WI - Indoor Localization System Using RSSI

Project Report

Giuliano Crescimbeni, 10712403
Politecnico di Milano

June 2025

1. Introduction

This report outlines the design, implementation, and assessment of a Wi-Fi-based indoor localization framework leveraging Received Signal Strength Indicator (RSSI) metrics. The localization strategy relies on data gathered during a preliminary offline stage, where RSSI measurements from various Wi-Fi access points were recorded at known positions within a predefined indoor layout. These reference points, distributed over a controlled environment, served as anchors for system calibration.

To ensure data consistency, several preprocessing operations were applied, such as eliminating anomalies, block-wise averaging of signal strengths, and assigning spatial identifiers aligned with a two-dimensional grid topology.

The study explores both deterministic and probabilistic localization paradigms. Deterministic strategies—including k-Nearest Neighbors (KNN) and Random Forest—were evaluated alongside probabilistic frameworks such as the Horus algorithm and Bayesian Networks. Performance assessment focused on predictive accuracy and positional estimation errors.

The comparative results provide insight into the strengths and limitations of each algorithm in the context of RSSI-based indoor positioning.

2. Data Acquisition - Offline Phase

To construct and evaluate the localization model, signal data was systematically collected during an offline acquisition phase. The experimental deployment was conducted within the author's personal residence. The house floor plan was segmented into a regular spatial grid, where each cell represented a measurement station, uniquely identified by a coordinate label reflecting its grid location.

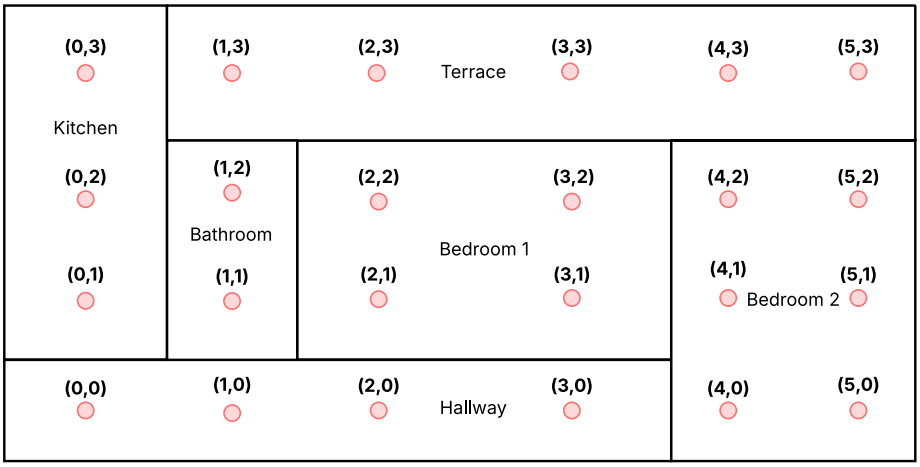

Figure 1: Sampling grid overlaid on the floor plan utilized for data acquisition.

At each designated coordinate, Wi-Fi packets were captured using Wireshark over a fixed interval of ninety seconds. The raw packet captures, initially stored in PCAP format, were later transformed into CSV files to streamline data handling and facilitate analysis.

Subsequent data preparation was carried out through tailored Python scripts developed for this project. The preprocessing pipeline included the following stages:

- **Data Cleaning**: Eliminated malformed or incomplete entries and harmonized MAC address formatting.
- **Dummy Entry Insertion**: Synthetic entries with low RSSI values were added to account for zones with zero signal reception, simulating realistic absence of access point coverage.

- **Data Augmentation**: Augmented the dataset through controlled resampling of existing measurements to mitigate class imbalance and enhance generalization capacity.

These processing steps contributed to a more coherent and statistically robust dataset, laying a solid foundation for the supervised learning procedures that followed.

