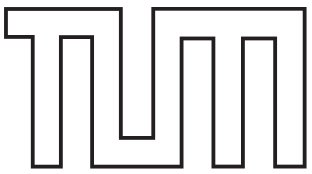

Technische Universität München
Chair of Media Technology
Prof. Dr.-Ing. Eckehard Steinbach

Master Thesis

Deep Learning-Based Indoor Localization Using CSI and RSSI in a Single Access Point Scenario

Author:	Sedki Ben Haouala
Matriculation Number:	03702786
Address:	Dorfstrasse 29
	85737 Muenchen
Advisor:	Prof. Dr.-Ing. Eckehard Steinbach
Begin:	01.04.2025
End:	25.09.2025

With my signature below, I assert that the work in this thesis has been composed by myself independently and no source materials or aids other than those mentioned in the thesis have been used.

München, September 22, 2025 _____

Place, Date Signature

München, September 22, 2025 _____

Place, Date Signature

Abstract

This work investigates the feasibility of deep learning-based indoor localization using commodity Wi-Fi hardware in a single access point (AP) setting. The experimental setup consists of an Archer Wi-Fi router as the transmitter and an ESP32-WROOM module as the receiver, constrained to a single antenna without MIMO support. Channel State Information (CSI) and auxiliary Received Signal Strength Indicator (RSSI) measurements were collected across a predefined spatial grid to capture multipath effects and signal variations in amplitude and phase. The localization task was formulated as a regression problem, where convolutional neural network (CNN) architectures were trained to predict continuous coordinates rather than discrete fingerprint classes. A multi-phase evaluation revealed that improper preprocessing and optimization choices significantly hinder generalization, while methodological corrections led to substantial performance improvements. Among the tested models, a hybrid CNN architecture that combined CSI amplitudes with RSSI achieved a median localization error of 1.193 m with 250 training samples per reference point, alongside 52% sub-meter accuracy. These results highlight that, under constrained hardware conditions, the effectiveness of CSI-based localization depends not only on architectural design but also critically on data quality, preprocessing rigor, and environment-specific sampling density.

Index Terms: Indoor localization (IL), Wi-Fi, Channel State Information (CSI), Received Signal Strength Indicator (RSSI), ESP32-WROOM, Convolutional Neural Networks (CNNs), Fingerprinting.

Contents

Chapter 1

Introduction

Indoor localization, the process of determining the position of a user or object within enclosed environments, has emerged as a critical research topic due to its wide range of applications, including healthcare, retail analytics, smart homes, industrial automation, and emergency response. Unlike outdoor environments, where the Global Positioning System (GPS) provides reliable and accurate positioning, indoor settings introduce unique challenges. Structural obstacles such as walls, ceilings, and furniture, combined with the absence of line-of-sight to satellites, render GPS ineffective indoors. As a result, the development of alternative indoor positioning systems (IPS) has become a priority for both academia and industry.

Several technologies have been proposed to address this challenge, each with inherent advantages and limitations. Vision-based approaches provide detailed localization but raise privacy concerns and often require costly infrastructure. Infrared (IR) and ultrasound systems can achieve high accuracy in constrained areas but are limited by range, susceptibility to interference, and deployment costs. Bluetooth Low Energy (BLE) and Ultra-Wideband (UWB) represent more practical solutions: BLE enables coarse proximity estimation at low cost, while UWB achieves fine-grained positioning via time-of-flight measurements, albeit with expensive hardware.

Wi-Fi-based localization offers a compelling balance between cost-effectiveness and practicality, owing to the ubiquity of Wi-Fi access points in modern buildings. Early methods relied primarily on RSSI fingerprinting, which provided coarse location estimates but suffered from instability and poor generalization. More recently, channel state information (CSI) has emerged as a powerful alternative, as it captures amplitude and phase variations across subcarriers, offering a fine-grained representation of multipath propagation. By exploiting this additional information, CSI-based methods have demonstrated significant improvements in localization accuracy under complex indoor conditions.

State-of-the-art efforts illustrate the potential of CSI for single-AP localization. CUPID [SLKC13], for instance, leveraged CSI to estimate the angle and distance of the direct

path, achieving median errors of approximately 5 m. S-Phaser [HLM+19] extended this concept by exploiting the richer information available from MIMO-enabled access points, using calibrated CSI phases to estimate direct path lengths and achieving median errors near 1.5 m. While these results highlight the benefits of advanced hardware configurations, they were conducted in larger experimental environments, making direct error comparisons to smaller-scale setups less meaningful.

In this work, we experimentally investigate the feasibility, robustness, and limitations of CSI-based indoor localization under a deliberately constrained scenario: a single-antenna access point with no MIMO support, using an ESP32-WROOM ship as a receiver. By constructing an offline reference grid of the study environment, we frame the task as a regression problem, predicting continuous user coordinates rather than performing discrete fingerprint classification. To address the limited variability imposed by commodity hardware, we systematically evaluate a set of convolutional neural network (CNN) architectures and feature combinations, including the integration of auxiliary RSSI measurements. Our analysis aims to quantify the trade-offs between architectural design, dataset size, and generalization performance, thereby situating deep learning-based approaches within the broader landscape of single-AP indoor localization research.

Chapter 2

Related Work

Indoor localization (IL) has become a prominent area of research over the past two decades due to its critical role in numerous applications. These include worker safety in industrial settings, healthcare monitoring for elderly populations, emergency response and disaster relief, and intelligent building management systems. The demand for accurate indoor positioning has further intensified with the proliferation of smart cities and the Internet of Things (IoT). While outdoor positioning is effectively supported by Global Navigation Satellite Systems (GNSS), such as GPS, GLONASS, Galileo, and BeiDou, these systems perform poorly indoors. Signal penetration through building materials is severely limited, leading to attenuation and multipath propagation. This fundamental limitation has motivated extensive research into alternative methods for dedicated indoor positioning.

Indoor localization techniques are commonly classified into active and passive approaches [KRSS22, SGLH18]. Active methods require users or objects to carry electronic devices such as smartphones, RFID tags, or wearables that communicate with anchor nodes in wireless sensor networks (WSNs). In contrast, passive methods, also termed device-free localization (DFL), do not require users to carry devices; instead, they infer presence or location by analyzing perturbations in the environment caused by human motion. This distinction is particularly relevant in safety-critical and healthcare applications, where device-free monitoring may be preferable. Within both categories, localization schemes can be further divided into RF-based and non-RF-based approaches.

Non-RF approaches include optical systems (e.g., vision- or LiDAR-based tracking). For example, [RHDR25] proposed a navigation framework that combines a fine-tuned ResNet-50 CNN with a large language model (LLM) to provide smartphone-based visual positioning and guidance. Other non-RF modalities include acoustic sensing (using ultrasonic or audible signals), inertial measurement units (IMUs) for dead reckoning, magnetic field fingerprinting [HGWY17], visible light positioning (VLP) using modulated LED light [ZYC+24], and pressure or thermal sensors embedded in building infrastructure. While these techniques offer unique advantages, they also face practical limitations such as line-

of-sight constraints, environment-dependent accuracy, high infrastructure costs, or privacy concerns.

In comparison, RF-based methods dominate both academic research and real-world deployment, largely due to the ubiquity of wireless technologies. Wi-Fi, Bluetooth, RFID, ZigBee, and Ultra-Wideband (UWB) have all been studied extensively as localization solutions. Among these, Wi-Fi is the most widely adopted, owing to its near-universal availability indoors and the lack of requirement for additional infrastructure in most environments. Wi-Fi localization typically relies on measuring received signal strength indicator (RSSI) values from multiple access points [KAR+20], making it straightforward to implement. However, its performance is highly sensitive to environmental changes and interference from other devices [A+22]. Bluetooth Low Energy (BLE) represents another widely deployed RF-based technology. By placing beacons throughout an environment, BLE-based localization provides location estimates at low cost and power consumption [RTR+21]. Nevertheless, high localization accuracy often demands a dense deployment of beacons, which can increase complexity and cost [RBS21].

In this study, we focus specifically on Wi-Fi-based indoor localization, which can be broadly categorized into two main approaches: RSSI-based methods and channel state information (CSI)-based methods.

Early Wi-Fi localization systems relied on RSSI fingerprinting. In these systems, a radio map is constructed during an offline phase by recording RSSI vectors at reference locations. In the online phase, user locations are inferred by matching observed RSSI vectors to the stored map. Several methods have been applied for this matching, including nearest-neighbor search [HZY+19, ZCW23], probabilistic inference [PHB18, PB19], and more recent deep learning (DL) techniques. For example, [HCN19] trained a deep neural network (DNN) using RSS fingerprints collected between an access point and a network interface card, achieving improved location estimation during the online phase. Similarly, [PH20] proposed a lightweight one-dimensional CNN that achieved high accuracy with reduced model complexity. [LCC19] highlighted the potential of DNNs for high-precision indoor localization, while also noting the lack of interpretability in such models.

Despite their simplicity and ease of deployment, RSSI-based methods face well-documented limitations. RSSI is a coarse measure of signal power, highly sensitive to multipath fading, device heterogeneity, and dynamic environmental changes. Maintaining acceptable accuracy often requires dense fingerprint maps and frequent recalibration, which constrains scalability. Even with enhancements such as interpolation, probabilistic modeling, and device calibration, sub-meter accuracy remains difficult to achieve in dynamic settings.

To address these limitations, CSI-based localization has emerged as a more powerful alternative. Unlike RSSI, CSI captures fine-grained amplitude and phase information for individual OFDM subcarriers. CSI provides a sampled channel frequency response that encodes multipath characteristics, enabling richer feature extraction. With proper calibration, CSI can be transformed into channel impulse responses (CIRs) and power delay

profiles (PDPs), supporting estimation of time-of-flight (TOF), angle-of-arrival (AOA), and multipath components. For instance, [WWM17] employed bimodal CSI comprising both AOA and amplitude information. Their network architecture processed two channels: one with AOA time images from three antennas, and another with CSI amplitude-time images from a single antenna, significantly improving localization accuracy. CSI thus enables both geometry-based localization and data-driven fingerprinting.

Deep learning has become a central tool for exploiting CSI in indoor localization. Convolutional neural networks (CNNs) have been used to map CSI matrices directly to spatial coordinates [FNL22, HCN19, TG17]. Typical CNN-based approaches treat CSI measurements from multiple transmitters as input matrices, with the network outputting location estimates [SRM22, SCP21]. Recurrent neural networks (RNNs) and long short-term memory (LSTM) models have also been applied to capture temporal dependencies in CSI sequences, improving performance in dynamic environments [YWKAO22]. Compared to RSSI, CSI is more resilient to interference and multipath effects, enabling more precise localization. However, CSI-based methods face challenges related to data complexity and the need for specialized hardware and software for measurement acquisition.

Recent studies suggest that hybrid pipelines combining signal processing with DL yield the most robust results. In such systems, calibrated CSI features (e.g., from phase sanitization or interpolation-enhanced methods) are first extracted, after which neural networks learn mappings that generalize across device heterogeneity and environmental variations. This reduces reliance on dense fingerprinting and enhances robustness against multipath propagation. More advanced approaches include the use of generative adversarial networks (GANs) to augment limited CSI datasets with realistic synthetic samples [LP21] [LLLW25], and autoencoder-based architectures for self-calibration and adaptive fingerprint reconstruction under environmental changes [ZLS+21] [LKS22]. Other work integrates DL with classical algorithms, for example, partitioning regions into sub-areas using DNNs before applying optimized k-nearest neighbor searches [DYWY19].

In summary, the trajectory of Wi-Fi indoor localization has progressed from RSSI-based fingerprinting toward CSI-based ranging and fingerprinting, and more recently toward hybrid DL-driven frameworks. RSSI remains attractive due to its simplicity and ubiquity, but its limitations have motivated increasing adoption of CSI. When combined with DL, CSI provides the fine-grained signal characterization and adaptability needed to meet the accuracy and scalability demands of modern applications. This thesis builds on these developments by jointly leveraging RSSI and CSI features with deep learning models to design a robust, scalable, and accurate framework for indoor localization.

Chapter 3

Theoretical Framework

3.1 Orthogonal Frequency-Division Multiplexing (OFDM)

Orthogonal Frequency-Division Multiplexing (OFDM) is a digital transmission technique widely used in modern wireless communication systems such as Wi-Fi (IEEE 802.11), Long-Term Evolution (LTE), and 5G. It divides the transmission channel into multiple orthogonal subcarriers, each transmitting a portion of the overall data stream [Min24]. The key advantage of OFDM lies in its robustness against inter-symbol interference (ISI) and frequency-selective fading, both of which are common in multipath propagation environments. The term orthogonal refers to the mathematical property of subcarriers being mutually independent, which allows them to overlap in frequency without interfering with each other. This enables efficient spectrum utilization and makes OFDM highly suitable for high-data-rate communication in challenging environments where signal degradation is common, such as indoors.

Fig. 3.1 [Min24] shows the frequency-domain representation of OFDM subcarriers. Each subcarrier has a sinc-shaped spectrum due to rectangular time-domain pulse shaping. Although the subcarriers overlap, the zeros of each function align with the peaks of all others. This ensures orthogonality, meaning that when sampled at the correct frequencies, no inter-carrier interference occurs. This property underpins OFDM ability to pack subcarriers closely together, achieving high spectral efficiency. This orthogonality condition implies that subcarriers must be spaced at intervals of

$$\Delta f = \frac{1}{T} \tag{3.1}$$

where T is the OFDM symbol duration, and Δf is the subcarrier spacing required to maintain orthogonality.

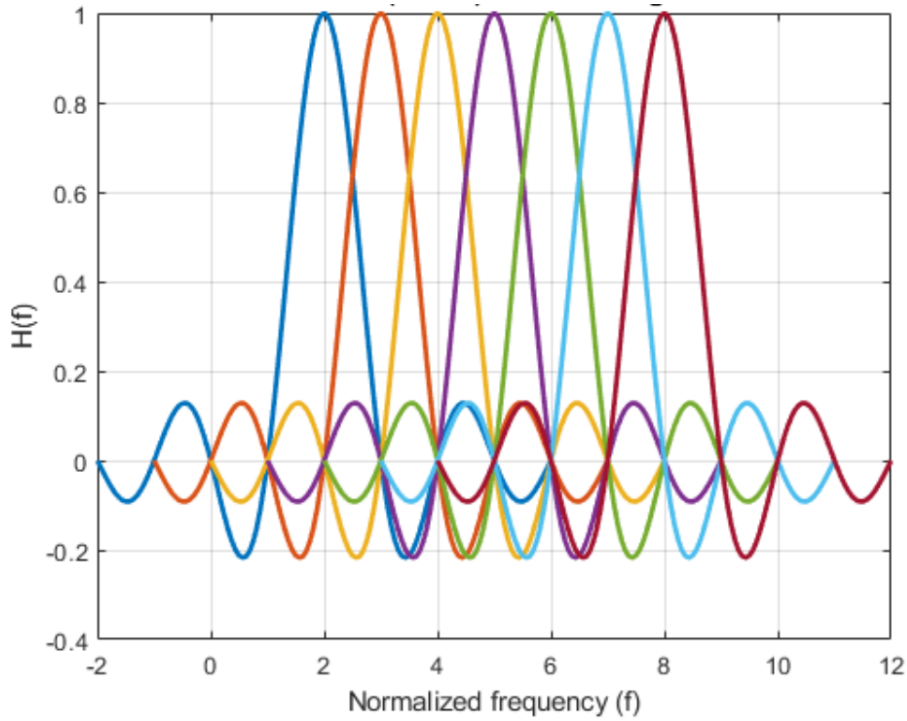

Figure 3.1: Multicarrier OFDM spectrum

The discrete-time OFDM signal is generated by mapping input symbols onto orthogonal subcarriers, which are the sinusoidal basis functions of the Discrete Fourier Transform (DFT). This process allows the subcarriers spectra to overlap while still being separable at the receiver, since the DFT basis functions are uncorrelated. At the transmitter, the Inverse Fast Fourier Transform (IFFT) efficiently implements this mapping by converting the data stream into a time-domain OFDM signal.

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[k] \cdot \exp\left(j \frac{2\pi k n}{N}\right), \quad n = 0, 1, \ldots, N-1 \tag{3.2}$$

where $X[k]$ represents the constellation symbol mapped to the k-th subcarrier and N is the total number of active subcarriers [TB12]. The orthogonality of the DFT basis functions ensures that each subcarrier energy is isolated during demodulation, despite the overlapping spectra illustrated in Fig. 3.1.

In indoor environments, multipath propagation creates frequency-selective fading. While this is a challenge for traditional modulation schemes, OFDM decomposes the channel into flat-fading subchannels, each carrying location-specific signatures. These frequency-domain channel responses form the foundation of Channel State Information (CSI), which can be exploited for indoor localization and sensing applications.

3.1.1 Channel State Information (CSI)

In indoor wireless environments, multipath propagation creates frequency-selective channels where transmitted signals interact with walls, furniture, and other obstacles. These interactions produce multiple propagation paths that arrive at the receiver with different delays, attenuations, and phase shifts as shown in Fig. 3.2 [YZL13]. While traditionally regarded as channel impairments, multipath components in fact carry rich spatial and temporal information that can be leveraged for localization and sensing.

OFDM is particularly well suited for capturing this information, as its frequency-domain representation decomposes the wireless channel into narrowband subcarriers that each experience approximately flat fading. This is because the combined effect of all these multipath components defines the channel's impulse response, whose Fourier Transform yields the complex Channel Frequency Response (CFR) measured at each subcarrier. The response of each subcarrier encapsulates location-specific signatures, collectively forming the Channel State Information (CSI). Unlike coarse power metrics such as RSSI, CSI provides a fine-grained view of the channel across frequency, enabling robust characterization of indoor multipath effects.

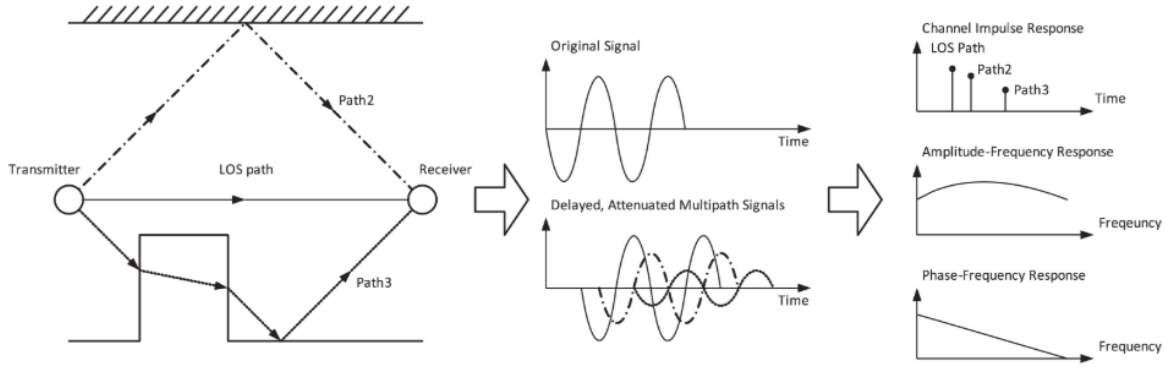

Figure 3.2: Multipath propagations and received signals

The fundamental relationship between the transmitted and received signal on the k-th subcarrier is given by:

$$Y[k] = H[k] \cdot X[k] + \eta[k] \tag{3.3}$$

where:

- $Y[k]$ is the **received symbol** on the k-th subcarrier.
- $X[k]$ is the **transmitted symbol** on the k-th subcarrier. For channel estimation, this is a known *pilot symbol*.

