve indoor localization, which is essential for navigation in areas where GPS is unreliable.

1. Dataset Exploration & Initial Findings

This section will walk through the data exploration with visualizations and code as well as present findings from these.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Descriptive analysis of the UJIIndoorLoc dataset

To start the dataset exploration, a descriptive statistic of all the columns was calculated. It seems that some Wi-Fi access points (WAPs) are not detected at all based on their minimum and maximum of the same value, meaning these WAPs provide no variation in signal strength. It is also seen that there are a total of five floors—floors 0 to 4—and there are 3 buildings—building 0 to building 2. I moved on to splitting the columns into RSSI columns and other columns.

A graph of different sizes and colors

AI-generated content may be incorrect.

Figure 2: Distribution plots of the non-RSSI columns

I plotted the distributions of the non-RSSI attributes separately using histograms. There was not much information to be gained besides that Floor 4 is not used as much as the other floors and Building 2 is used a little under twice as much as the other buildings.

A diagram of a number of attributes

AI-generated content may be incorrect.

Figure 3: Correlation matrix of the non-RSSI columns

Checking the correlation of the other columns using a correlation matrix, we can see that building ID is highly correlated with longitude and latitude. While expected, this doesn’t provide much useful information since longitude and latitude already define a building’s position. Floor is not correlated to any of the other columns. This suggests that Wi-Fi signal strength is essential for predicting floor level, as no other attribute can help predict it.

A screenshot of a graph

AI-generated content may be incorrect. A screenshot of a graph

AI-generated content may be incorrect. Figure 4: Pair plots of non-RSSI columns colors by floor and building number

To go deeper, pair plots are created color-coded according to Floor and BuildingID. Since there was difficulty with the syntax, OpenAI’s GPT 4o model was used to write the code for this part. There is no strong distinction between floors when looking at positional features, which is similar to the findings in the correlation matrix. This can confirm that floor level cannot be easily inferred from longitude, latitude, or other non-RSSI features hence why Wi-Fi Fingerprinting exists. BuildinID, however, has a clear divide on the longitude and latitude. However, since we're using RSSI to indicate these longitude and latitude, it is not as useful as mentioned before.

A graph of colored lines

AI-generated content may be incorrect.

Figure 5: Scatter plot of record by floor and by building based of coordinates

To predict floor, building number, and coordinates, I wanted to understand these positional attributes better with visualization. I plotted scatter plots for each floor and buildings to see the layout of the space we are working with. The first feature that stood out was that only Building 2 has a 4th floor. This explains my previous observation of having less records in Floor 4 and more in Building 2.

A graph with different colored squares

AI-generated content may be incorrect.

A graph of different colored squares

AI-generated content may be incorrect.

Figure 6: Scatter plot of record by floor and by building based of coordinates

Next, I sought to figure out the frequency distribution of all floors and buildings. This confirmed that there are more records in building 2 and we can see that only building 2 has a 4th floor. Even without the 4th floor, building 2 still appears the most frequent among all buildings.

A graph of a graph of a graph

AI-generated content may be incorrect.

Figure 7: Distribution plots and RSSI values across all WAP features

To create a better model, it is crucial to understand the RSSI values across all WAPs. To create the distribution plots better and more accurately, I changed the no signal value from 100 to negative 105. The distribution of detected signals is overshadowed by the no signal value, so it is removed. If we plot the distribution again without those missing signals, we can see better distribution skewed to the right.