The system selected four MAC addresses for analysis—those that appeared most consistently across the entire grid. The corresponding packet intensity distributions are visualized in the following heatmaps:

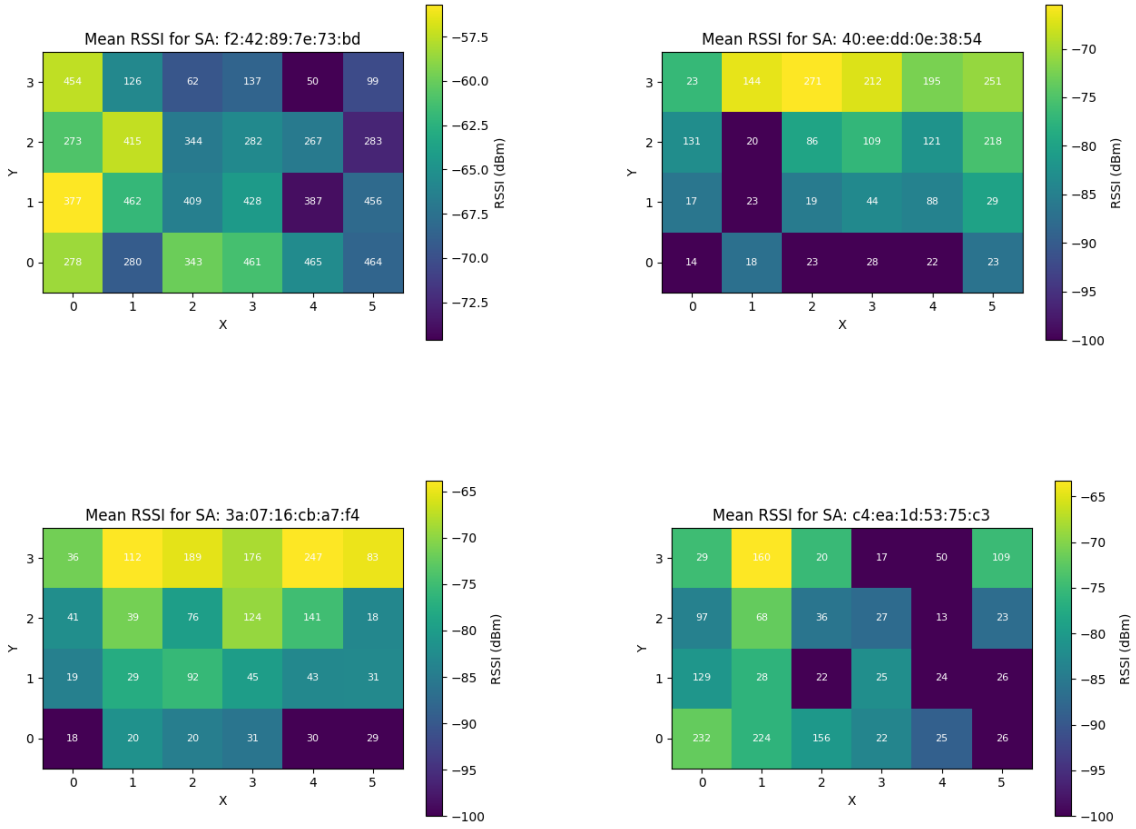

Figure 2: Packet occurrence heatmaps for the four most prevalent MAC addresses across the sampling grid.

3. Machine Learning Algorithms

To benchmark the effectiveness of various approaches for indoor positioning, four machine learning models were selected. These models were grouped according to their underlying methodology:

- **Probabilistic Models:**
  - **Horus:** Relies on statistical signal distribution profiles and uses a distance-weighted inference mechanism.
  - **Bayesian Network:** Encodes the conditional dependencies between signal features and location via a probabilistic graphical structure. Parameters were estimated using Maximum Likelihood, and inference was carried out through variable elimination.
- **Deterministic Models:**
  - **K-Nearest Neighbors (KNN):** A non-parametric classifier that assigns labels by voting among the k most similar training instances in the RSSI space.
  - **Random Forest:** An ensemble approach aggregating multiple decision trees to produce the most frequent classification outcome.

For every method, the dataset was partitioned using a consistent protocol: 70% of the data served for training, while 30% was allocated to validation. This split ensured uniform evaluation criteria across all experiments.