- $H[k]$ is the **true Channel State Information (CSI)** for the k-th subcarrier. It is a complex value representing the amplitude and phase distortion introduced by the channel.
- $\eta[k]$ is the additive **noise** on the k-th subcarrier.

In practice, the true CSI $H[k]$ is unknown. It must be estimated at the receiver. This is achieved using the known pilot symbols. The standard method is the Least Squares (LS) estimator, which calculates the estimated CSI, $\hat{H}[k]$, as:

$$\hat{H}[k] = \frac{Y[k]}{X[k]}, \quad X[k] \neq 0 \tag{3.4}$$

This estimate $\hat{H}[k]$ provides a snapshot of the channel's behavior at the subcarrier frequency but is influenced by noise: $\hat{H}[k] = H[k] + \frac{\eta[k]}{X[k]}$.

Decomposition into Amplitude and Phase

The estimated CSI $\hat{H}[k]$ is a complex number. It can be represented in its Cartesian form with in-phase (real) and quadrature (imaginary) components, or more intuitively for sensing, in its polar form:

$$\hat{H}[k] = \mathrm{Re}\{\hat{H}[k]\} + j \cdot \mathrm{Im}\{\hat{H}[k]\} = |\hat{H}[k]| \cdot e^{j\angle\hat{H}[k]} \tag{3.5}$$

The magnitude (amplitude) response $|\hat{H}[k]|$ and phase response $\angle\hat{H}[k]$ are extracted as follows:

$$\text{Amplitude:} \quad A[k] = |\hat{H}[k]| = \sqrt{(\mathrm{Re}\{\hat{H}[k]\})^2 + (\mathrm{Im}\{\hat{H}[k]\})^2} \tag{3.6}$$

$$\text{Phase:} \quad \phi[k] = \angle\hat{H}[k] = \mathrm{atan2}(\mathrm{Im}\{\hat{H}[k]\}, \mathrm{Re}\{\hat{H}[k]\}) \tag{3.7}$$

where:

- $A[k]$ represents the signal attenuation (or gain) on the k-th subcarrier, indicative of the overall signal strength and path loss.
- $\phi[k]$ represents the phase shift, which is highly sensitive to the path length and thus to subtle changes in the propagation environment, such as those caused by movement.

This complex value $\hat{H}[k]$, and its derived components $A[k]$ and $\phi[k]$ for all subcarriers, form the multi-dimensional data source for our Indoor localization algorithms in this work, providing a rich description of the wireless channel between transmitter and receiver.

3.1.2 Received Signal Strength Indicator (RSSI)

While OFDM provides the physical-layer foundation to cope with multipath fading by decomposing the wireless channel into orthogonal, flat-fading subcarriers, early approaches to indoor localization relied on simpler, power-based metrics. The most widely adopted among these is the Received Signal Strength Indicator (RSSI), a scalar measurement representing the total integrated power of the received signal across the entire channel bandwidth.

Despite its historical prevalence and ease of acquisition, RSSI is highly susceptible to multipath-induced fading, shadowing, and temporal variations as shown in figure 3.3 [YZL13] leading to poor accuracy and stability in complex indoor environments. Consequently, its role in modern Indoor Localization (IL) systems is often limited to providing a coarse proximity estimate or serving as a supplementary feature to more robust, high-dimensional sensing modalities like CSI.

To define RSSI, we start by understanding the received complex baseband voltage In a typical indoor environment [YZL13]. It is the superposition of the transmitted signal arriving at the receiver through multiple propagation paths, each with a distinct delay, attenuation, and phase shift (Fig. 3.2). it can thus be expressed as

$$V = \sum_{i=1}^{N} V_i e^{-j\theta_i}, \tag{3.8}$$

where V_i and θ_i are the amplitude and phase of the i-th multipath component, and N is the total number of components.

The RSSI is then defined as the received signal power in decibels:

$$RSSI = 10 \log_{10} \left(|V|^2 \right). \tag{3.9}$$

As a superposition of multipath contributions, RSSI fluctuates significantly even for static links. Minor variations in relative phase among multipath components may lead to constructive or destructive interference, producing RSSI swings of 5 to 7 dB in typical indoor settings. These fluctuations manifest at both fine temporal scales (seconds) and coarser scales (hours), limiting the robustness of RSSI-based positioning.

Despite these instabilities, RSSI is mapped into the distance from the transmitter by the prevalent Log-normal Distance Path Loss (LDPL) [YZL13] (Fig. 3.3):

$$\mathrm{PL}(d)[\mathrm{dB}] = \mathrm{PL}(d_0) + 10n \log \left(\frac{d}{d_0} \right) + X_\sigma, \tag{3.10}$$

where $\mathrm{PL}(d)$ denotes the measured path loss at distance d. $\mathrm{PL}(d_0)$ is the average path loss at reference point d_0 and n is the path loss exponent. X_σ is a zero-mean normal random variable reflecting the attenuation in decibel caused by shadowing.

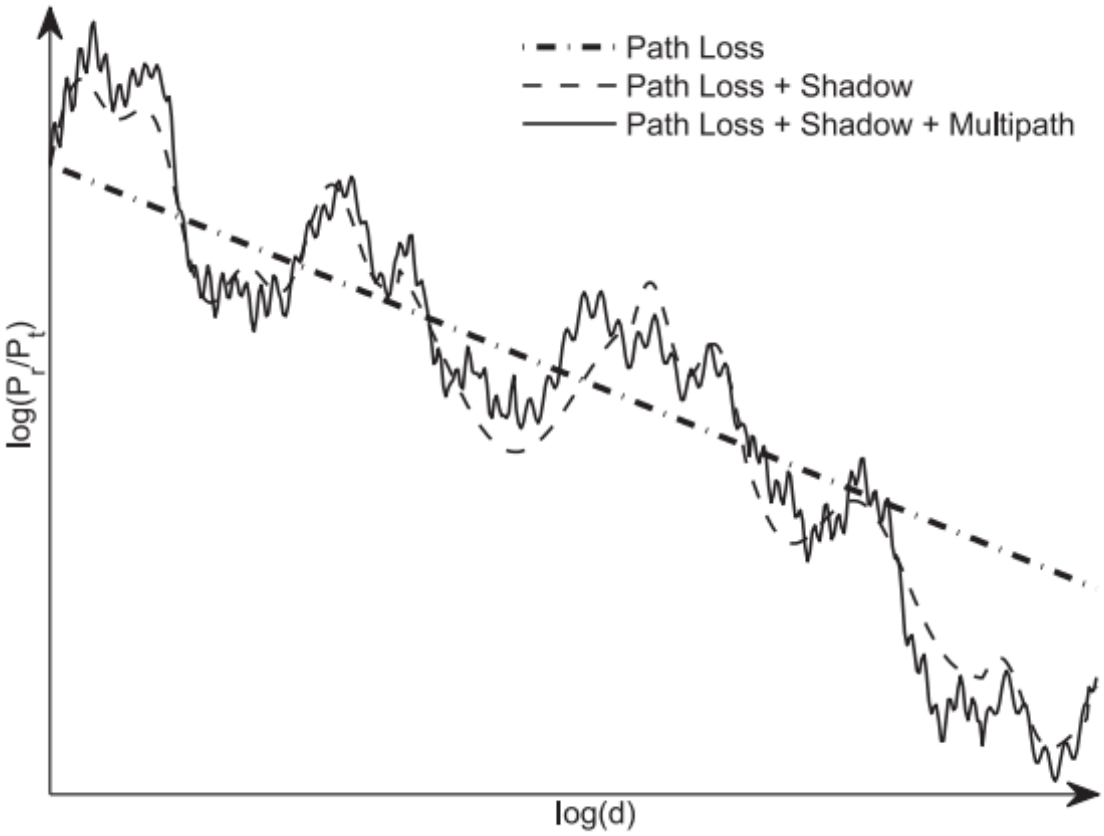

Figure 3.3: RSSI deterioration over distance and multipath induced fluctuations

The LDPL model characterizes the variation of received signal power over distance due to path loss and shadowing. Path loss stems from the dissipation of transmission power in the propagation channel, while shadowing results from the obstacles that attenuate signal power through absorption, reflection, scattering, and diffraction.

3.2 Fingerprinting Indoor Localization

The fingerprinting (FP) method is one of the most widely adopted approaches for indoor localization and serves as the foundation of this work. It relies on constructing unique signal 'fingerprints' at specific locations within the environment, which are subsequently exploited for position estimation. As illustrated in Fig. 3.4, the method operates in two phases. In the offline phase, signal measurements are systematically collected at reference points (RPs) to build a database of fingerprints. In the online phase, real-time measurements are compared against this database to infer the user's position.

Although FP is predominantly formulated as a classification problem, where real-time

measurements are matched to the closest reference point, it can also be extended to a regression framework. In this case, the model predicts continuous location coordinates, enabling estimation of positions not explicitly represented in the training set. Such an approach is particularly valuable when the deployment grid is sparsely populated with RPs due to cost, physical restrictions, or other practical constraints. While still categorized as a fingerprinting method, this formulation allows for greater generalization beyond discrete reference locations.

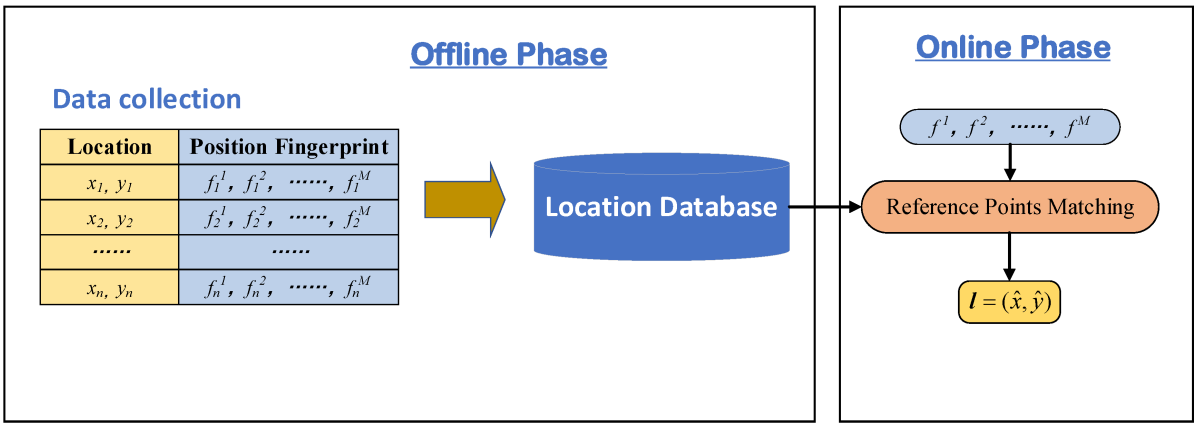

Figure 3.4: Illustration of the fingerprinting localization process

3.2.1 Offline Phase

The offline phase constitutes the foundation of the fingerprinting method, during which the system characterizes the radio environment of the indoor space. This phase begins with a site survey in which signal measurements are systematically collected across the target area. Reference points (RPs) are strategically selected, typically using a grid-based approach that partitions the area into regular sections, with the spacing between RPs determined by the desired accuracy and the complexity of the environment (Fig. 3.5).

At each reference point (RP) (x_i, y_i), multiple signal measurements are collected from all accessible access points (APs) or beacons within range. These measurements constitute the fingerprint of that location and are expressed as a feature vector

$$\mathbf{f_i} = (f_i^1, f_i^2, \ldots, f_i^M),$$

where M denotes the number of detectable transmitters (Tx). Unlike conventional fingerprinting methods that aggregate measurements from multiple APs, this work restricts evaluation to a single-AP scenario in order to examine the feasibility of accurate position estimation and to assess the limits of achievable accuracy under this constraint.

Indoor localization techniques rely on different categories of signal metrics, which can be broadly divided into signal-characteristic and geometric metrics. Signal-characteristic

metrics, such as the Received Signal Strength Indicator (RSSI) or Channel State Information (CSI), capture radio channel properties without requiring explicit geometric modeling. These metrics are widely used in fingerprinting-based approaches due to their ease of collection and compatibility with commodity hardware. In contrast, geometric metrics, including Angle of Arrival (AoA), Time of Arrival (ToA), and Round-Trip Time (RTT), exploit spatial propagation characteristics to estimate distances or angles relative to transmitters. While geometric metrics can achieve high accuracy, they typically require specialized hardware, precise synchronization, or antenna arrays, which limit their practicality in many deployments. In this study, we focus on CSI as the primary signal-characteristic metric, as it provides finer-grained channel information compared to RSSI and enables more robust fingerprinting in multipath-rich indoor environments.

To generate Fingerprints, we realize multiple measurements at each RP to account for wireless signal variability and to construct a database mapping location coordinates to their corresponding fingerprints, thereby enabling accurate retrieval during the online phase. However, the offline phase remains labor-intensive and sensitive to environmental dynamics, often necessitating recalibration when conditions such as furniture placement, wall modifications, or AP repositioning change. The experimental setup, measurement environment, data collection strategy, and database structure used in this work are detailed in subsequent sections.

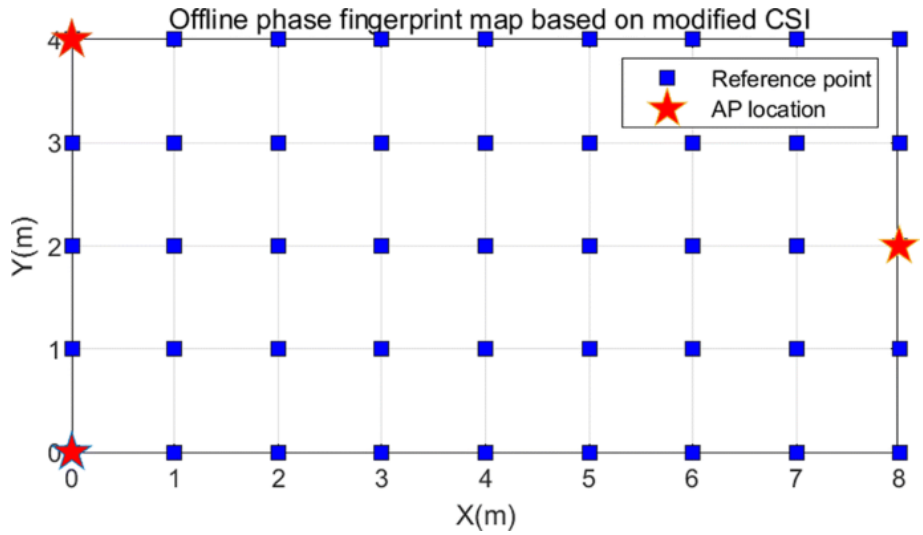

Figure 3.5: Example of the creation of a reference Grid in the offline phase (not our particular work)

3.2.2 Online Phase

The online phase utilizes the fingerprint database constructed during the offline phase to estimate the position of a target device. This phase begins when a mobile device measures

its current radio environment, producing a fingerprint

$$\mathbf{f_p} = (f_p^1, f_p^2, \ldots, f_p^M),$$

which contains the same signal characteristics recorded during calibration. Preprocessing operations such as signal filtering and normalization are applied to ensure consistency with offline data. To mitigate short-term variability, temporal smoothing techniques, such as averaging over multiple successive measurements, are often employed to reduce the impact of fast fading, noise, and transient interference.

The central task of the online phase is the matching process, in which the current fingerprint is compared with stored entries in the database to generate a location estimate. Classical approaches rely on similarity measures, including Euclidean distance, Manhattan distance, and cosine similarity, to identify the most likely reference points. Enhancements such as weighted distance functions and k-nearest neighbor (KNN) extensions have been proposed to improve robustness in multipath-rich environments. Despite their simplicity, these methods often struggle to generalize in dynamic environments and may suffer from degraded accuracy when the grid of reference points is sparse.

Recent advances have introduced data-driven approaches, particularly deep learning models, to replace hand-crafted similarity measures. Convolutional neural networks (CNNs) have demonstrated the ability to capture nonlinear mappings between signal features and spatial coordinates by directly learning discriminative representations from Channel State Information (CSI) and Received Signal Strength Indicator (RSSI) data. Compared to traditional matching, CNN-based methods exhibit greater adaptability to channel variations and improved localization accuracy in complex propagation environments. In this work, the matching stage is implemented using a CNN-based architecture, with the design, training procedure, and evaluation methodology presented in subsequent sections.

As illustrated in Fig. 3.4, the output of the online phase is a location estimate expressed as two-dimensional coordinates $\mathbf{l} = (\hat{x}, \hat{y})$ in the environmental coordinate system. Depending on the application, the system may also provide auxiliary information such as confidence indicators, error bounds, or uncertainty estimates, which are particularly important in safety-critical or quality-of-service-driven use cases.

As noted earlier, while fingerprinting is most commonly formulated as a classification problem, in which the current fingerprint is matched to the closest reference point, it can also be extended to a regression framework. In this formulation, the model predicts continuous coordinates, enabling the estimation of positions that do not coincide exactly with labeled reference points. Although the problem remains categorized as fingerprinting, the regression-based formulation enhances generalization and supports more flexible localization in real-world environments.

Fingerprinting remains one of the most widely deployed techniques for indoor localization due to its ability to exploit existing wireless infrastructure, its algorithmic maturity, and its proven effectiveness in real-world deployments. It also supports multi-signal fusion,

allowing heterogeneous sources such as WiFi, cellular, and inertial sensors to be integrated for improved robustness and accuracy.

Nevertheless, several challenges persist. The offline calibration phase is resource-intensive, requiring extensive measurement campaigns and periodic recalibration to maintain accuracy in the presence of environmental changes such as furniture rearrangement, wall modifications, or access point repositioning. Furthermore, the stochastic variability of radio signals, influenced by human mobility, interference, and multipath dynamics, continues to limit consistency. At scale, the computational requirements of online matching, even with efficient CNN architectures, remain a significant consideration for real-time deployment. These challenges underscore the need for more adaptive, resilient, and cost-efficient fingerprinting frameworks that balance accuracy, scalability, and practicality.