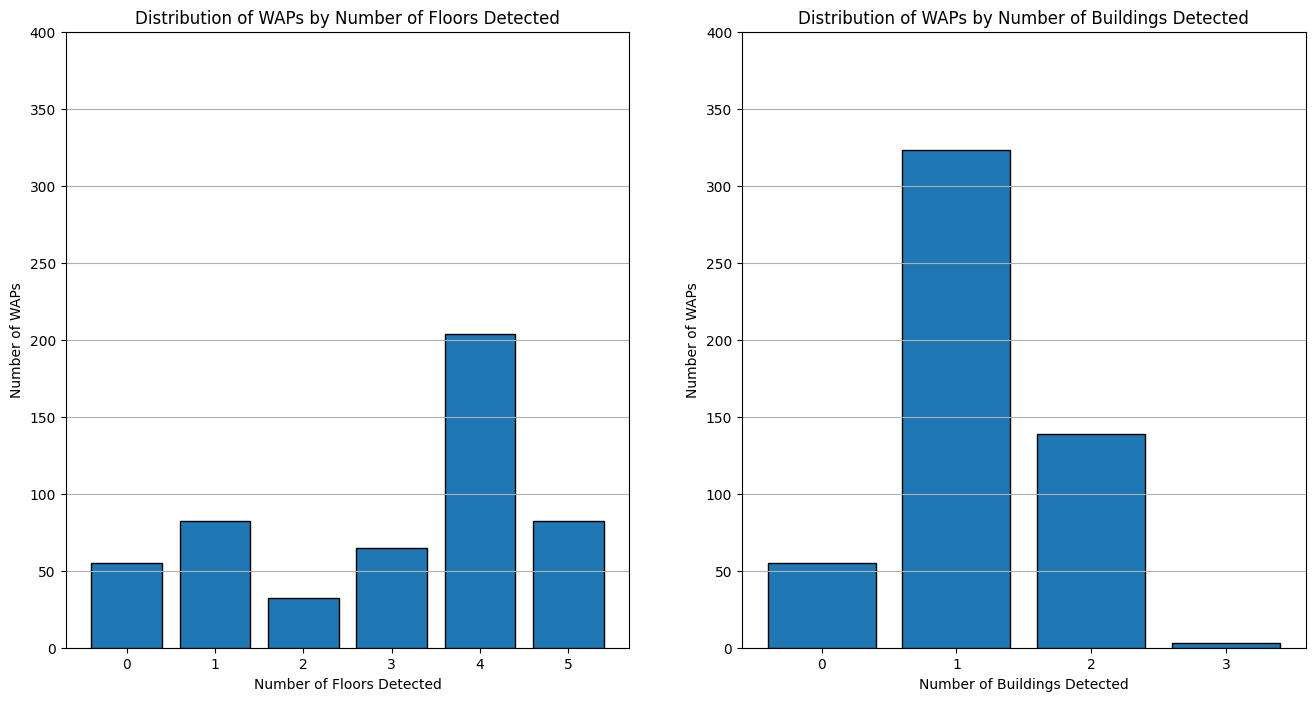


Figure 8: Detecting the RSSI on each WAP per building and floor

To find out whether Wi-Fi access points are floor specific or building specific, the visualization above was ran. It shows that signals from one WAP can appear on multiple floors and multiple buildings. RSSI travels through floors better as more WAP signals travel to 4 floors while it doesn’t travel through building as much as most RSSI signals stay in the same building. Moreover, there were 55 WAPs that did not receive any signals at all, and these features are dropped.

A graph of a bar graph

AI-generated content may be incorrect.

Figure 9: Boxplot of RSSI signals on each WAP per floor

A graph of different sizes and colors

AI-generated content may be incorrect.

Figure 10: Boxplot of RSSI signals on each WAP per building

To learn more about the distribution of the signal strength across floors and buildingIDs on different WAPs, boxplots on the first few RSSI columns are plotted. Most WAPs show very clear high signals among floors and buildings while others like WAP008 do not.