4. Experimental Results

The results below summarize the predictive behavior of the four localization techniques. Each subsection reports the confusion matrix, accuracy score, and mean spatial error (in meters) obtained on the test set.

4.1. Horus

Horus estimates position based on probabilistic similarity, assigning weights inversely proportional to the Euclidean distance from known RSSI samples.

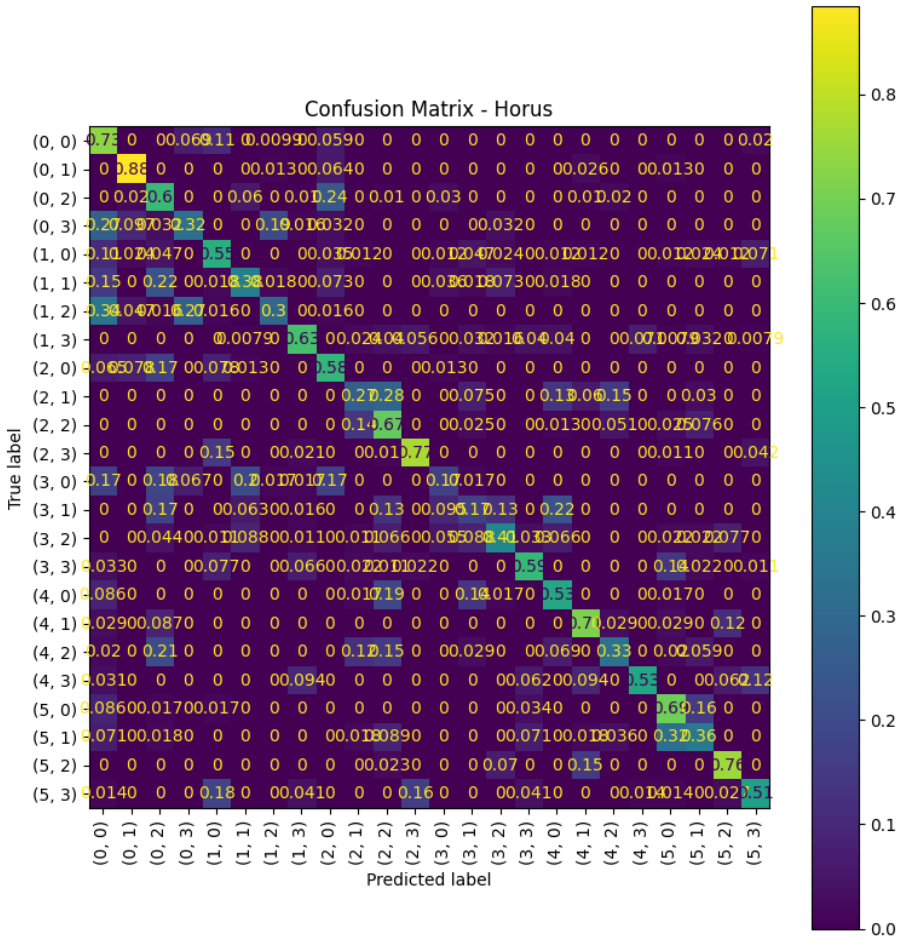

Figure 3: Confusion Matrix - Horus

Accuracy: 0.5377
Average Localization Error: 1.07 m

4.2. Bayesian Network

After discretizing signal values into quartile bins, a Bayesian Network was trained to model the joint probability distribution. Inference was performed using variable elimination.