Chapter 4

Data Collection And Analysis

4.1 Data Collection

In the subsequent sections, we describe the dataset construction process based on Received Signal Strength Indicator (RSSI) and Channel State Information (CSI) measurements collected in the student laboratory located on the second floor of Building 9, Media Technology Department, TUM main campus. The architectural characteristics of the environment, the deployed equipment, and the adopted measurement procedures are outlined in detail.

4.1.1 Experimental Setup

The experiment was conducted in a student laboratory with the shape of an irregular quadrilateral containing one right angle, as illustrated in Fig. 4.3. The laboratory is furnished with clusters of tables, chairs, and computing equipment, resulting in a complex propagation environment with multiple obstacles and reflective surfaces. Consequently, the wireless channel exhibits strong multipath effects, including reflection, refraction, and scattering, which directly influence RSSI and CSI measurements.

A local coordinate system was established by aligning the x- and y-axes with the right-angled corner of the room, with the origin placed at this corner. Reference points (RPs) were distributed at one-meter intervals along both axes to form a measurement grid. Adjustments were made to exclude positions obstructed by furniture or equipment. The Wi-Fi access point (AP) was placed at an elevation of approximately 1.20 m, as shown in Figs. 4.1 and 4.2, to emulate a realistic deployment scenario and ensure adequate coverage across the measurement area.

Figure 4.1: Real layout of the lab

Figure 4.2: Access point position in the lab

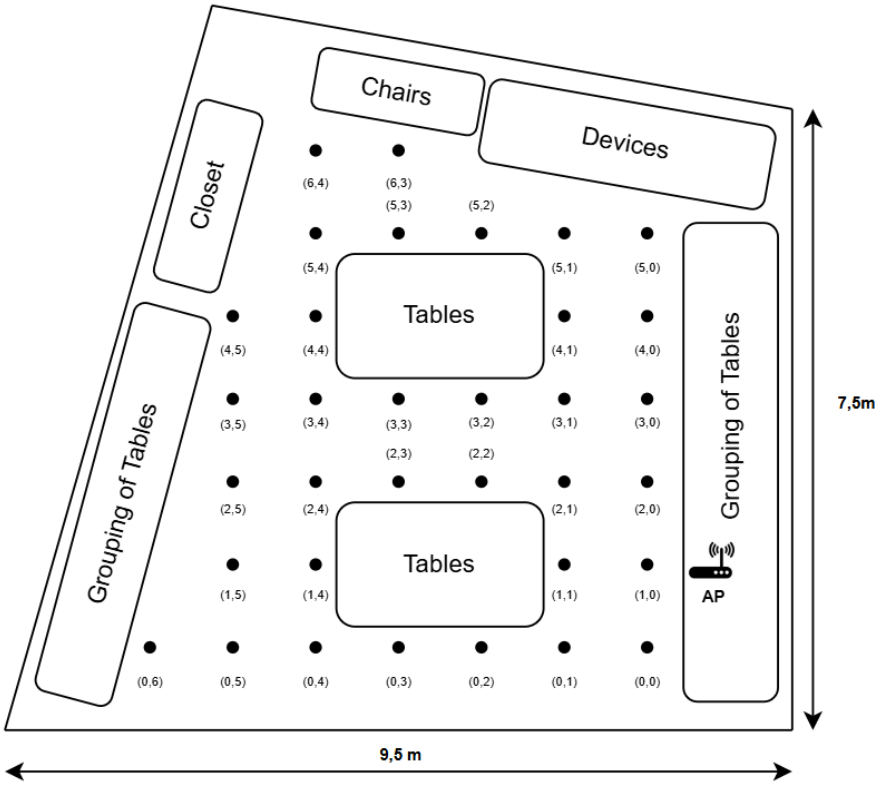

Figure 4.3: Reference grid layout of the lab

4.1.2 Software and Equipment

The experimental system consisted of a TP-LINK Archer A5 AC1200 router configured as the Wi-Fi access point and an ESP32 WROOM-32 microcontroller used as the receiver. The router transmitted packets over the 2.4 GHz band, while the ESP32 captured both RSSI and CSI features for each packet.

Initially, the study aimed to investigate per-antenna transmission behavior by treating each router antenna as an independent transmitter for multiple-input multiple-output (MIMO) channel analysis. However, this was not feasible with the off-the-shelf TP-LINK Archer A5. While the device operates as a dual-band router employing a 2×2 MIMO configuration with four external antennas and the chipset supports MU-MIMO and beamforming, phase and amplitude control is fully encapsulated within the MediaTek system-on-chip (SoC). As a result, the antennas are internally bound to their respective RF chains and cannot be individually accessed, configured, nor does the device provide raw per-antenna channel state information.

Consequently, the Archer A5 was treated as a conventional access point, and the study focused on aggregate link behavior rather than per-antenna channel characterization. This adaptation ensured compatibility with the ESP32 measurement framework and preserved experimental feasibility.

Figure 4.4: Transmitter: TP-LINK Archer A5 AC1200 access point

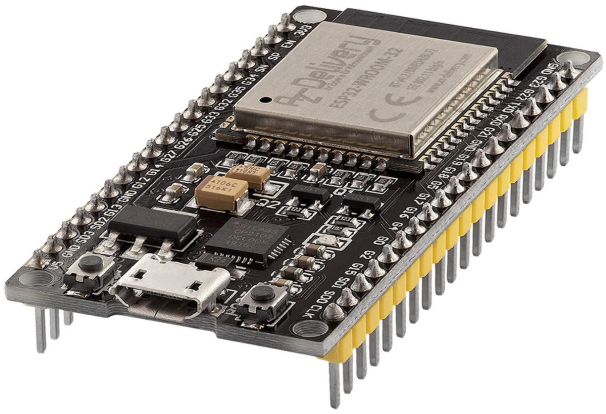

Figure 4.5: Receiver: ESP32 WROOM-32 microcontroller

The ESP32 WROOM-32 module was selected as the receiving device due to its integrated support for CSI extraction. It is a compact, low-power platform widely used in IoT applications, operating at 2.4 GHz under IEEE 802.11b/g/n standards. The module is powered through a 5V USB supply, with 3.3 V input/output pins and a minimum current requirement of 500 mA. Its integrated PCB antenna and compact dimensions ($56 \times 28 \times 13$ mm)

facilitated unobtrusive placement at reference points within the grid. For this study, the ESP32's ability to capture CSI across 64 subcarriers represented its key advantage, enabling fine-grained channel analysis beyond RSSI measurements alone.

4.1.3 Experimental Procedure

Data Collection Process:

The experimental setup was configured to enable reliable and fine-grained acquisition of Channel State Information (CSI) from the Wi-Fi link. Data transmission was conducted using IEEE 802.11n on the 2.4 GHz band, specifically on channel 2 with a 40 MHz bandwidth. This configuration provided a dense distribution of subcarriers, thereby enhancing channel resolution and improving the accuracy of multipath characterization. The access point (TP-Link Archer A5) was positioned at an elevation of 1.20 m within the laboratory to emulate a realistic deployment scenario and ensure uniform spatial coverage across the measurement grid.

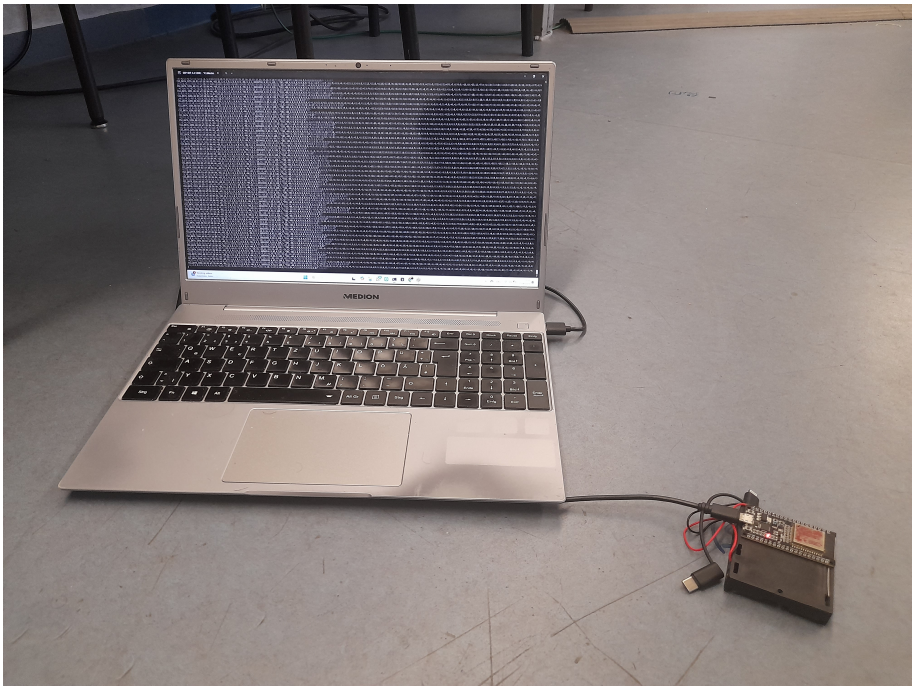

Figure 4.6: CSI data collection from a single reference point

On the receiver side, CSI extraction was enabled through custom ESP32 firmware [HB20]. Console output was redirected to a dedicated UART interface operating at a baud rate of 921600, which minimized data loss during high-throughput logging. Real-time handling was supported by FreeRTOS with a tick rate of 1000 Hz, yielding millisecond-level timestamp resolution for each CSI packet. The receiver (ESP32 WROOM-32) was connected via USB to a laptop (Medion, 2018, Intel-based, 8 GB RAM, 1 TB SSD, Windows 11),

where the high baud rate ensured efficient transfer and storage of large CSI volumes. This configuration ensured both temporal granularity and data integrity, providing a robust foundation for subsequent analysis.

After establishing these parameters, CSI measurements across the reference grid were conducted following a systematic procedure:

- **Firmware Initialization:** Prior to each measurement, the ESP32 was flashed with the appropriate firmware [HB20], ensuring that the receiver was consistently configured to capture CSI packets at each grid location.
- **Measurement Setup:** The ESP32 was connected to the laptop via a shielded USB extension cable to minimize interference from surrounding objects or the operator'This project aims to develop a comprehensive dataset for future indoor localization research by collecting both Channel State Information (CSI) and Received Signal Strength Indicator (RSSI) data within a controlled indoor environment. Using two ESP32 boards,one as a Wi-Fi transmitter and the other as a receiver. We will gather CSI and RSSI measurements from 50 predefined points on a spatial grid. At each location, variations in wireless signal behavior, including amplitude, phase, and signal strength, will be recorded, capturing the multipath effect caused by signal propagaation within the space.

  This dataset will provide valuable insight into how wireless signals travel and interact with indoor spaces, offering a rich foundation for training machine learning models to classify and localize objects or individuals. The combination of CSI and RSSI fingerprints at each grid point enhances the precision and robustness of localization models, addressing key challenges such as signal interference, multipath propagation, and environmental obstructions. Ultimately, this dataset will serve as a crucial resource for advancing the development of accurate and cost-effective indoor localization algorithms.

  Index Terms: Convolutional neural network (CNN),Channel State Information (CSI), ESP32, Indoor Localization, Fingerprinting.
- **Data Storage:** At each reference point, CSI packets were logged and stored in comma-separated values (.csv) format, facilitating structured post-processing. Each file contained detailed CSI information corresponding to a unique (x, y) coordinate in the grid.

In total, the raw dataset consisted of 39 files, each corresponding to a distinct reference point. At every point, 750 CSI samples were recorded, yielding a consistent dataset suitable for both statistical and machine learning-based analysis. The table below summarizes all the information provided per recorded raw sample.

The data acquisition tool employed in this work [HB20] outputs not only CSI values but also auxiliary metadata describing experimental parameters and static environmental properties. While many of these parameters (e.g., router configuration, sampling rate, antenna

identifiers) remain constant across the dataset and are not directly leveraged in our analysis, they are preserved in the raw files to ensure completeness and reproducibility. A detailed overview of these fields is provided in [LOL24].

Field	Description
Seq	The order of packets captured or transmitted.
Timestamp	The exact time when the packets was captured or received.
target_seq	The sequence number associated with the activity.
Target	The activity performed by the volunteer.
Mac	The MAC address of the Tx.
Rssi	RSSI represents the strength of the signal captured by the Rx.
Rate	The speed of communication link between devices, typically measured in Mbps (Megabits per second).

4.2 Data Analysis

Prior to designing and training an indoor localization model, it is essential to characterize the wireless channel, as the performance of fingerprinting approaches depends critically on the richness, stability, and distinctiveness of extracted features. In particular, the effectiveness of Received Signal Strength Indicator (RSSI)-based fingerprints and, more importantly, Channel State Information (CSI)-based fingerprints is determined by the degree to which these metrics capture spatial and frequency-selective channel properties.

This section presents a characterization of the wireless channel in the laboratory environment. The analysis begins with large-scale RSSI spatial distributions, proceeds to fine-grained CSI subcarrier variability, and concludes with an interpretation of how these characteristics influence localization feasibility and system design.

4.2.1 RSSI Spatial Distribution Analysis

RSSI is a widely employed yet inherently coarse metric in indoor localization. Its value attenuates with distance but is highly sensitive to multipath, obstacles, and shadowing. Consequently, spatial distribution analysis of RSSI provides qualitative insight into the feasibility of fingerprinting while highlighting localized anomalies due to propagation effects.

From the grid measurements (Fig. 4.7):

- **RSSI range:** -17 dB (from -64 dBm to -47 dBm).
- **Strongest signals:** Coordinates $(1, 0)$ and $(2, 0)$, with values near -48 to -49 dBm.
- **Weakest signals:** Coordinates $(4, 5)$ and $(5, 0)$, with values down to -63 and -64 dBm.

Field	Description
sig_mode	The modulation and encoding scheme used for transmitting data.
Mcs	The Modulation and Coding Scheme (MCS) specifies the combination of modulation and error correction coding used in wireless communication.
Cwb	The channel bandwidth or the range of frequencies allocated for communication.
smoothing	This field indicates whether smoothing is applied during transmission.
not_sounding	This field indicates whether a sounding frame is being used for channel estimation and feedback.
aggregation	This field indicates whether multiple data frames are combined into a single transmission unit.
stbc	This field indicates whether Space-Time Block Coding (STBC) is employed to improve the reliability of wireless communication by transmitting redundant data across multiple antennas.
fec_coding	The type of Forward Error Correction (FEC) coding used, if any.
sgi	This field indicates whether Short Guard Interval (SGI) is enabled to reduce the guard interval between symbols in wireless communication.
noise_floor	The level of background noise in the communication channel
ampdu_cnt	The number of A-MPDU (Aggregate MAC Protocol Data Unit) transmitted or received.
channel_primary	The specific frequency band used for transmission
channel_secondary	The secondary channel used to increase data transmission rate, if applicable.
local_timestamp	The timestamp relative to the local time of the device capturing the packet.
ant	The antenna used for transmitting or receiving the packet.
sig_len	The size of the signal or frame in bytes.
rx_state	The operational state of the Rx.
len	The length of the packet in bytes.
first_word	The information about the first word or header of the packet.
data	The actual payload or content of the packet. In this context, data refers to the CSI subcarriers captured.

The bottom-left region of the grid exhibits higher signal strength, consistent with the AP's location. Signal decay toward the top and right of the grid reflects natural path loss. However, deviations from ideal radial contours are evident: for example, at $(2, 4)$, the RSSI drops sharply to -61 dBm despite moderate distance from the AP. This anomaly is attributable to destructive interference and obstruction, as the point lies in a cluttered area with furniture and computing equipment.

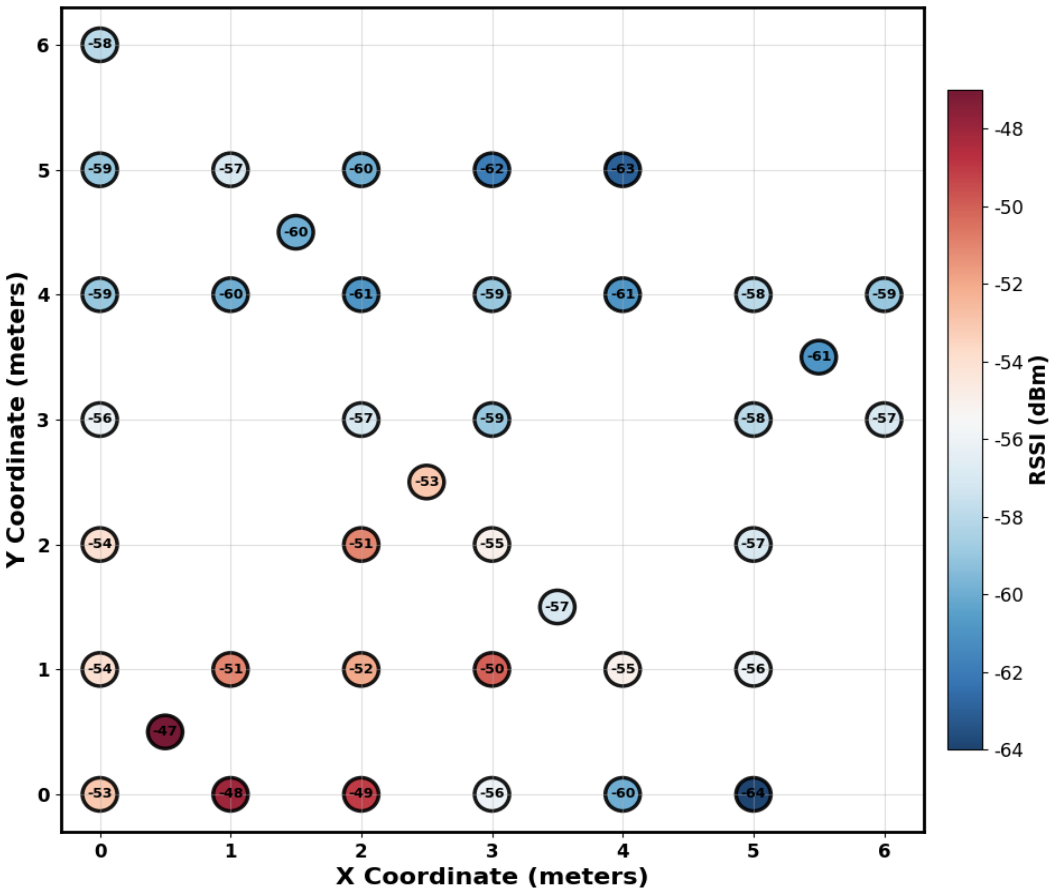

Figure 4.7: Measured RSSI spatial distribution across the laboratory grid

Overall, the RSSI field exhibits both a global decay trend and localized irregularities. This combination provides discriminative gradients for coarse localization while simultaneously exposing limitations due to multipath variability. Such behavior highlights the need for fine-grained features such as CSI to achieve higher robustness and accuracy.