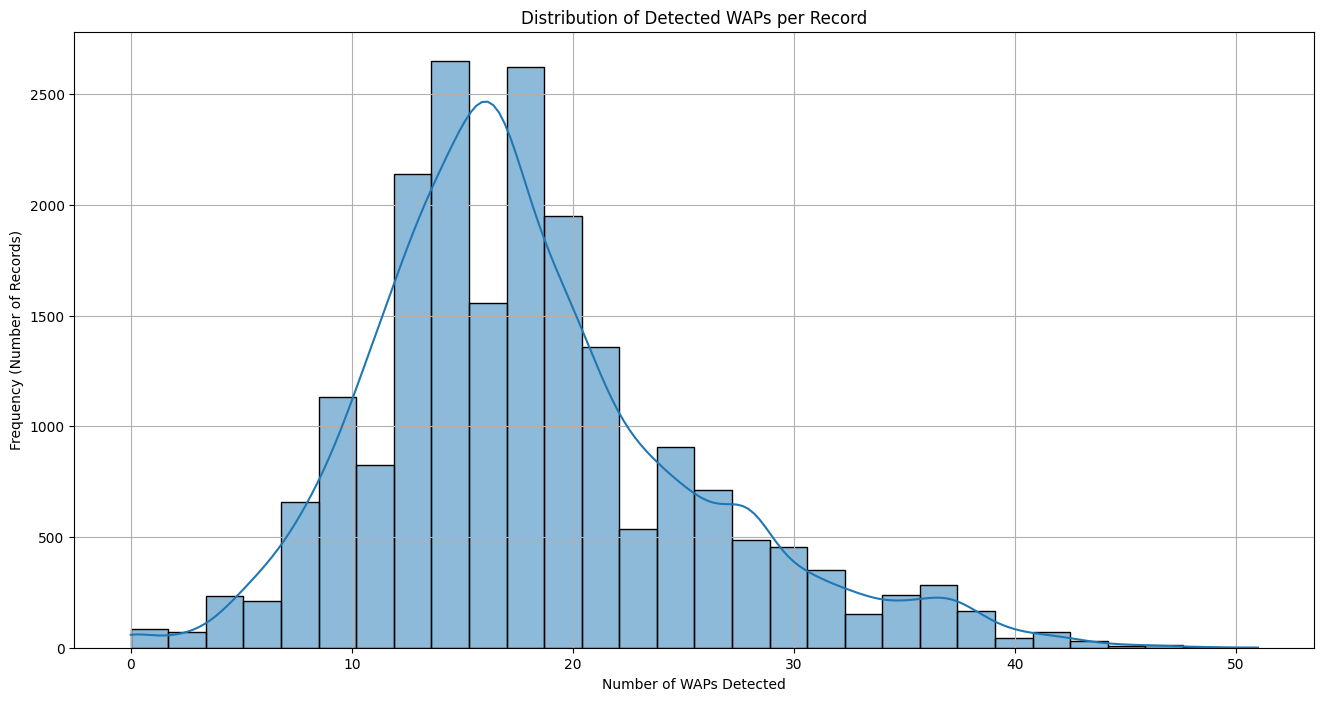


Figure 10: Displaying the distribution of the number of detected WAP for all records

Initially, I thought that each record would only show a detected signal from one WAP, but it turns out that’s not the case—some records detect signals from multiple WAPs at the same time. This can create a challenge when making predictions because the overlapping signals might introduce noise or make it harder to distinguish between locations.

Right now, my plan is to feature engineer by figuring out the strongest signal for each WAP and mapping it to a specific building or floor and filtering the strongest signal per record, since multiple WAPs are detected in a single record. To verify my approach, I took more time to deepen my understanding of Indoor localization, Wi-Fi signals, and Wi-Fi fingerprinting. This will help refine the way we process the data and ensure that the model captures the most useful signals for accurate indoor localization.

1. Further Research to Improve Feature Understanding

I first confirmed that one device can detect multiple WAP signals simultaneously. This happens during initial network scanning, network reconnection, roaming, and periodical background scanning [4]. Wi-Fi signals can travel through multiple rooms and floors, but their strength and quality are significantly affected. Signal strength decreases with distance and obstacles. Wi-Fi signals attenuate, meaning the signals weaken due to the noisy environment. That is why we have many RSSI values in our dataset. Sometimes these signals bounce off surfaces causing an interference, but these RSSI values can help approximate the distance with triangulation-based methods and estimations. For reliable connectivity, signal strengths between -50 dBm and -70 dBm are ideal [5], [6]. I will incorporate this into my data exploration by filtering the signals above -70dBm.