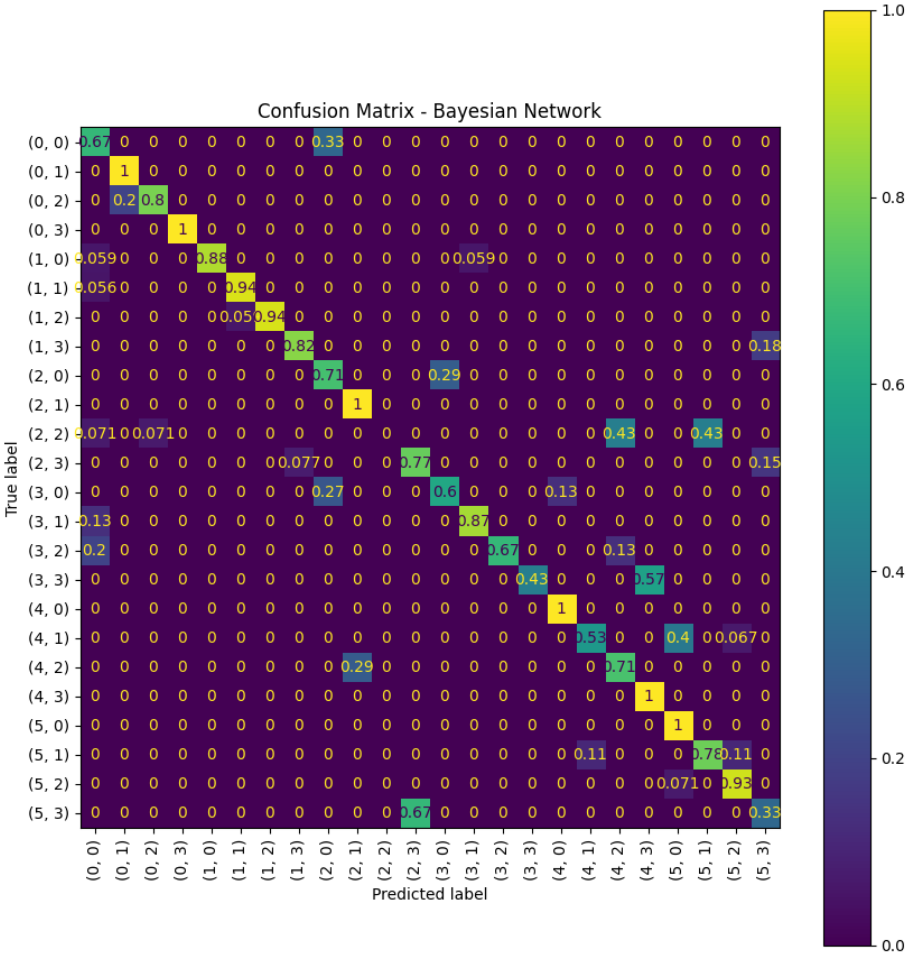

Figure 4: Confusion Matrix - Bayesian Network

Accuracy: 0.7722
Mean Localization Error: 0.46 m

4.3. K-Nearest Neighbors (KNN)

The KNN classifier was evaluated across a range of k values (1 to 20), with the optimal value chosen based on the highest observed classification accuracy.