4.2.2 CSI Amplitude Variation Across Subcarriers

In multipath-rich environments, the received signal is a superposition of the direct line-of-sight (LOS) component and multiple reflected, diffracted, or scattered paths. This induces frequency selectivity, whereby different subcarriers in an OFDM system experience distinct fading conditions. As a result, CSI reveals amplitude and phase variations across subcarriers that cannot be captured by RSSI alone.

Figures 4.8 and 4.9 illustrate measurements obtained in an unobstructed corridor environment dominated by LOS propagation. In this case, amplitude variation across subcarriers was limited, with smooth spectral behavior and high correlation among adjacent subcarriers, consistent with channels exhibiting minimal multipath.

By contrast, the present dataset collected in the student laboratory (Fig. 4.10) exhibits pronounced amplitude fluctuations across subcarriers. Although some degree of correlation among neighboring subcarriers remains, the overall spectral response is highly irregular, reflecting strong multipath and scattering effects introduced by furniture, electronic equipment, and the non-rectangular room geometry.

Figure 4.8: Corridor in a minimal multipath scenario

Interpretation

The combined evidence from RSSI and CSI analysis classifies the laboratory as a **rich multipath environment**, with the following defining features:

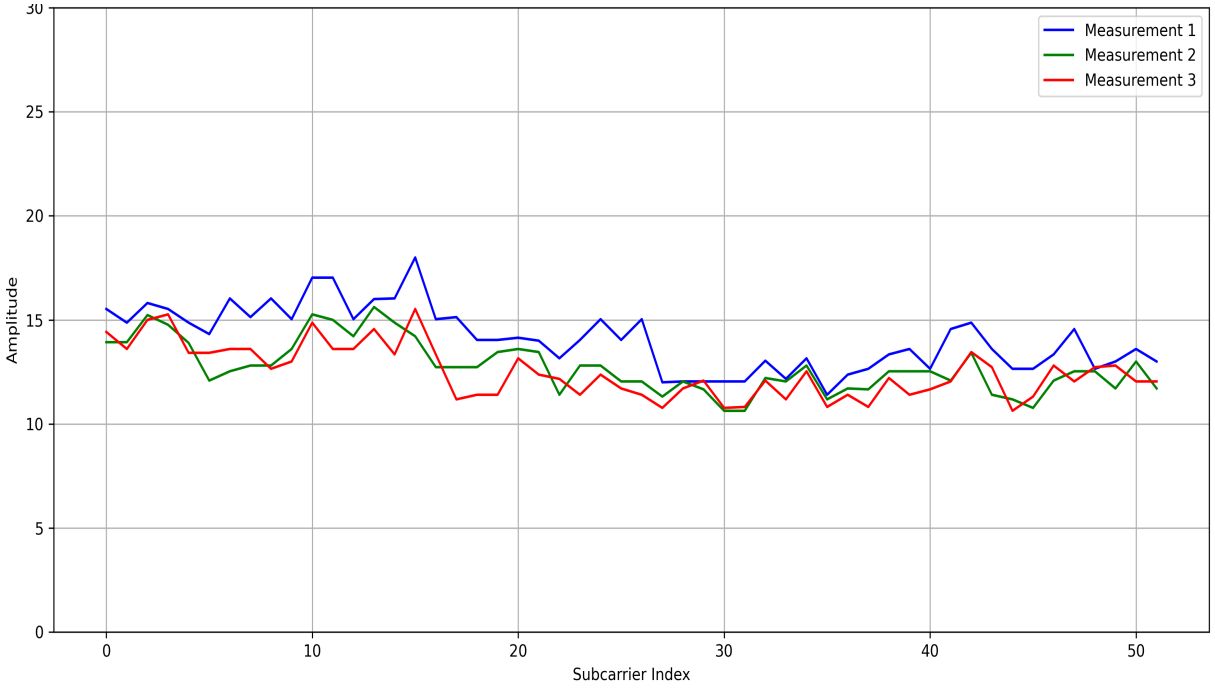

Figure 4.9: Subcarrier amplitude distribution in LOS-dominated corridor environment

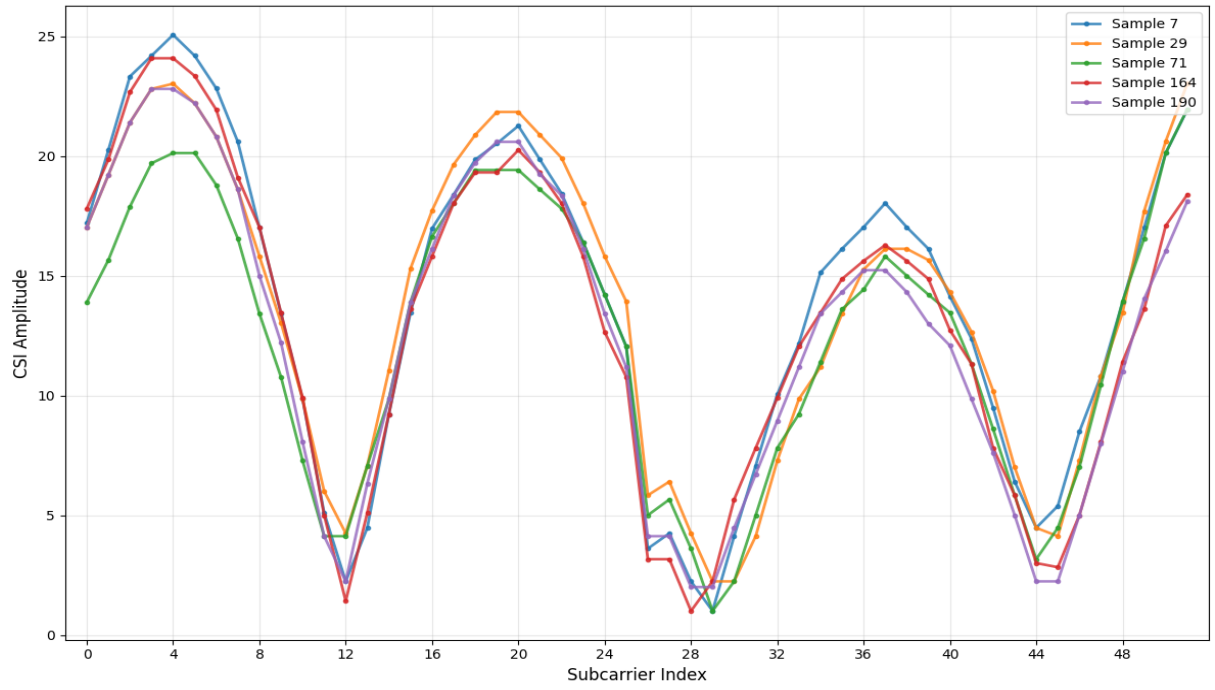

Figure 4.10: Subcarrier amplitude variation in multipath-rich laboratory environment

1. Strong scattering and appreciable delay spread, indicative of multiple propagation paths.

2. Dominance of non-line-of-sight propagation, leading to irregular amplitude fluctuations across subcarriers.

These characteristics imply that while RSSI can provide coarse localization capability, robust fingerprinting in this environment requires CSI-based methods capable of exploiting fine-grained frequency-selective channel variations.

Chapter 5

Indoor Localization Methodology

5.1 Data Processing and Preprocessing

CSI Data Preprocessing

Prior to model development, the raw Channel State Information (CSI) data was processed to retain only location-relevant features. Many recorded fields, including `rate`, `sig-mode`, and `aggregation`, describe static properties of the wireless link or device configuration and do not vary across spatial locations. For instance, the modulation and coding scheme (MCS) and channel bandwidth (CBW) are fixed by the device settings and carry no information about the propagation environment. Similarly, error-correction or smoothing parameters reflect implementation details of the communication system rather than spatial channel characteristics. Inclusion of these features would add redundancy without improving the discriminative power of the dataset.

Timestamp fields, both global and local, were also excluded. While temporal information is critical in dynamic applications such as human activity recognition, the present study involves a static grid of measurement points. Consequently, the sequence or absolute timing of packets does not contribute meaningful localization information. By restricting the feature set to CSI subcarriers and RSSI, the dataset remains focused on spatially informative attributes.

After isolating the relevant 128-element CSI array and RSSI values, additional preprocessing steps were applied to remove non-informative subcarriers:

- **Metadata subcarriers:** The first two subcarriers typically encode system information such as channel identifiers or device IDs. Discarding them removes four numerical values (real and imaginary components) that do not vary with location.
- **Guard-band subcarriers:** Subcarriers at the edges of the array serve as guard bands to mitigate inter-channel interference. As these carry no transmitted informa-

tion, they were excluded.

- **DC subcarrier:** The central subcarrier at zero frequency is conventionally set to zero and does not encode propagation information; it was therefore discarded.

Following this reduction, 104 values remained, corresponding to the real and imaginary components of 52 effective subcarriers. To facilitate analysis, complex CSI values were converted into amplitude and phase components using standard trigonometric relations:

$$|H_k| = \sqrt{\Re(H_k)^2 + \Im(H_k)^2}, \quad \angle H_k = \arctan\left(\frac{\Im(H_k)}{\Re(H_k)}\right), \tag{5.1}$$

where H_k denotes the complex CSI value of the k-th subcarrier.

The CSI magnitude (amplitude) was retained without modification as the primary feature for model training, reflecting the frequency-dependent attenuation across the channel. Phase measurements were sanitized using the S-Phaser Interpolation Elimination Method (IEM) [HLM+19] to remove hardware- and sampling-induced linear offsets:

$$\phi_k = \tilde{\phi}_k - (aK_k + b), \tag{5.2}$$

where $\tilde{\phi}_k$ is the raw measured phase, K_k is the subcarrier index re-centered around the DC subcarrier ($-26, \ldots, -1$ and $1, \ldots, 26$), and a, b are least-squares fitted constants.

To further stabilize the feature space, both amplitude and calibrated phase sequences were lightly smoothed using moving averages across consecutive packets. Additionally, outlier rejection was applied to discard measurements with abnormally high variance across subcarriers. These steps reduce short-term noise while preserving location-dependent variations, enhancing model convergence and robustness, particularly under limited training data conditions.

5.2 Dataset Construction

For the regression-based localization task, the dataset was partitioned into training, validation, and testing subsets. The 39 reference points were divided such that 5 grid points were reserved for testing, strategically distributed to ensure spatial coverage and unbiased evaluation. The remaining 34 grid points were split approximately 80:20 into 27 training points and 7 validation points, with the latter used for hyperparameter optimization and model selection (Fig. 5.1).

Formally, the partitioning can be expressed as:

$$N_{\text{total}} = 39, \quad N_{\text{train}} = 27, \quad N_{\text{val}} = 7, \quad N_{\text{test}} = 5.$$

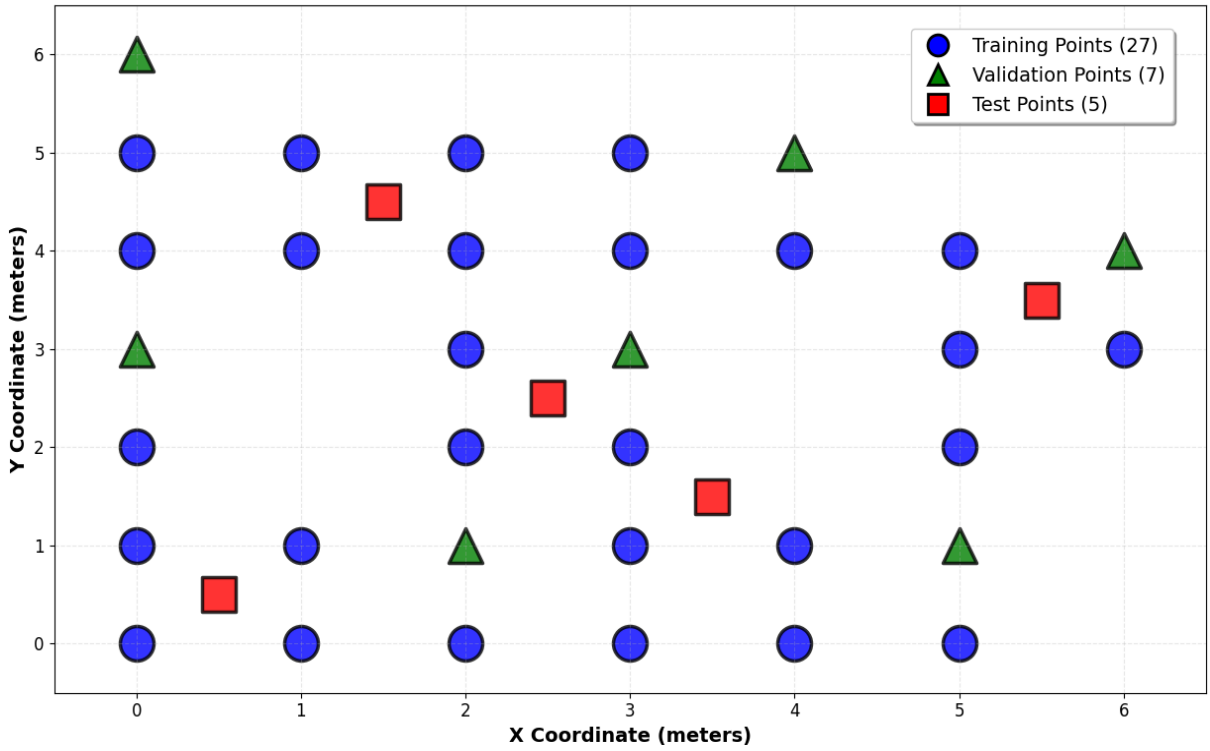

Figure 5.1: Partitioning of grid points for training, validation, and testing

After processing, at each grid point, at least 750 samples were available to capture signal variability and measurement noise. To evaluate the impact of sample size on model performance, the dataset was also organized into subsets containing 250, 500, and 750 samples per point, enabling systematic analysis of model accuracy, stability, and generalization across different data volumes.

The laboratory environment exhibits rich multipath propagation, moderate spatial diversity, and statistically stable CSI features over time. These characteristics justify the adoption of deep learning techniques, as conventional regression or interpolation methods may fail to capture subtle yet distinctive spatial variations across the grid.

5.3 Proposed Indoor Localization Solutions

Prior work has demonstrated the effectiveness of convolutional neural networks (CNNs) for CSI-based indoor localization. For example, [HCN19] trained a deep neural network on Wi-Fi RSS and CSI fingerprints to estimate user positions in residential environments, while [PH20] proposed a lightweight one-dimensional CNN that achieved good accuracy with reduced computational complexity. Similarly, [SRM22] and [SCP21] represented CSI measurements from multiple transmitters as input matrices, allowing CNNs to learn spatial mappings directly to device coordinates. These studies confirm the potential of CNN-driven

CSI localization, outperforming traditional fingerprinting approaches and demonstrating robustness against multipath and interference. However, most existing solutions assume access to multiple-input multiple-output (MIMO) systems and rely on several transmitters deployed across the environment. In contrast, this work addresses the more constrained setting of achieving reliable accuracy using a single Wi-Fi access point and a low-cost ESP32 receiver with a single antenna. To this end, we evaluate and compare multiple CNN-based architectures to identify the most effective model for this deployment scenario.

Architectural Components of Convolutional Neural Networks

We opted for Convolutional Neural Networks (CNNs) this indoor localization task, as CNNs leverage convolutional operations to capture local correlations and hierarchical feature representations, enabling effective extraction of both spatial and frequency patterns from CSI and RSSI measurements. Through the application of multiple convolutional and pooling layers, the network learns discriminative features that encode location-specific channel fingerprints, even in environments dominated by multipath propagation and non-line-of-sight conditions. This hierarchical feature extraction enhances the model ability to generalize across spatially proximate locations while preserving high localization accuracy. In the following sections, we present the theoretical foundations of CNNs, detailing the structure and function of their constituent layers and their role in feature learning.

A Convolutional Neural Network (CNN) is a deep learning architecture designed to process structured data such as images, time-series, or spatial matrices. In the context of indoor localization, CNNs are frequently employed to learn spatial and frequency-domain features from Channel State Information (CSI) matrices or Received Signal Strength Indicator (RSSI) maps. A typical CNN consists of two major parts: (i) feature extraction layers, which include convolutional, activation, and pooling layers, and (ii) classification or in our case regression layers, usually composed of fully connected and output layers. Figure 5.2 illustrates this structure.

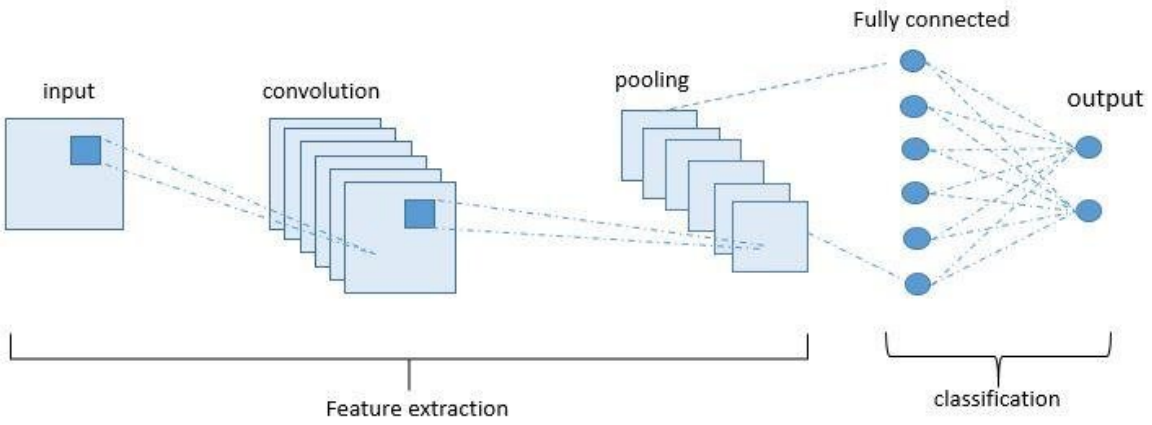

Figure 5.2: The core architectural components of a CNN.

Convolutional Layer

The convolutional layer constitutes the fundamental computational unit of a Convolutional Neural Network (CNN). Unlike fully connected layers, which perform dense matrix multiplications over the entire input, convolutional layers employ local connectivity and parameter sharing to capture spatially localized features. Specifically, a small learnable kernel (or filter) slides over localized regions of the input, applying an element-wise multiplication followed by summation. This mechanism reduces the number of trainable parameters while enhancing the ability to capture hierarchical feature structures.

In the context of indoor localization, the input may take the form of a Channel State Information (CSI) matrix, an RSSI spatial heatmap, or other signal-derived representations. The convolutional kernels operate as feature extractors that learn to detect spatial and frequency-domain correlations. For instance, in CSI data, kernels may emphasize frequency-selective fading patterns that arise due to multipath propagation, while in RSSI heatmaps, they capture localized attenuation gradients associated with transmitter proximity and obstacles.