In each record, we can capture multiple RSSI values from multiple WAP and we can use it to help approximate the distance with triangulation-based methods. Since some records show multiple strong signals, we will explore on keeping the top few strongest signals for a record instead of oversimplifying to the top strongest signal. I might also filter out the WAPs with higher threshold in RSSI.

This deeper understanding of Wi-Fi fingerprinting will guide feature engineering and model selection, ensuring we use the most reliable signals for accurate indoor localization.

1. EARLY Data CLEANING AND PREPARATION

With more understanding of the dataset and WiFi-fingerprinting, the dataset was prepared for initial model creation. Some data cleaning was already done for the visualization such as turning non-detected RSSI values from 100 to -115 and deleting the WAP columns that were never detected. This section aims to continue that.

The columns were first checked for missing values and from the DataFrame information, there were no null values. This was confirmed by applying a simple code to check for missing values. A MinMaxScaler was used to scale the RSSI columns and the coordinate columns. MinMaxScaler does not assume normal distribution and it works well with distance-based models like SVM and CNNs [7] which will be explored in the later part of this project. The other positional values (floor and buildingId) were encoded into separate columns to, again, cater to distance-based models.

The RSSI columns are plentiful originally containing 520 attributes—465 after the deletion of never detected WAP. These columns could be further narrowed down with Principal component analysis (PCA). PCA reduces the number of features in large datasets to principal components that retain most of the original information [8]. The threshold used is 0.95 which keeps enough components to preserve 95% of the variance. This reduced the RSSI columns from 465 to 167.

USE OF A.I. Tools

While the use of A.I. tools for this project is kept to a minimum, these tools are in the following aspects: creating the pair plots in data exploration, using the stack method in the histograms of RSSI values, combining different plots in subplots with idx, and iterating over rows for detecting RSSI in WAP columns are assisted by OpenAI’s GPT 4o model. The formatting of references is also aided by online A.I. tools like Scribbr’s citation generator and, again, GPT4o.

REFERENCES

1. Magda Chelly and Nel Samama. 2009. New techniques for indoor positioning, combining deterministic and estimation methods. In Proceedings of the European Navigation Conference - Global Navigation Satellite Systems (ENC-GNSS 2009), Naples, Italy, 1–12. <https://hal-01367483>.
2. E. Cheng. 2025. Location fingerprinting—What is it, and should you choose it as your IPS technology? Pointr. Retrieved February 5, 2025, from <https://www.pointr.tech/blog/location-fingerprinting-what-is-it-should-you-choose-it>.
3. J. Torres-Sospedra, R. Montoliu, A. Martinez-Us, T. Arnau, and J. Avariento. 2014. UJIIndoorLoc, UCI Machine Learning Repository. <https://doi.org/10.24432/C5MS59>.
4. Android Open Source Project. n.d. Wi-Fi AP/AP concurrency. Android Open Source Project. Retrieved from <https://source.android.com/docs/core/connect/wifi-ap-ap-concurrency>.
5. MetaGeek. n.d. Wi-Fi Signal Strength Basics. MetaGeek. Retrieved from <https://www.metageek.com/training/resources/wifi-signal-strength-basics/>.
6. M. Pierce. n.d. WiFi Signal Strength: A No-Nonsense Guide. TechGrid. Retrieved from <https://techgrid.com/blog/wifi-signal-strength>
7. N. Lang. 2024. What is the MinMax Scaler? Data Basecamp. Retrieved February 16, 2025, from <https://databasecamp.de/en/ml/minmax-scaler-en>.
8. IBM. 2024. What is PCA? Retrieved February 16, 2025, from <https://www.ibm.com/think/topics/principal-component-analysis>.

1. [↑](#footnote-ref-1)