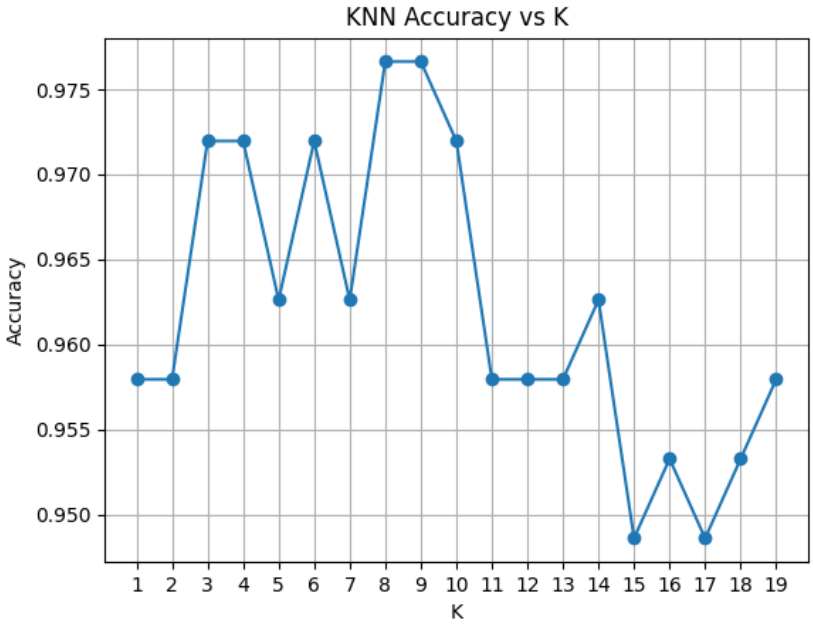

Figure 5: Accuracy vs. K for KNN algorithm

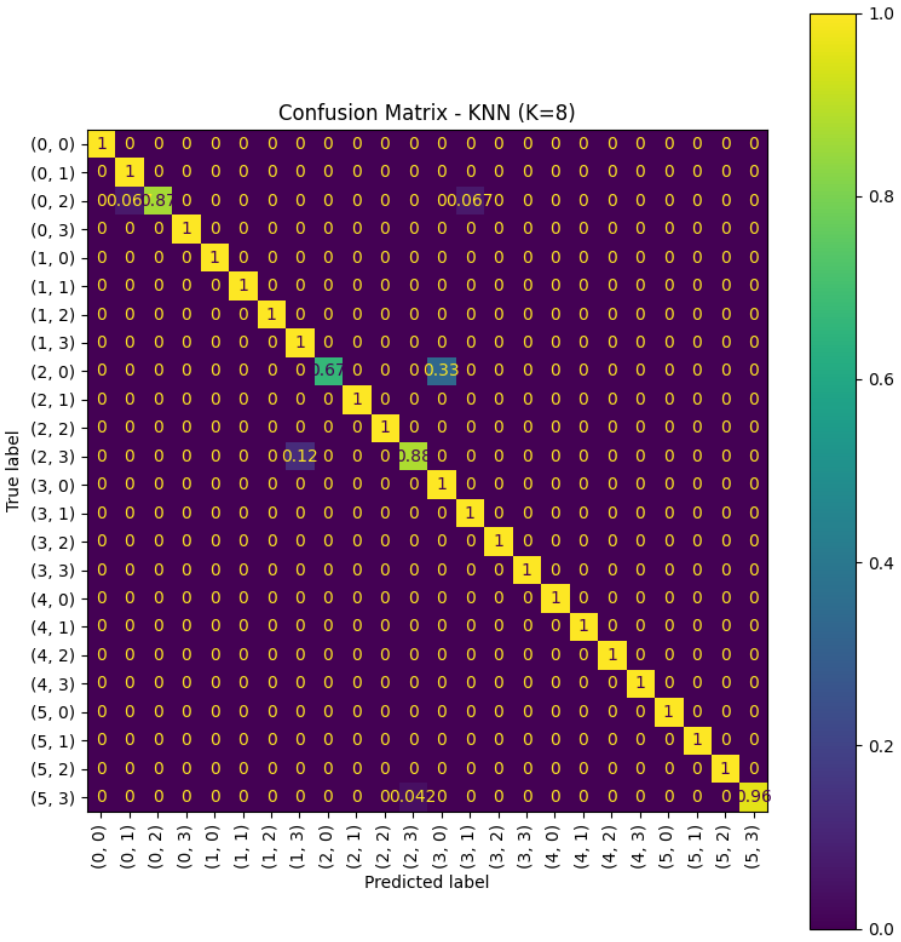

Figure 6: Confusion Matrix - KNN

Accuracy: 0.9766
Mean Localization Error: 0.09 m

4.4. Random Forest

A Random Forest model was trained with default hyperparameters. The classification results are visualized below.