Mathematically, the two-dimensional discrete convolution of an input matrix X with a kernel A is defined as

$$(X * A)(m, n) = \sum_i \sum_j X(i, j)\, A(m - i, n - j), \tag{5.3}$$

where (m, n) denotes the output feature map coordinates, $X(i, j)$ represents the input values, and A is the kernel. By sliding the kernel across the input, the operation generates a feature map that highlights distinctive local patterns such as sharp amplitude transitions or frequency-domain distortions.

An important property of convolutional layers is *translation invariance*, meaning that features detected in one region of the input are equally recognizable in another. This is particularly valuable for localization tasks, where similar multipath effects may occur at different spatial locations within the environment. Additionally, deeper convolutional stacks progressively capture higher-level abstractions: from low-level frequency shifts in the first layers to more complex propagation signatures in deeper layers.

Activation Functions

To introduce non-linearity into the network, activation functions are applied after convolutional or fully connected layers. This step is essential, as linear operations alone cannot approximate complex nonlinear mappings such as those encountered in indoor localization. The most widely adopted function is the Rectified Linear Unit (ReLU), defined as

$$\mathrm{ReLU}(z) = \begin{cases} z, & z \geq 0, \\ 0, & z < 0, \end{cases} \tag{5.4}$$

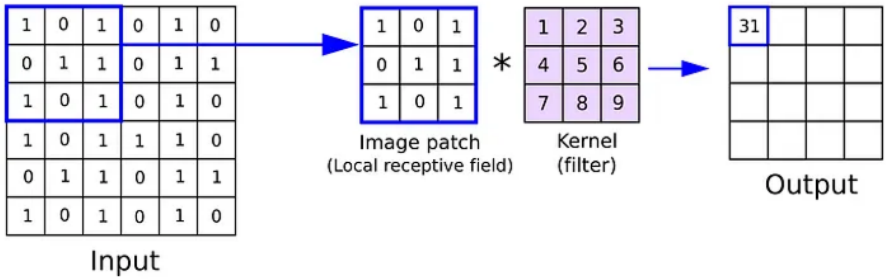

Figure 5.3: Illustration of element-wise multiplication of a kernel and an input patch during convolution.

where z denotes the pre-activation value. ReLU enforces sparsity by suppressing negative activations while retaining positive responses, thereby emphasizing salient features and improving computational efficiency due to its simplicity.

Despite its advantages, ReLU is prone to the "dying ReLU" problem, where neurons become inactive if their outputs remain consistently negative. Variants such as *Leaky ReLU* introduce a small slope for negative values to maintain gradient flow, while *Parametric ReLU* (PReLU) learns this slope as a parameter. For tasks requiring bounded activations, functions such as the *Sigmoid* and *Hyperbolic Tangent (Tanh)* are used, although they are more susceptible to vanishing gradients. More recently, advanced functions such as the *Gaussian Error Linear Unit (GELU)* and *Swish* have demonstrated improved performance by combining smoothness with nonlinearity, enabling richer gradient propagation.

In the context of fingerprinting-based indoor localization, the choice of activation function can directly affect the regression accuracy of position estimates. Functions like ReLU and its variants are particularly suitable, as they allow the model to capture dominant propagation features while maintaining robustness to local fluctuations and multipath-induced distortions. For final regression outputs, however, linear activations are often employed to preserve the unbounded nature of spatial coordinates.

Pooling Layer

Although convolutional layers are effective in extracting rich spatial and spectral feature maps, they remain sensitive to minor spatial shifts, local noise, or abrupt fluctuations in the input. Pooling layers mitigate these sensitivities by performing a controlled downsampling of feature maps. By aggregating local neighborhoods into a single representative value, pooling provides a degree of translational invariance and reduces the dimensionality of the feature space, thereby lowering computational complexity and mitigating overfitting.

The most widely used variant is *max pooling*, which preserves the most dominant activation within a local region. Formally, given a sliding window of size $m \times n$ with stride s, max

pooling computes

$$y = \max\{x_{ij}\}, \quad (i,j) \in \text{window}, \tag{5.5}$$

where x_{ij} denotes the activation value at position (i, j). This operation retains the strongest local response while discarding weaker activations, effectively emphasizing salient features such as edges or peaks in the learned representation.

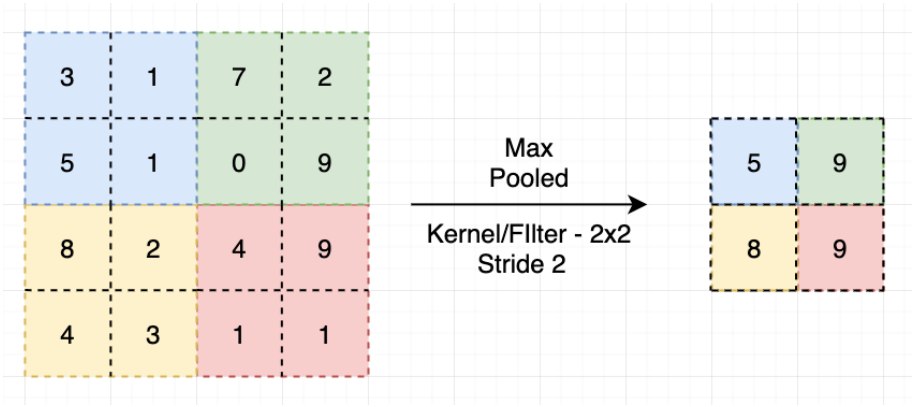

Figure 5.4: Illustration of element down-sampling in Max pooling

Alternative strategies include average pooling, which computes the mean of each window, and global average pooling (GAP), which condenses an entire feature map into a single representative value. GAP is increasingly used in place of fully connected layers in modern CNNs, as it reduces overfitting.

Fully Connected Layer

Following feature extraction through convolutional and pooling layers, the resulting multi-dimensional feature maps are typically transformed into a one-dimensional vector representation via a flattening operation. This vector is then processed by one or more fully connected (dense) layers, in which each neuron is connected to all activations from the preceding layer. Such dense connectivity allows the network to integrate and combine low-level features into higher-level abstractions, ultimately supporting robust decision-making or regression outputs.

Formally, the operation of a fully connected layer is expressed as

$$z = W \cdot x + b, \tag{5.6}$$

where x is the input feature vector, W denotes the weight matrix, and b represents the bias term. The output z is then typically passed through a non-linear activation function, enabling the model to capture complex feature interactions.

In the context of indoor localization, fully connected layers serve to integrate spatial and frequency-domain patterns extracted by earlier convolutional layers. For instance,

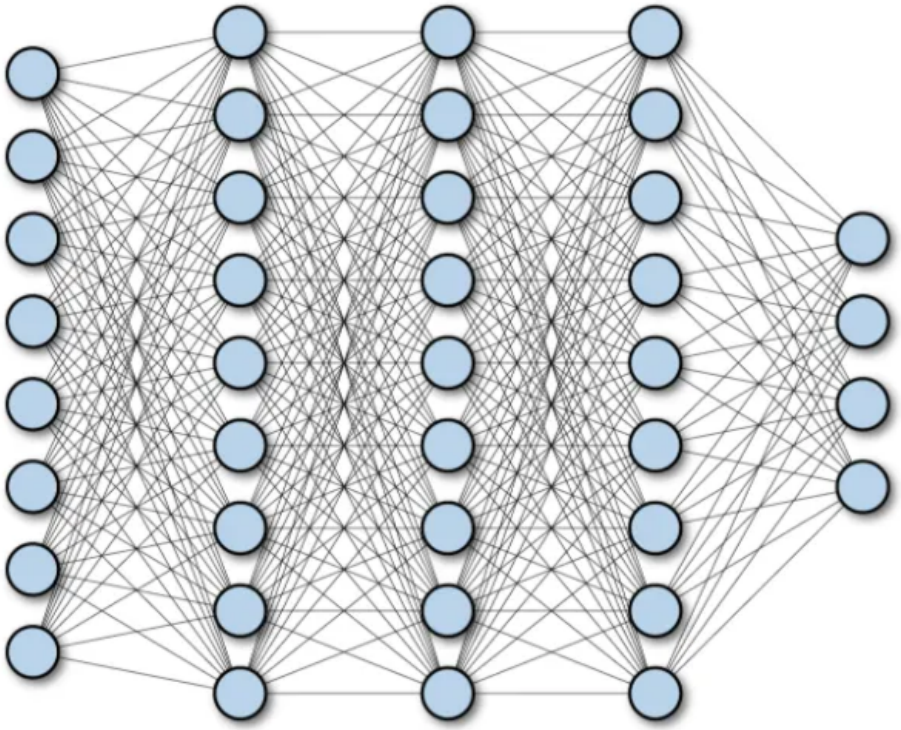

Figure 5.5: Illustration of fully connected layers

frequency-selective fading characteristics identified at the subcarrier level can be combined with spatial attenuation gradients, producing compact feature embeddings that are directly predictive of user location. When formulated as a regression task, the final dense layer often maps these embeddings into two-dimensional coordinate estimates $(\hat{x}, \hat{y})$ within the reference environment.

Output Layer

The design of the output layer is directly determined by the chosen formulation of the indoor localization problem. When localization is framed as a discrete multi-class classification task, the goal is to identify the most likely reference point (RP) from a finite set of candidates in the fingerprint database. In this case, the softmax activation function is commonly employed to map the network outputs into a probability distribution over K discrete classes:

$$\sigma(\vec{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}, \quad i = 1, \ldots, K, \tag{5.7}$$

where $\vec{z}$ denotes the vector of raw outputs (logits), and K corresponds to the number of reference points. The predicted class is then taken as the index with the maximum probability, while the associated probability values provide a natural confidence measure. This formulation is consistent with the traditional fingerprinting paradigm, in which each RP corresponds to a labeled fingerprint in the offline database.

Alternatively, fingerprinting can be formulated as a regression problem, where the objective is to predict continuous two-dimensional coordinates $\mathbf{l} = (\hat{x}, \hat{y})$ directly. This approach bypasses the need to map measurements to discrete RPs and instead leverages the spatial continuity of signal features to infer positions in previously unseen locations. A linear output layer without activation is typically adopted in this case, preserving the unbounded numerical range required for coordinate estimation. This regression-based perspective is particularly useful in scenarios where the measurement grid is sparse, or where cost and environmental constraints prevent dense reference point deployment.

Both formulations have distinct advantages. Classification-based approaches benefit from robustness to noise and straightforward interpretability, as the output directly maps to known reference points. Regression-based approaches, on the other hand, offer finer granularity and the ability to interpolate between reference points, which is especially valuable in multipath-rich environments where continuous spatial variation in channel characteristics can be exploited. In practice, hybrid strategies such as coarse classification followed by regression refinement are increasingly explored to balance robustness with accuracy in fingerprinting-based indoor localization.

Additional Considerations

Modern CNN architectures often incorporate regularization and optimization strategies, In our work in particular we used strategies such as:

- **Dropout**: Randomly disabling neurons during training to prevent overfitting.
- **Batch Normalization**: Normalizing activations to stabilize learning and accelerate convergence.
- **Advanced Optimizers**: Using methods such as Adam or RMSProp instead of plain stochastic gradient descent.

These additions are particularly relevant in indoor localization, where CSI and RSSI data exhibit high dimensionality and variability due to multipath fading and environmental dynamics.

Neural Network Architectures for CSI-Based Localization

In this work, we formulate the indoor localization problem using WiFi Channel State Information (CSI) and received signal strength indicator (RSSI) as a regression task rather than as a classification problem. Instead of predicting the discrete index of a reference point (RP), the objective is to directly estimate the continuous two-dimensional coordinates (x, y) in meters within the accessible area of a 9.5×7.5 m indoor laboratory environment. To this end, multiple learning pipelines were proposed and systematically compared.

In all models, CSI amplitudes were complemented by sanitized and calibrated phases to ensure that both magnitude and phase carried meaningful spectral information. Model

training was performed using the Adam optimizer with early stopping based on validation error to prevent overfitting. Specifically, training was terminated if the validation loss failed to improve for 15 consecutive epochs.

The above configuration represents the initial training setup and baseline set of parameters adopted in this study. Based on the empirical results and observed model performance, several modifications and refinements were subsequently introduced to improve localization accuracy and stability. These adjustments are detailed in the performance evaluation section, where their impact is quantitatively assessed.

5.3.1 Baseline Convolutional Neural Network (CNN)

Input Structure

As proposed in the dataset formation section, we deployed a single transmitter in a static environment and collected CSI fingerprints at 39 reference points distributed across the measurement area. 27 of them were deployed for training. For each RP, we obtained 52 subcarrier responses, consisting of both amplitude and phase information. Thus, every training instance was represented as a matrix of dimension 52×2, where one channel captured amplitude and the other captured phase. Each fingerprint $\mathbf{X}$ was paired with $\mathbf{y}$ the corresponding ground-truth spatial coordinates, yielding supervised training pairs of the form

$$\big(\mathbf{X}, (x, y)\big). \tag{5.8}$$

where $\mathbf{X} \in \mathbb{R}^{52\times 2}$ and $\mathbf{y} = (x, y) \in \mathbb{R}^2$.

Input remark For the baseline CNN we intentionally retain both CSI amplitude and CSI phase for every sample. The two-channel input (amplitude, phase) forms the feature channels for 1-D convolution along the subcarrier axis. Prior to training each channel is standardized (per-subcarrier zero-mean, unit-variance across the training set) to ensure stable optimization and comparable dynamic ranges between amplitude and phase.

Layer	Type	Filter / Kernel	Activation	Output Shape	Parameters
0	Input	–	–	(52, 2)	0
1	Conv1D	32 filters, kernel=5, padding=same	ReLU	(52, 32)	352
2	BatchNormalization	–	–	(52, 32)	128
3	MaxPooling1D	pool size=2	–	(26, 32)	0
4	Dropout	rate=0.2	–	(26, 32)	0
5	Conv1D	64 filters, kernel=3, padding=same	ReLU	(26, 64)	6,208
6	BatchNormalization	–	–	(26, 64)	256
7	MaxPooling1D	pool size=2	–	(13, 64)	0
8	Dropout	rate=0.2	–	(13, 64)	0
9	GlobalAveragePooling1D	–	–	(64)	0
10	Dense	128 units	ReLU	(128)	8,320
11	Dropout	rate=0.3	–	(128)	0
12	Dense	64 units	ReLU	(64)	8,256
13	Dropout	rate=0.2	–	(64)	0
14	Dense	2 units	Linear	(2)	130

Table 5.1: Layer configuration of the proposed baseline CNN

Network Architecture

We designed a one-dimensional convolutional neural network (CNN) to learn the mapping between CSI fingerprints and physical coordinates. The input representation was reshaped

so that the subcarrier index formed the sequential axis for convolution, while amplitude and phase served as complementary feature channels. The rationale for this design was that multipath propagation and reflections manifest as correlated variations across adjacent subcarriers, which can be effectively modeled by convolutional filters.

The first convolutional layer employed 32 filters with a kernel size of 5, enabling the network to capture medium-range spectral dependencies over neighborhoods of adjacent subcarriers. This stage was intended to extract smooth variations and correlated fading patterns that characterize the environment. Batch normalization, max pooling with a pool size of 2, and a dropout rate of 0.2 were applied to stabilize training and mitigate overfitting.

A second convolutional layer with 64 filters and a kernel size of 3 was then introduced to learn finer-grained spectral patterns, such as sharp notches or localized phase transitions, which provide additional discriminatory power. Again, as shown in Table 5.1, normalization, pooling, and dropout were applied to preserve robustness.

Following convolutional feature extraction, we employed a Global Average Pooling (GAP) operation to compress each feature map into a single representative value. The use of GAP, as opposed to a fully connected flattening layer, significantly reduced the number of parameters, which was critical given the relatively small number of reference points. Moreover, this design encouraged the network to focus on the overall spectral signature rather than memorizing specific subcarrier responses.

The pooled features were subsequently processed by two fully connected layers with 128 and 64 neurons, respectively, each using ReLU activations and accompanied by dropout (0.3 and 0.2). These dense layers introduced nonlinear transformations that allowed the model to capture the complex, non-linear mapping between CSI-derived features and physical coordinates. Finally, a linear output layer with two units provided the predicted coordinates $(\hat{x}, \hat{y})$.

Training Objective

The baseline CNN is trained as a regression model with mean squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \|(x_i, y_i) - (\hat{x}_i, \hat{y}_i)\|_2^2, \tag{5.9}$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote ground truth and predicted coordinates, respectively.

Summary

The baseline CNN learns to transform CSI fingerprints into high-level spectral features through hierarchical convolutional layers, compresses these features using global pooling, and maps them to continuous two-dimensional coordinates through dense layers. By explicitly formulating localization as a regression problem and optimizing for Euclidean distance

in physical space, the model leverages spectral signatures of the multipath channel to provide fine-grained position estimates with strong generalization capabilities.

5.3.2 Hybrid Convolutional Neural Network with RSSI Integration (H-CNN)

Input Structure

The hybrid model omits phase measurements, which are more susceptible to noise and hardware-induced corruption, and instead exploits two complementary signal modalities collected at each reference point (RP): (i) the CSI amplitude across 52 subcarriers, and (ii) a scalar RSSI value corresponding to the same Wi-Fi packet. The CSI amplitude encodes fine-grained frequency-selective fading and multipath structure, capturing the localized variations of the wireless channel, whereas the RSSI provides coarse-grained range information and aggregate path loss, effectively complementing the CSI features. Prior to model training, the CSI amplitudes are standardized on a per-subcarrier basis to zero-mean and unit-variance across the training set, and the RSSI is similarly normalized to the same scale. This preprocessing ensures numerical consistency between the two modalities, facilitating their effective fusion within the learning pipeline.

$$\mathbf{X}_{\text{CSI}} \in \mathbb{R}^{52\times 1}, \qquad \mathbf{X}_{\text{RSSI}} \in \mathbb{R}.$$

Supervised training pairs take the form

$$\big((\mathbf{X}_{\text{CSI}}, \mathbf{X}_{\text{RSSI}}),\ (x, y) \big), \tag{5.10}$$

where $(x, y) \in \mathbb{R}^2$ denotes the ground-truth coordinates associated with the fingerprint.

Network Architecture

We implement a dual-branch neural architecture that processes CSI and RSSI in specialized pathways and fuses their representations for regression to Cartesian coordinates.

The design rationale mirrors the baseline CNN: convolutions model correlated frequency-domain structure (multipath signatures), while an auxiliary branch encodes coarse distance information provided by RSSI.

CSI branch (multi-scale convolution):

The CSI branch employs a multi-scale convolutional block consisting of two parallel paths that operate on the amplitude-only CSI vector. The first path uses 32 filters with kernel size 3 to capture local spectral variations (e.g., frequency-selective fading and narrowband notches). The second path uses 32 filters with kernel size 7 to capture broader spectral structure that reflects more global channel response characteristics. Each path applies batch normalization and max pooling (pool size = 2) to stabilize optimization and reduce temporal (subcarrier) dimensionality. Path outputs are concatenated along the channel

Layer	Type	Filter / Kernel	Activation	Output Shape	Parameters
0	CSI Input	–	–	(52, 1)	0
1	RSSI Input	–	–	(1)	0
2	Conv1D (Path 1)	32 filters, kernel=3, padding=same	ReLU	(52, 32)	128
3	Conv1D (Path 2)	32 filters, kernel=7, padding=same	ReLU	(52, 32)	256
4	BatchNormalization	–	–	(26, 32) each	128
5	MaxPooling1D	pool size=2	–	(26, 32) each	0
6	Concatenate	–	–	(26, 64)	0
7	Conv1D	64 filters, kernel=3, padding=same	ReLU	(26, 64)	12,352
8	GlobalAveragePooling1D	–	–	(64)	0
9	Dense (CSI)	128 units	ReLU	(128)	8,320
10	Dense (RSSI)	32 units	ReLU	(32)	64
11	Dense (RSSI)	32 units	ReLU	(32)	1,056
12	Dense (RSSI)	32 units	ReLU	(32)	1,056
13	Concatenate	–	–	(160)	0
14	Dense	256 units	ReLU	(256)	41,216
15	Dropout	rate=0.3	–	(256)	0
16	Dense	128 units	ReLU	(128)	32,896
17	Dropout	rate=0.2	–	(128)	0
18	Dense	64 units	ReLU	(64)	8,256
19	Dense	2 units	Linear	(2)	130

Table 5.2: Layer configuration of the proposed hybrid CNN with RSSI integration

axis and further processed by a convolutional layer with 64 filters (kernel size 3) to learn higher-order feature interactions. A Global Average Pooling (GAP) operation then reduces each feature map to a single scalar, yielding a compact CSI feature vector which is projected by a 128-unit dense layer to form the final CSI representation.

RSSI branch (fully connected):

The RSSI branch comprises a lightweight fully connected stack. The scalar RSSI input is passed through three dense layers of 32 units each, with ReLU activations, enabling nonlinear calibration and implicit compensation for device- or orientation-dependent bias. The output of the RSSI branch is a 32-dimensional feature vector that complements the CSI representation.

Feature fusion and regression head:

We fuse the modality-specific representations by concatenating the 128-dimensional CSI feature vector with the 32-dimensional RSSI feature vector to obtain a 160-dimensional joint descriptor. The joint vector is processed by three dense layers (256, 128, 64 units) with ReLU activations; dropout (rates 0.3 and 0.2) is applied in the first two dense layers to mitigate overfitting. The final linear output layer contains two units and yields the coordinate prediction $(\hat{x}, \hat{y})$. A condensed layer-wise configuration of the hybrid model is presented in Table 5.2.

Training Objective

The hybrid network is trained end-to-end using gradient-based optimization with the mean squared error (MSE) loss identical to Equation 5.9.

Because the input modalities have different dynamic ranges and units, preprocessing is applied prior to training: CSI amplitudes are normalized per subcarrier to zero-mean and unit-variance across the training set, and RSSI values are standardized to the same scale. This normalization ensures stable gradient propagation and allows the fusion layers to learn meaningful relative weightings of CSI and RSSI features. During training, validation localization error is monitored to guide hyperparameter selection and to perform early stopping, thereby reducing the risk of overfitting.

Summary

The hybrid CNN with RSSI integration extends the baseline by combining fine-grained spectral signatures (CSI) with coarse-grained distance cues (RSSI) in a principled dual-branch architecture. The CSI branch extracts multi-scale frequency-domain features through parallel convolutional paths and GAP compression, while the RSSI branch provides complementary global signal-strength information via a compact fully connected stack. The concatenated representation is mapped to continuous two-dimensional coordinates through a small regression head trained with Euclidean loss. This multi-modal

fusion strategy is designed to improve localization robustness to measurement noise and environmental variation relative to single-modality baselines.

5.3.3 Attention-based Convolutional Neural Network (A-CNN)

Input Structure

The attention-based model extends the baseline CNN by leveraging both CSI amplitude and phase to improve localization accuracy. Each training instance comprises amplitude and phase measurements across all 52 subcarriers, forming a matrix $\mathbf{X} \in \mathbb{R}^{52\times 2}$. The amplitude channel captures the magnitude variations across subcarriers, reflecting frequency-selective fading and multipath structure, whereas the phase channel encodes frequency-dependent phase shifts, aiding in disambiguating complex multipath effects. Prior to attention processing, both channels are normalized per subcarrier to zero mean and unit variance, ensuring consistent dynamic ranges. Each fingerprint $\mathbf{X}$ is paired with the corresponding ground-truth coordinates $\mathbf{y} = (x, y)$, producing supervised training pairs:

$$(\mathbf{X}, \mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^2. \tag{5.11}$$

The novelty of this approach lies in the self-attention mechanism, which adaptively assigns weights to different subcarriers, allowing the network to focus on the most informative frequency components for precise localization. This enables the model to capture long-range dependencies across the subcarrier spectrum, enhancing robustness in multipath-rich indoor environments.

Network Architecture

We retain a convolutional backbone similar to the baseline CNN to extract local and medium-range spectral features from the CSI input. Specifically, two sequential Conv1D layers (64 filters, kernel sizes 5 and 3) capture medium- and fine-grained subcarrier patterns, followed by batch normalization to stabilize training.

The extracted feature map $\mathbf{H} \in \mathbb{R}^{52\times 64}$ is then processed by a self-attention module, which implements learnable query, key, and value transformations:

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}\mathbf{W}_V, \tag{5.12}$$

$$\mathbf{A} = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\Big), \tag{5.13}$$

$$\mathbf{Z} = \mathbf{A}\mathbf{V}, \tag{5.14}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{64\times 64}$ are learnable parameters and $d_k = 64$.

The attention output $\mathbf{Z}$ is combined with the original features via a residual connection and layer normalization:

$$\mathbf{H}' = \mathrm{LayerNorm}(\mathbf{H} + \mathbf{Z}).$$

This allows the network to emphasize the most informative subcarriers for each spatial location while preserving the underlying spectral structure.

The attention-enhanced features are then compressed using Global Average Pooling (GAP) and passed through three fully connected layers (256, 128 units with dropout rates 0.3 and 0.2) before producing the predicted coordinates $(\hat{x}, \hat{y})$ via a linear output layer.

Layer	Type	Filter / Kernel	Activation	Output Shape	Parameters
0	Input	–	–	(52, 2)	0
1	Conv1D	64 filters, kernel=5, padding=same	ReLU	(52, 64)	704
2	BatchNormalization	–	–	(52, 64)	256
3	Conv1D	64 filters, kernel=3, padding=same	ReLU	(52, 64)	12,352
4	BatchNormalization	–	–	(52, 64)	256
5	Dense (Query)	64 units	Linear	(52, 64)	4,160
6	Dense (Key)	64 units	Linear	(52, 64)	4,160
7	Dense (Value)	64 units	Linear	(52, 64)	4,160
8	Dot Product	QÂ·K^T/â64	–	(52, 52)	0
9	Softmax	–	Softmax	(52, 52)	0
10	Dot Product	AttentionÂ·V	–	(52, 64)	0
11	Add	Residual connection	–	(52, 64)	0
12	LayerNormalization	–	–	(52, 64)	128
13	GlobalAveragePooling1D	–	–	(64)	0
14	Dense	256 units	ReLU	(256)	16,640
15	Dropout	rate=0.3	–	(256)	0
16	Dense	128 units	ReLU	(128)	32,896
17	Dropout	rate=0.2	–	(128)	0
18	Dense	2 units	Linear	(2)	258

Table 5.3: Layer configuration of the proposed attention-based CNN

Training Objective

The attention-based CNN is trained end-to-end with mean squared error (MSE) loss identical to Equation 5.9. The self-attention weights are optimized jointly with the convolutional backbone to minimize Euclidean localization error.

Summary

The attention-based CNN extends the baseline convolutional architecture by incorporating a self-attention mechanism to adaptively weight subcarriers according to their spatial relevance. This enables the network to selectively emphasize the most informative frequency components, improving localization accuracy and robustness in complex multipath environments while preserving the hierarchical feature extraction capabilities of the convolutional backbone.

5.3.4 Multi-Scale Convolutional Neural Network (MS-CNN)

Input Structure

The multi-scale CNN is designed for amplitude-only CSI; for each sample we retain the magnitude response across all 52 subcarriers and discard phase. Similar to other models, phase is omitted due to its high sensitivity to synchronization offsets and hardware calibration errors, which can degrade robustness and performance. The single-channel input emphasizes spectral envelope and magnitude patterns across local and global frequency scales. Formally,

$$\mathbf{X} \in \mathbb{R}^{52\times 1},$$

and each training instance is paired with the corresponding ground-truth coordinates (x, y), forming supervised training pairs:

$$\big(\mathbf{X}, (x, y)\big), \tag{5.15}$$

where $(x, y) \in \mathbb{R}^2$. Prior to training each subcarrier is standardized across the training set.

Network Architecture

The multi-scale design is motivated by the fact that indoor multipath effects manifest at different spectral resolutions: localized fades appear over 2–3 subcarriers, while global channel envelopes span tens of subcarriers. To capture these heterogeneous dependencies, we employ three parallel convolutional paths: kernel size 3 for local fading variations, kernel size 7 for medium-range multipath patterns, and kernel size 15 for global spectral envelopes.

Layer	Type	Filter / Kernel	Activation	Output Shape	Parameters
0	Input	–	–	(52, 1)	0
1	Conv1D (Path 1)	32 filters, kernel=3, padding=same	ReLU	(52, 32)	128
2	Conv1D (Path 2)	32 filters, kernel=7, padding=same	ReLU	(52, 32)	256
3	Conv1D (Path 3)	32 filters, kernel=15, padding=same	ReLU	(52, 32)	512
4	BatchNormalization	–	–	(52, 32) each	384
5	MaxPooling1D	pool size=2	–	(26, 32) each	0
6	Concatenate	–	–	(26, 96)	0
7	Conv1D	128 filters, kernel=3, padding=same	ReLU	(26, 128)	36,992
8	BatchNormalization	–	–	(26, 128)	512
9	GlobalAveragePooling1D	–	–	(128)	0
10	Dropout	rate=0.3	–	(128)	0
11	Dense	256 units	ReLU	(256)	33,024
12	Dropout	rate=0.3	–	(256)	0
13	Dense	128 units	ReLU	(128)	32,896
14	Dropout	rate=0.2	–	(128)	0
15	Dense	2 units	Linear	(2)	258

Table 5.4: Layer configuration of the proposed multi-scale CNN

Each path applies batch normalization and max pooling (pool size=2) to stabilize training and reduce dimensionality. Outputs from the three paths are concatenated along the channel axis, forming a 96-dimensional multi-scale feature representation.

A subsequent convolutional layer with 128 filters (kernel size 3) integrates the concatenated features, learning higher-order interactions between local and global scales. Global Average Pooling compresses the resulting feature maps, producing a compact representation for regression.

The final stages consist of two dense layers (256 and 128 units with dropout rates 0.3 and 0.2) followed by a linear output layer of two units to produce the coordinate prediction $(\hat{x}, \hat{y})$.

Training Objective

The MS-CNN is trained end-to-end with MSE loss as in Equation 5.9, optimizing for Euclidean localization accuracy.

Summary

The MS-CNN introduces multi-scale convolutional branches that simultaneously capture fine-grained, medium-range, and global spectral patterns in CSI amplitude. This design is tailored to indoor localization, where multipath effects arise at diverse frequency scales. By combining multi-scale features through concatenation and hierarchical integration, the MS-CNN provides improved robustness relative to single-scale convolutional models.

5.3.5 Residual CNN for CSI-Based Indoor Localization (RCNN)

Input Structure

Similar to the multi-scale CNN, the residual CNN operates on amplitude-only CSI fingerprints. Each sample consists of magnitude responses across 52 subcarriers:

$$\mathbf{X} \in \mathbb{R}^{52\times 1}.$$

Inputs are standardized on a per-subcarrier basis to zero mean and unit variance across the training set to ensure comparability across subcarriers and stable gradient dynamics. Each fingerprint $\mathbf{X}$ is paired with the corresponding ground-truth coordinates:

$$\big(\mathbf{X}, (x, y)\big), \qquad (x, y) \in \mathbb{R}^2. \tag{5.16}$$

Network Architecture

The RCNN architecture introduces residual skip connections to enable deeper networks while mitigating vanishing gradient issues. The model begins with a 1-D convolutional

layer (32 filters, kernel size = 7, ReLU activation) to extract low-level spectral features, followed by batch normalization to stabilize optimization.

Three residual blocks are then stacked with increasing channel depths (32, 64, 128). Each residual block comprises two Conv1D layers (kernel size = 3, same padding), each followed by batch normalization and ReLU. Skip connections add the block input to its output, encouraging feature reuse and preserving gradient flow. When the number of channels changes across blocks, a 1×1 convolution projection is applied to align dimensions. After the first and second residual blocks, max pooling with pool size 2 is applied to reduce resolution along the subcarrier dimension.

The final residual block outputs a feature map of size $(13, 128)$, which is compressed via Global Average Pooling to form a compact feature vector. This vector is passed through two fully connected layers (256 and 128 units, ReLU activations) with dropout regularization (0.3 and 0.2, respectively) before reaching a linear output layer of 2 units that predicts the spatial coordinates $(\hat{x}, \hat{y})$.

Layer	Type	Filter / Kernel	Activation	Output Shape	Parameters
0	Input	–	–	(52, 1)	0
1	Conv1D	32 filters, kernel=7, padding=same	ReLU	(52, 32)	256
2	BatchNorm	–	–	(52, 32)	128
3–6	Residual Block 1	32 filters, kernel=3	ReLU	(52, 32)	6,272
7	MaxPooling1D	pool size=2	–	(26, 32)	0
8–11	Residual Block 2	64 filters, kernel=3	ReLU	(26, 64)	24,896
12	MaxPooling1D	pool size=2	–	(13, 64)	0
13–16	Residual Block 3	128 filters, kernel=3	ReLU	(13, 128)	99,072
17	GlobalAveragePooling1D	–	–	(128)	0
18	Dense	256 units	ReLU	(256)	33,024
19	Dropout	rate=0.3	–	(256)	0
20	Dense	128 units	ReLU	(128)	32,896
21	Dropout	rate=0.2	–	(128)	0
22	Dense	2 units	Linear	(2)	258

Table 5.5: Layer configuration of the proposed residual CNN

Training Objective

The RCNN is trained end-to-end with mean squared error (MSE) loss, consistent with other models in this work:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left\| (x_i, y_i) - (\hat{x}_i, \hat{y}_i) \right\|_2^2. \tag{5.17}$$

This formulation explicitly penalizes Euclidean distance errors between predicted and ground-truth positions. Early stopping is applied based on validation error to mitigate overfitting.

Summary

The RCNN leverages residual connections to enable deeper model capacity while preserving stable gradient flow. By learning hierarchical representations of CSI amplitude from fine-grained local fading patterns to high-level global abstractions, we expect residual CNN to improve robustness and localization accuracy compared to shallower CNN baselines.

Summary of Proposed Architectures

We introduced a family of neural architectures that formulate CSI-based indoor localization as a supervised regression task. The proposed models span increasing levels of architectural complexity:

- A lightweight baseline 1-D CNN that extracts local spectral features and directly regresses to (x, y) coordinates.
- A hybrid dual-branch network that fuses amplitude-only CSI with RSSI, combining fine-grained spectral cues with coarse distance priors.
- An attention-augmented CNN that applies self-attention across subcarriers to adaptively emphasize frequency components most informative for localization.
- A multi-scale CNN that captures spectral structures ranging from local fading variations to global envelope characteristics.
- A deeper residual CNN that leverages skip connections to improve gradient flow, enabling hierarchical feature reuse and enhanced representational capacity.

These architectures collectively explore complementary strategies for enhancing localization performance: multi-modality fusion, attention-driven feature selection, multi-scale representation learning, and residual-based deep modeling. In the following section, we empirically evaluate these models in terms of localization accuracy (mean, median, and percentile errors with cumulative distribution functions), robustness to environmental variation, and computational efficiency.

5.4 Performance Study

Having outlined the architectural designs, we now turn to their empirical evaluation. This section benchmarks the proposed models under controlled experimental conditions, focusing on localization accuracy, quantified through mean, median, and percentile errors as well as cumulative distribution functions (CDFs)as well as the robustness to environmental setup and data quality.

Unlike many prior works that exploit MIMO configurations or specialized hardware, all experiments here are conducted with a single commercial access point and a single-antenna

WROOM32 receiver. This setup highlights the feasibility and limitations of deep learning-based localization in resource-constrained deployments. The following results provide a comparative assessment of how different design choice, attention, multi-scale processing, residual learning, and modality fusion, translate into empirical performance gains under these constraints.

5.4.1 Performance of Classical Localization Algorithms

Prior to evaluating deep learning-based approaches, we established a set of classical baselines using the collected CSI amplitudes (52 subcarriers) yielding a 52-dimensional feature vector $\mathbf{x} \in \mathbb{R}^{52}$. Classical methods remain important baselines because of their interpretability, low implementation overhead, and widely understood failure modes. In the following we summarize the algorithms evaluated, describe their training and inference procedures, and report empirical performance on the measurement grid. Results are shown as cumulative distribution functions in Fig. 5.6 and summarized in Tables 5.6 5.7 and 5.8

K-Nearest Neighbors (KNN)

KNN infers a test location by finding the k nearest training fingerprints in Euclidean feature space and aggregating their coordinates. Concretely, given a query $\mathbf{x}$ we compute $d_i = \|\mathbf{x} - \mathbf{x}_i\|_2$ for every stored training sample $\mathbf{x}_i$, select the k samples with smallest d_i, and output the arithmetic mean of their (x, y) coordinates. We evaluated $k \in \{1, 3, 5, 9\}$ to trade off variance and bias.

Empirically, KNN produced low performance under our single-antenna, amplitude-only features. The $k = 5$ variant provided the best median error among the tested k values, with a median localization error of 2.713 m (Table 5.6). Smaller k (e.g., $k = 1$ or $k = 3$) yielded larger median errors and more variable outliers, reflecting the sensitivity of nearest-neighbor methods to training-sample sparsity and measurement noise on a coarse grid.

Model	Median (m)	Mean (m)	1m Acc	2m Acc
k-NN (k=1)	3.606	3.898	12.1	21.1
k-NN (k=3)	3.256	3.369	15.7	29.8
k-NN (k=5)	2.713	3.043	17.8	28.1
k-NN (k=9)	3.101	3.420	13.0	20.7

Table 5.6: k-Nearest Neighbors Models

Inverse Distance Weighting (IDW)

IDW produces a continuous coordinate estimate by weighting all training coordinates according to their feature-space proximity to the query. For each training sample we compute

a weight

$$w_i = \frac{1}{(d_i)^p + \varepsilon},$$

where $d_i = \|\mathbf{x} - \mathbf{x}_i\|_2$, p is a power parameter, and $\varepsilon = 10^{-6}$ avoids division by zero. The estimated location is the weighted average of all training coordinates. We tested $p \in \{1, 2, 4\}$.