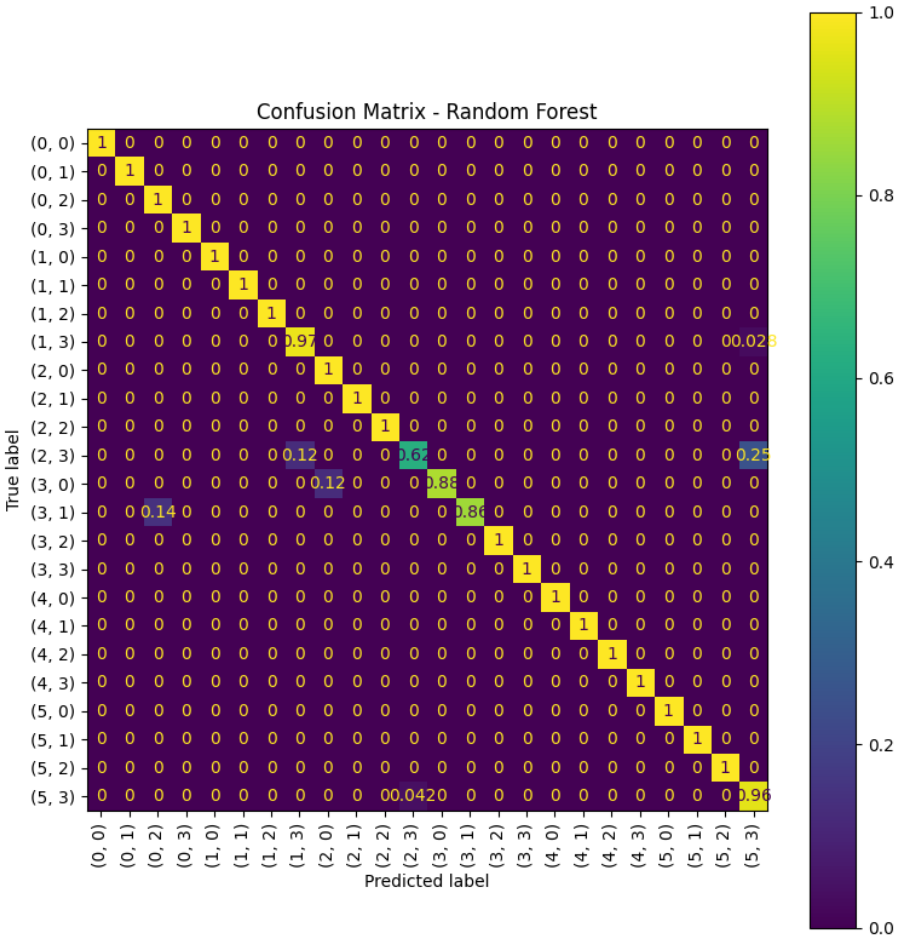

Figure 7: Confusion Matrix - Random Forest

Accuracy: 0.9672
Mean Localization Error: 0.08 m

5. Discussion

Table 1 consolidates the empirical findings for each evaluated model, reporting accuracy and average localization error.

Table 1: Performance Summary of Localization Algorithms

Algorithm	**Accuracy**	**Mean Localization Error (m)**
Horus	0.5377	1.07
Bayesian Network	0.7722	0.46
K-Nearest Neighbors	**0.9766**	**0.09**
Random Forest	0.9672	0.08

The empirical results highlight a marked contrast in effectiveness between simpler distance-based models and more structured probabilistic methods. Both KNN and Random Forest achieved remarkably high accuracy and sub-meter localization errors, suggesting their strong compatibility with the features derived from the RSSI dataset.

The Bayesian Network yielded respectable results, showing robustness despite the discretization process applied to signal values. Its strength lies in capturing probabilistic relationships, though its granularity is inherently limited compared to continuous approaches.

Horus showed the lowest performance among the group. While rooted in solid probabilistic principles, the algorithm may have struggled due to insufficient signal differentiation between neighboring positions and the influence of spatial noise on signal distributions.

In summary, deterministic models like KNN and Random Forest proved exceptionally effective for the considered scenario, delivering both precision and reliability. Probabilistic techniques, while offering valuable interpretability, may require richer datasets to compete at the same level.

6. Conclusion

The investigation confirms that algorithm selection plays a pivotal role in the effectiveness of indoor localization systems. In structured environments with well-processed signal data, deterministic approaches such as KNN or Random Forest may yield superior results. However, probabilistic models retain value when interpretability or uncertainty modeling is required.