IDW smooths estimates across the entire dataset and can mitigate single-sample noise, but it does not model structured correlations across subcarriers induced by multipath. In our experiments IDW attained median errors in the 2.8–3.0 m range; the $p = 2$ configuration provided the best trade-off for two-meter accuracy (Table 5.7).

Model	Median (m)	Mean (m)	1m Acc	2m Acc
IDW (p=1)	2.907	2.797	18.8	26.8
IDW (p=2)	2.831	2.572	14.0	27.8
IDW (p=4)	3.008	3.112	9.4	20.8

Table 5.7: Inverse Distance Weighting (IDW) Models

Probabilistic (Gaussian) Fingerprinting

The probabilistic baseline models the distribution of fingerprints observed at each reference point as a multivariate Gaussian. During training we grouped samples by reference location and estimated the per-location mean vector and covariance matrix. At the positioning phase, the log-likelihood of a query $\mathbf{x}$ under each reference distribution is computed, and the reference point with maximum likelihood is selected; its coordinates are returned as the estimate. This is a maximum-likelihood scheme that outputs a discrete reference-point location contrarily to the continuous localization we are aiming for, but even so, this method gave the best median error among the classical baselines, with a median of 2.686 m and improved robustness to outliers compared to nearest-neighbor variants (Table 5.8). Its discrete output, however, is constrained by the reference-grid resolution and thus cannot directly yield sub-grid continuous corrections without additional interpolation.

Model	Median (m)	Mean (m)	1m Acc	2m Acc
Probabilistic	2.686	2.895	19.3	29.4

Table 5.8: Probabilistic Model

Discussion

Across the evaluated classical algorithms the best median localization error was approximately 2.69 m (probabilistic), while the best KNN variant (k=5) reached 2.71 m. These

results illustrate several recurring limitations of classical fingerprinting under our constrained, single-antenna experimental conditions:

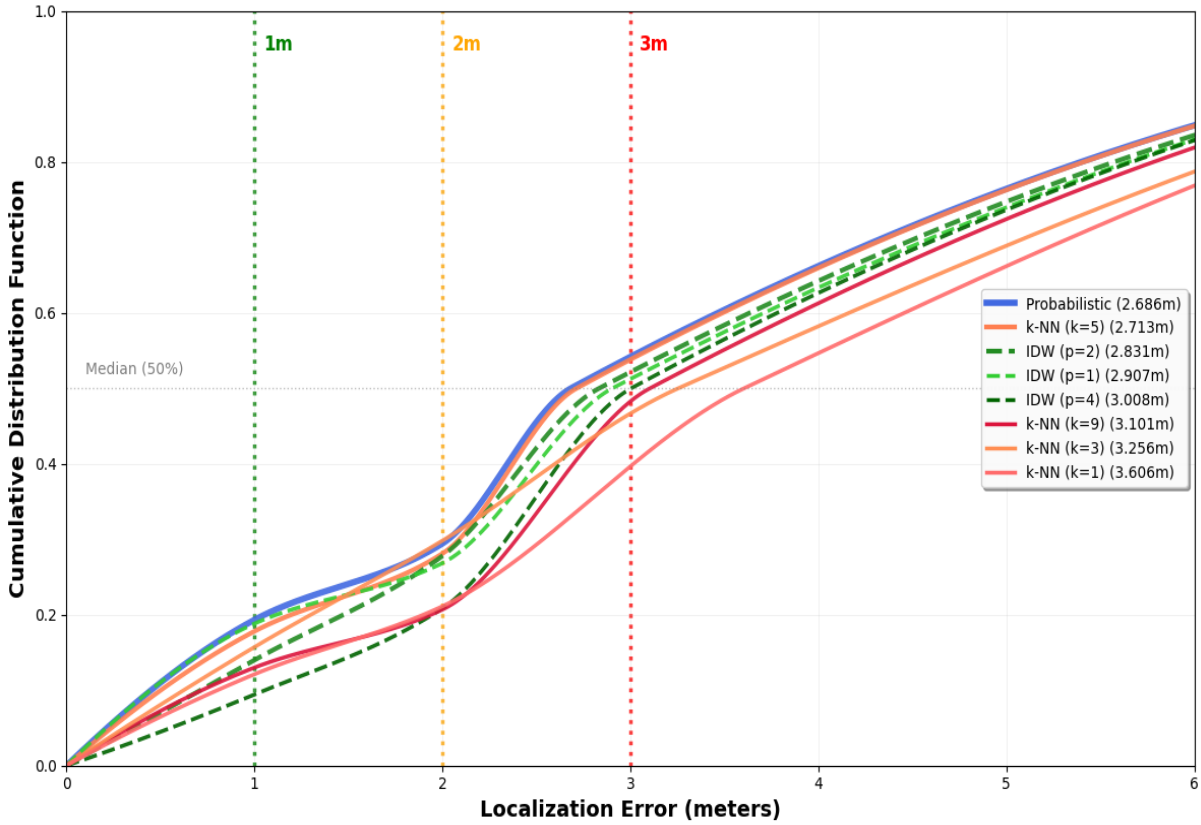

Figure 5.6: Cumulative Distribution Functions of our proposed Models

- **Dependence on grid density and training coverage:** All methods degrade as the spatial sampling becomes sparse; coarse grids limit the attainable resolution irrespective of the estimator used.
- **Limited exploitation of CSI structure:** The evaluated algorithms operate on the 52-D feature vector in a largely isotropic manner (Euclidean distances or Gaussian likelihoods) and do not capture structured inter-subcarrier correlations or frequency-domain patterns that arise from multipath propagation.
- **Sensitivity to noise and outliers:** Nearest-neighbor methods are particularly vulnerable to individual noisy captures; averaging or weighting (larger k, IDW) can reduce variance but at the cost of increased bias.

Taken together, these observations motivate the subsequent exploration of convolutional neural network (CNN) architectures. CNNs can learn hierarchical, non-linear feature transformations that exploit inter-subcarrier structure and spatially coherent patterns in CSI, offering a path to improved accuracy and robustness under the hardware and sampling

constraints considered in this work.

5.4.2 Performance Of Deep Learning Based Solutions

The first experimental phase of the performance evaluation examined five proposed network architectures: (i) a baseline convolutional neural network with stacked convolutional layers (B-CNN), (ii) a multi-scale convolutional network employing filters of varying temporal widths to capture channel state information (CSI) fluctuations at multiple resolutions (MS-CNN), (iii) an attention-augmented network designed to emphasize discriminative CSI features (A-CNN), (iv) a hybrid architecture that fuses CSI with auxiliary received-signal-strength indicator (RSSI) features (H-CNN), and (v) a deeper residual network with skip connections to facilitate gradient propagation (R-CNN). Each architecture was trained using datasets of increasing size, namely 250, 500, and 750 samples per grid point. Performance was evaluated on a fixed test set comprising 750 CSI samples corresponding to intermediate positions.

The baseline model (B-CNN) demonstrated moderate consistency, producing median localization errors in the range 2.2–2.5 m; however, it exhibited limited sub-meter accuracy. The multi-scale variant (MS-CNN), despite its design to capture diverse temporal scales, did not improve upon the baseline and produced median errors consistently above 2.6 m. The attention-augmented model (A-CNN), which leverages both amplitude and phase information, achieved a median error of 1.807 m when trained with 750 samples per grid point; its performance, however, degraded substantially with reduced training set sizes (250 and 500 samples), indicating pronounced sensitivity to the amount of training data.

The hybrid model (H-CNN), the only architecture that incorporates RSSI as an auxiliary input, outperformed the other models under small- and medium-sized training regimes. Notably, H-CNN attained a median error of 1.610 m for the 500-sample condition and achieved 46% and 68% accuracy within a 1 m and 2 m thresholds. Its performance slightly declined for the largest training set (750 samples), suggesting potential issues related to data imbalance or overfitting. The residual network (R-CNN) yielded incremental improvements as the training set size increased but did not match the best-performing models across all data regimes.

Overall, these results indicate clear dataset-size dependence for architectures that exploit richer feature sets (A-CNN and H-CNN), also that multiscale filtering alone (MS-CNN) is insufficient to guarantee improved localization accuracy in the evaluated scenarios.

Despite the architectural diversity, all models exhibited a common limitation: validation performance did not consistently improve as training progressed, suggesting underlying generalization challenges.

Generalization Issues Diagnosis

To improve model performance, several parameters and configurations were systematically varied. The analysis first focused on understanding the underlying causes of poor general-

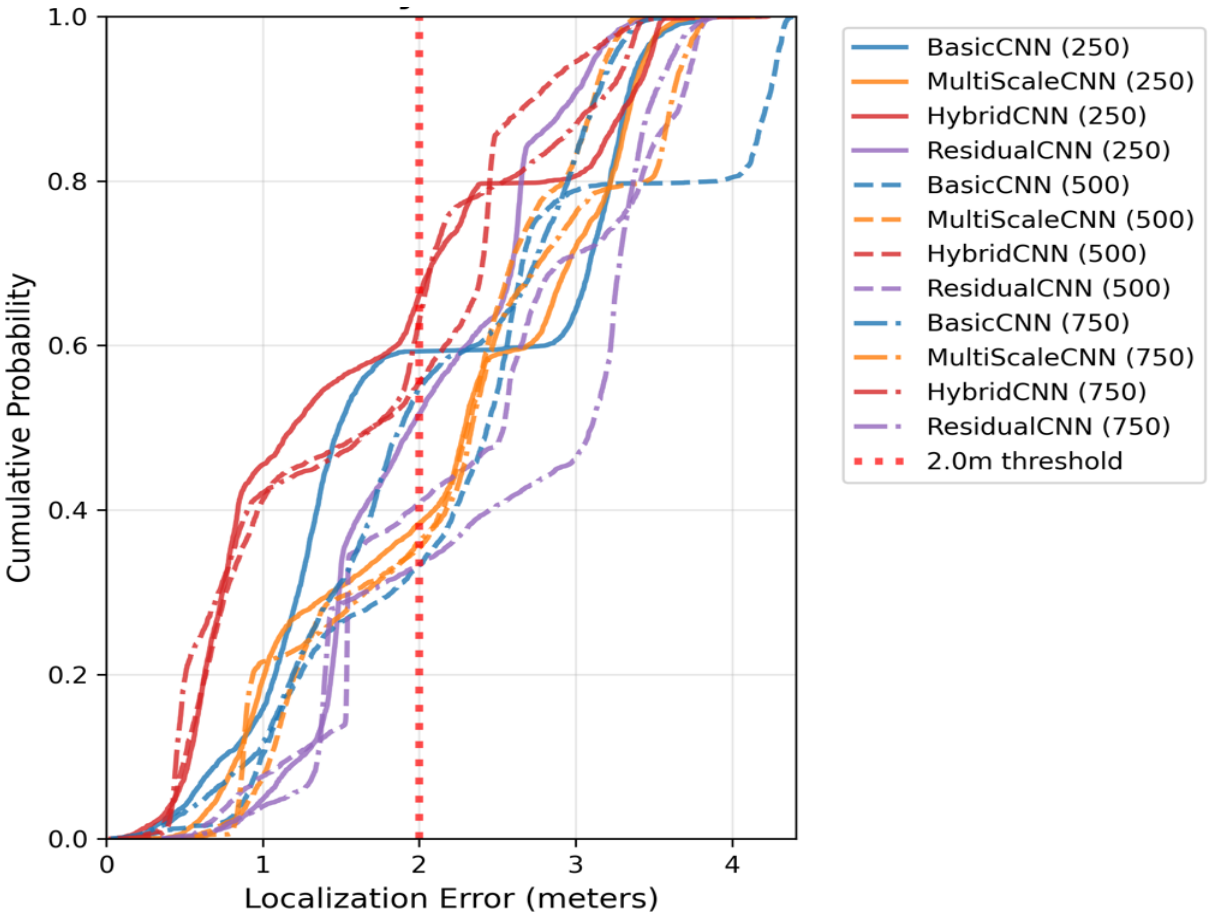

Figure 5.7: Cumulative Distribution Functions of our proposed Models

ization. A critical flaw identified was a subtle form of data leakage: normalization scalers were fitted on the full dataset rather than exclusively on the training subset. This inadvertently allowed information from the test distribution to influence the training process, thereby compromising the model's ability to generalize to unseen data.

A second issue concerned the learning rate. The commonly adopted value of 0.001 in CNN applications proved unsuitable for CSI regression tasks, where target coordinates vary smoothly and require fine-grained updates. With this setting, training exhibited oscillatory convergence, indicating that the step size was excessively aggressive for the task.

Finally, the choice of loss function also contributed to instability. Although a custom Euclidean distance loss is intuitively aligned with localization error, it introduced optimization challenges compared to the simpler and numerically stable mean squared error (MSE). MSE not only facilitates smoother gradient dynamics but also remains the established standard for regression problems in the literature.

Retraining with improved parameters

To address the limitations diagnosed above, a series of targeted corrections were applied. Data leakage was eliminated by fitting scalers exclusively on training data, ensuring a proper separation between training and evaluation. The learning rate was reduced by a factor of five to 0.0002, which yielded smoother convergence. Regularization was strengthened by combining L2 weight decay with increased dropout, balancing model complexity against overfitting risks. The unstable Euclidean loss was replaced with MSE, improving numerical stability. Finally early stopping patience was increased, allowing models more time to converge under stricter regularization.

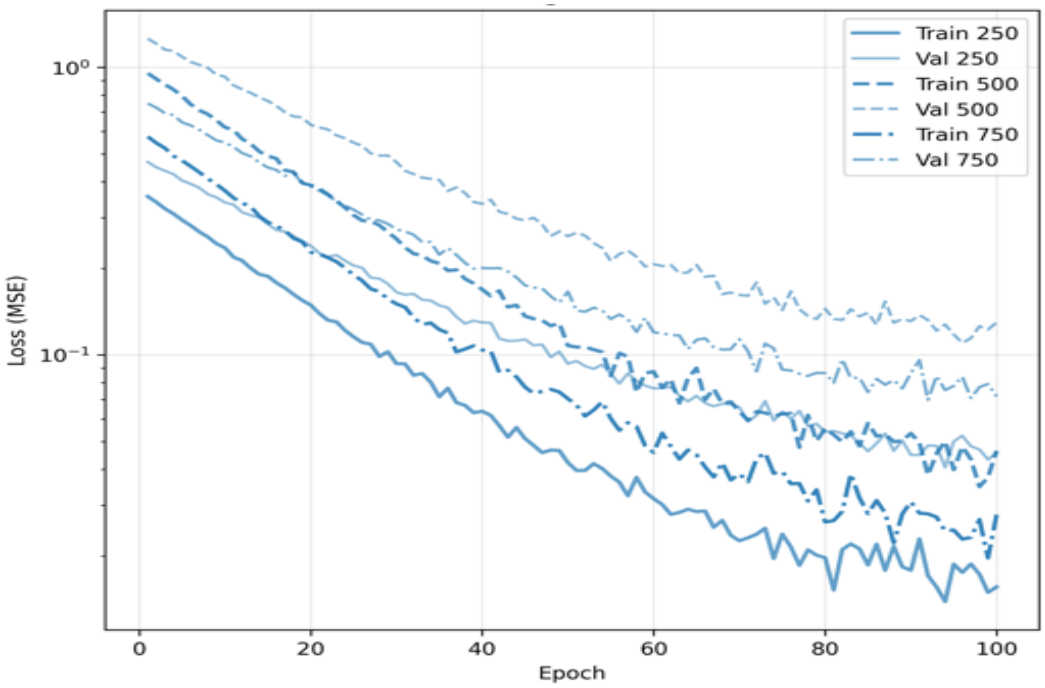

Figure 5.8: Learning curve for BCNN stopped after 100 epochs

These methodological corrections had a substantial impact on performance. Among all architectures, H-CNN remained the best-performing model, achieving a median localization error of 1.193 m with only 250 training samples. This result not only represented the lowest error across all experiments but also yielded a sub-meter accuracy rate of 51.5% and accuracy within 2 m exceeding 76%. Relative to its Phase 1 counterpart, this constituted a marked improvement in both error reduction and reliability. The baseline B-CNN also demonstrated notable gains, reducing its error from 2.524 m to 1.686 m on the smallest dataset. In contrast, the residual network (R-CNN) failed to benefit from the adjustments. Under stronger regularization, its median error continued to increase with larger training sets, degrading further from its already uncompetitive Phase 1 performance. These results suggest that deeper residual architectures are less suitable for CSI-based localization under the limited data conditions considered, whereas shallower architectures can exploit the

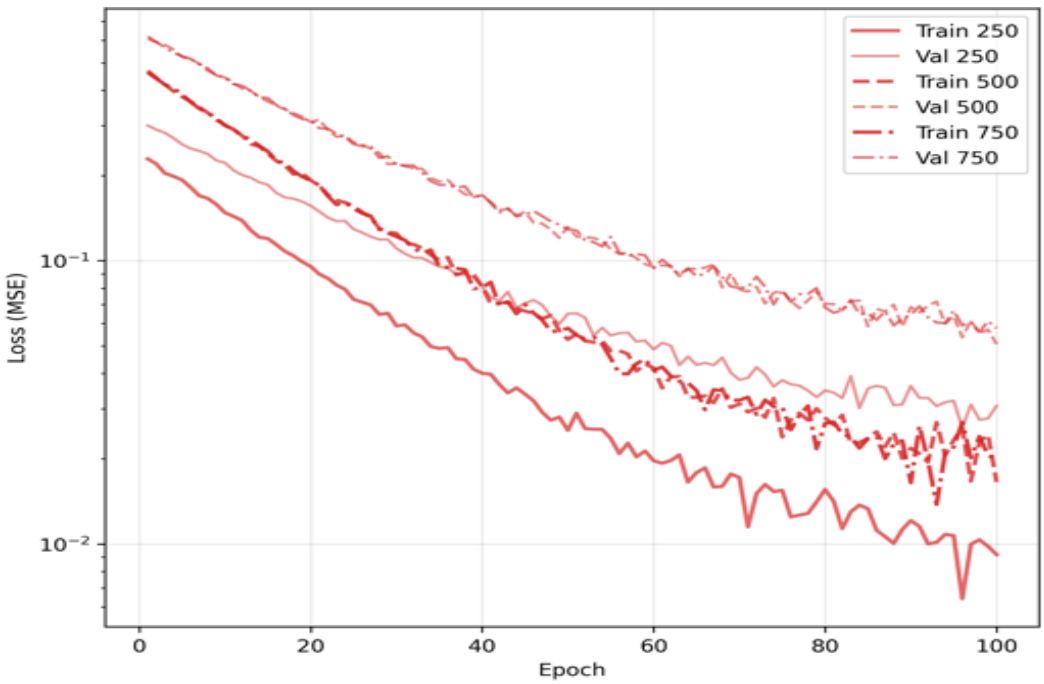

Figure 5.9: Learning curve for HCNN stopped after 100 epochs

available training samples more effectively.

The magnitude of improvement observed after correcting the experimental flaws underscores the critical role of methodological rigor in CSI-based deep learning. The hybrid model's consistently strong results, particularly with the inclusion of RSSI as a complementary feature, further emphasize that performance gains arise less from architectural novelty alone and more from appropriate data handling and optimization choices, which ultimately determine whether models generalize effectively.

Figures 5.8 and 5.9 present the learning curves for B-CNN and H-CNN, respectively, illustrating the evolution of both training and validation mean squared error (MSE) during optimization. In both cases, the MSE decreases steadily over the course of training, confirming effective learning of the underlying mapping from CSI (and RSSI in the hybrid case) to spatial coordinates. However, after approximately 100 epochs, a divergence between training and validation error becomes apparent. While the training error continues to decrease monotonically, the validation error begins to rise, indicating the onset of overfitting. For this reason, training was deliberately cut at 100 epochs to preserve generalization performance.

Effects of spatial distribution on accuracy

For furthe analysis, we now use the H-CNN model trained on 250. Beyond global error statistics, spatial analysis of test point predictions provides critical insight into model behavior. In Fig 5.10 Predictions at three points, (0.5,0.5), (1.5,4.5), and (3.5,1.5) demon-

strated excellent generalization, with sub-meter accuracy rates of 100%. These locations were centrally positioned within the training grid and benefited from interpolation across multiple nearby training points.

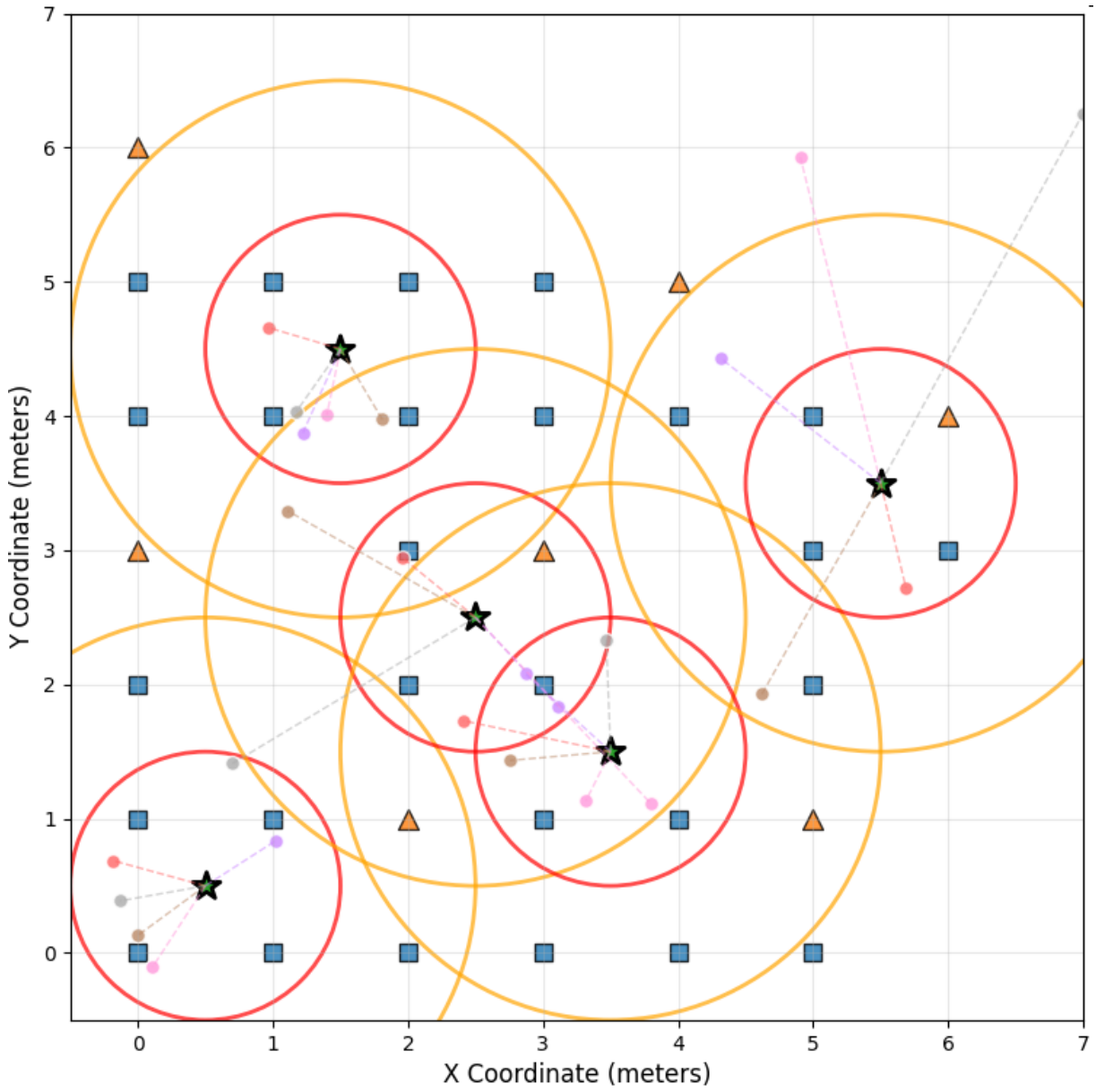

Figure 5.10: Hybrid Model positioning Of Test Points

At the test locations (2.5,2.5) and (5.5,3.5), performance was mixed. Although the minimum positioning error reached 0.618m, a larger proportion of predictions fell outside the 1-m accuracy range. Notably, results revealed a degradation in generalization capability, even in regions that contained training points. These locations either lay between the two primary table clusters where no reference points were accessible or near the periphery of

the training space, which was dominated by furniture and electronic devices.

These observations are consistent with known properties of indoor RF propagation. Channel State Information (CSI) variations are highly sensitive to multipath geometry, and when test points fall outside the convex hull of the training set, models are forced to extrapolate in regimes where propagation behavior is poorly constrained. This accounts for the discrepancy between the high accuracy achieved at interpolated points and the reduced accuracy at extrapolated ones.

Aggregate statistics over 25 predictions confirmed these trends: the mean error was 1.570m, the median error was 1.176m and the proportion of sub-meter predictions was 52%. While acceptable in this constrained setting, these results underscore the limitations of the available data, particularly when relying on a single antenna and a single access point.

Further Attempts to improve accuracy

In the final phase, we varied hyperparameters such as learning rate, batch size, and dropout rate; however, the resulting models did not surpass the best-performing configuration. The closest result achieved was a median localization error of 1.469 m with 750 training samples, attaining sub-meter accuracy in approximately 36% of cases. Although this represents an improvement over the baseline CNN, it remained inferior to the systematically optimized Hybrid CNN trained on smaller datasets. These findings indicate that the primary limitation arises not from optimization strategy, but from the expansiveness, scale, and spatial representativeness of the dataset. While advanced training schedules or gradient clipping can improve convergence, they cannot compensate for insufficient training coverage in multipath-dominated indoor environments.

5.4.3 Performance Comparison With Similar Works

This study presented a multi-phase analysis of CNN-based indoor localization using WiFi CSI and RSSI in a single-AP setting. Through systematic architectural comparisons, identification of methodological flaws, and subsequent corrections, the results demonstrate that the richness and quality of training data are as critical as architectural design. Among the evaluated models, the Hybrid-CNN, which integrates CSI amplitude with auxiliary RSSI input, consistently achieved the most competitive performance, reaching median errors as low as 1.193 m and attaining nearly 52% sub-meter accuracy. Notably, its performance exhibited dependence on the spatial region of the laboratory and the density of training points therein, underscoring the sensitivity of CSI-based deep learning models to data distribution and sampling coverage.

The spatial analysis that followed highlighted a critical limitation: CNNs generalize effectively through interpolation but fail at extrapolation, particularly near the boundaries of the training grid and the inaccessible zones inside it. This reflects the underlying physics of multipath propagation, where channel variations outside the training distribution cannot be captured by data-driven models alone.

As highlighted throughout this work, the objective was to examine the limits of a constrained setup in terms of the hardware used for radio frequency transmission. Prior efforts have explored this challenge under varying assumptions. For instance, CUPID [SLKC13] exploited CSI to extract the angle and distance of the direct path from a single AP, achieving a median localization error of approximately 5 m. More recently, S-Phaser [HLM+19] leveraged the richer information provided by MIMO-enabled hardware to estimate direct path lengths using calibrated CSI phases, achieving a median error of 1.5 m even in non-line-of-sight scenarios. While this performance highlights the benefits of multi-antenna diversity, our work deliberately restricts itself to a single-antenna configuration, representing a more constrained and cost-efficient setting. Taken together, these findings suggest that CNN-based architectures, when properly optimized, can attain acceptable performance under such restrictions, but their success is restricted by the data quality and environment-specific training coverage.

Chapter 6

Conclusion and Future Work

In this work, we investigated the use of convolutional neural network (CNN) architectures for regression-based indoor localization using CSI and RSSI measurements acquired from a single-antenna WROOM32 receiver. We first evaluated classical machine learning approaches, including K-Nearest Neighbors (KNN), probabilistic fingerprinting, and inverse distance weighting (IDW). While these traditional methods offer simplicity, low computational cost, and moderate accuracy, their performance was limited by sparse reference point deployment, coarse spatial resolution, and the inability to capture complex, non-linear correlations inherent in multipath propagation.

To overcome these limitations, we proposed and evaluated multiple deep learning architectures, including a baseline CNN, a hybrid CNN integrating RSSI, a multi-scale CNN, an attention-based CNN, and a residual CNN. These models exploit amplitude and phase CSI features, as well as RSSI, to learn hierarchical and non-linear representations that map channel fingerprints to physical coordinates. Our results demonstrate that deep learning solutions consistently outperform classical ML algorithms, providing improved localization accuracy and robustness, even under the constraints of a single access point and single-antenna receiver.

Despite these improvements, the performance of all models remains constrained by the quality and richness of the available measurements. Single-antenna CSI provides limited spatial diversity, and the sparse reference grid restricts the model's ability to generalize fine-grained spatial variations. Consequently, the localization accuracy reaches a practical limit in this scenario. Future work could explore MIMO systems or multi-access point deployments, which provide additional spatial diversity and richer measurement modalities, enabling models to capture more detailed multipath characteristics and achieve substantially higher localization precision.

Overall, this study highlights both the potential and the current limitations of deep learning-based indoor localization with commodity hardware. The proposed CNN-based architectures provide a flexible and effective framework for leveraging CSI and RSSI in-

formation, but further improvements will require richer data, higher-density deployments, and potentially hybrid multi-modal sensing approaches to fully exploit the complexity of indoor wireless environments.

List of Figures

List of Tables

Bibliography

[A+22] A. O. Adikpe et al. A review on wireless fidelity co-location technology adopted indoors for technology-based contact tracing. *Jordan Journal of Electrical Engineering*, 8(2):134, 2022.

[DYWY19] P. Dai, Y. Yang, M. Wang, and R. Yan. Combination of dnn and improved knn for indoor location fingerprinting. *Wireless Communications and Mobile Computing*, 2019:1–9, 2019.

[FNL22] X. Feng, K. A. Nguyen, and Z. Luo. A survey of deep learning approaches for wifi-based indoor positioning. *Journal of Information and Telecommunication*, 6(2):163–216, 2022.

[HB20] S. M. Hernandez and E. Bulut. Lightweight and standalone iot based wifi sensing for active repositioning and mobility. In *2020 IEEE 21st International Symposium on âA World of Wireless, Mobile and Multimedia Networksâ (WoWMoM)*, pages 277–286, Cork, Ireland, 2020.

[HCN19] C. H. Hsieh, J. Y. Chen, and B. H. Nien. Deep learning-based indoor localization using received signal strength and channel state information. *IEEE Access*, 7:33256–33267, 2019.

[HGWY17] X. Huang, S. Guo, Y. Wu, and Y. Yang. A fine-grained indoor fingerprinting localization based on magnetic field strength and channel state information. *Pervasive and Mobile Computing*, 41(17):150–165, 2017.

[HLM+19] Shuai Han, Yi Li, Weixiao Meng, Cheng Li, Tianqi Liu, and Yanbo Zhang. Indoor localization with a single wi-fi access point based on ofdm-mimo. *IEEE Systems Journal*, 13(1):964–975, 2019.

[HZY+19] Minh Tu Hoang, Yizhou Zhu, Brosnan Yuen, Tyler Reese, Xiaodai Dong, Tao Lu, Robert Westendorp, and Michael Xie. A soft range limited k-nearest neighbours algorithm for indoor localization enhancement. *arXiv preprint arXiv:1908.11480*, 2019. Submitted August 29, 2019.

[KAR+20] K. A. Kordi, A. Alhammadi, M. Roslee, M. Y. Alias, and Q. Abdullah. A review on wireless emerging iot indoor localization. In *2020 IEEE 5th*

International Symposium on Telecommunication Technologies (ISTT), pages 82–87, Shah Alam, Malaysia, 2020. IEEE.

[KRSS22] K. Karthikeyan, M. Radhika, K. Sivaprakash, and E. Sarayu. An empirical analysis of fingerprinting based device-free localization using iot. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1898–1903, Coimbatore, India, 2022. IEEE.

[LCC19] S. J. Liu, R. Y. Chang, and F. T. Chien. Analysis and visualization of deep neural networks in device-free wi-fi indoor localization. *IEEE Access*, 7:69379–69392, 2019.

[LKS22] S. H. Lee, W. Y. Kim, and D. H. Seo. Automatic self-reconstruction model for radio map in wi-fi fingerprinting. *Expert Systems with Applications*, 192(5):116455, 2022.

[LLLW25] Jie Lin, Hsun-Yu Lee, Ho-Ming Li, and Fang-Jing Wu. Ligen: Gan-augmented spectral fingerprinting for indoor positioning. *arXiv preprint arXiv:2508.03024*, 2025. Submitted August 5, 2025.

[LOL24] Zhe-âYu Lim, Lee-Yeng Ong, and Meng-Chew Leow. Radio frequency-based human activity dataset collected using esp32 microcontroller in line-of-sight and non-line-of-sight indoor experiment setups. *Data in Brief*, 57:111101, 2024.

[LP21] C. Lim and J. Paek. Cost reduction in fingerprint-based indoor localization using generative adversarial network. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1024–1026, Jeju Island, Korea, Republic of, 2021. IEEE.

[Min24] MiniCircuits. The basics of orthogonal frequency-division multiplexing (ofdm). Blog post, Feb 2024. Accessed 2025-08-21.

[PB19] K. S. V. Prasad and V. K. Bhargava. Rss-based positioning in distributed massive mimo under unknown transmit power and pathloss exponent. In *2019 IEEE 90th Vehicular Technology Conference*, pages 1–5, Honolulu, HI, USA, 2019. IEEE.

[PH20] A. Poulose and D. S. Han. Hybrid deep learning model based indoor positioning using wi-fi rssi heat maps for autonomous applications. *Electronics*, 10(1):2, 2020.

[PHB18] K. S. V. Prasad, E. Hossain, and V. K. Bhargava. Machine learning methods for rss-based user positioning in distributed massive mimo. *IEEE Transactions on Wireless Communications*, 17(12):8402–8417, 2018.

[RBS21] L. Reichert, S. Brack, and B. Scheuermann. A survey of automatic contact tracing approaches using bluetooth low energy. *ACM Transactions on Computing for Healthcare*, 2(2):1–33, 2021.

[RHDR25] Keyan Rahimi, Md. Wasiul Haque, Sagar Dasgupta, and Mizanur Rahman. Vision-based localization and llm-based navigation for indoor environments. *arXiv preprint arXiv:2508.08120*, pages 1–20, 2025. Submitted August 12, 2025.

[RTR+21] S. Roy, R. J. J. Tiang, M. B. Roslee, M. T. Ahmed, and M. P. Mahmud. Quad-band multiport rectenna for rf energy harvesting in ambient environment. *IEEE Access*, 9:77464–77481, 2021.

[SCP21] N. Singh, S. Choe, and R. Punmiya. Machine learning based indoor localization using wi-fi rssi fingerprints: An overview. *IEEE Access*, 9:127150–127174, 2021.

[SGLH18] X. Sun, X. Gao, G. Y. Li, and W. Han. Single-site localization based on a new type of fingerprint for massive mimo-ofdm systems. *IEEE Transactions on Vehicular Technology*, 67(7):6134–6145, 2018.

[SLKC13] S. Sen, J. Lee, K. H. Kim, and P. Congdon. Avoiding multipath to revive inbuilding wifi localization. In *Proc. Int. Conf. Mobile Syst., Appl. Services (MobiSys)*, pages 249–262, Taipei, Taiwan, June 2013.

[SRM22] A. Sobehy, E. Renault, and P. MÃ¼hlethaler. Generalization aspect of accurate machine learning models for csi-based localization. *Annals of Telecommunications*, 77(5–6):345–357, 2022.

[TB12] M. Tayebi and M. Bouziani. Performance of ofdm in radio mobile channel. *Lecture Notes in Computer Science*, 7340:129–136, 2012.

[TG17] G. Tuna and V. C. Gungor. A survey on deployment techniques, localization algorithms, and research challenges for underwater acoustic sensor networks. *International Journal of Communication Systems*, 30(17):e3350, 2017.

[WWM17] X. Wang, X. Wang, and S. Mao. Resloc: Deep residual sharing learning for indoor localization with csi tensors. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6, Montreal, QC, Canada, 2017. IEEE.

[YWKAO22] J. Yu, P. Wang, T. Koike-Akino, and P. V. Orlik. Multi-modal recurrent fusion for indoor localization. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5083–5087, Singapore, 2022. IEEE.

[YZL13] Z. Yang, Z. Zhou, and Y. Liu. From rssi to csi: Indoor localization via channel response. *ACM Computing Surveys*, 46(2):1–32, 2013.

[ZCW23] X. Zheng, R. Cheng, and Y. Wang. Rssi-knn: A rssi indoor localization approach with knn. In *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 600–604, Changchun, China, 2023. IEEE.

[ZLS+21] C. Zhou, J. Liu, M. Sheng, Y. Zheng, and J. Li. Exploiting fingerprint correlation for fingerprint-based indoor localization: A deep learning based approach. *IEEE Transactions on Vehicular Technology*, 70(6):5762–5774, 2021.

[ZYC+24] Zhiyu Zhu, Yang Yang, Mingzhe Chen, Caili Guo, Julian Cheng, and Shuguang Cui. A survey on indoor visible light positioning systems: Fundamentals, applications, and challenges. *arXiv preprint arXiv:2401.13893*, pages 1–34, 2024. Submitted January 25, 2024.