Indoor Localization: Predicting Position and Floor from RSSI Fingerprints

Francesca Sacchetti
MS in Data Science
Fordham University
New York, NY, USA
fsacchetti@fordham.edu

***Abstract*— Indoor localization is increasingly important for navigation, safety, and many smart-building applications. This project explores WiFi fingerprinting using the UJIIndoorLoc dataset to predict building, floor, and spatial coordinates from WiFi Received Signal Strength Indicator (RSSI) features. After preprocessing 520+ WiFi access point signals, multiple machine learning models were trained for both classification and regression tasks, including K-Nearest Neighbors (KNN), Random Forests, and XGBoost. Dimensionality reduction via Principal Component Analysis (PCA) was applied to improve speed and reduce potential overfitting. Results show near-perfect performance for building classification (up to 100% accuracy) and strong performance for floor classification (82% – 88%). Coordinate regression achieved low mean absolute error (MAE) (Longitude MAE ≈ 6.46 – 7.45, Latitude MAE ≈ 5.91 – 7.35, Floor MAE ≈ 0.22 – 0.36). These results highlight the effectiveness of classical machine learning models for WiFi localization and demonstrate that the dataset provides highly separable signal patterns. Future work includes exploring neural networks, converting coordinates to real-world meter units, and validating models on external environments.**

Keywords*— *Indoor localization, WiFi fingerprinting, RSSI, machine learning, PCA.

I. Introduction

Indoor localization has become an important component of modern navigation systems, smart-building infrastructure, emergency response, and location services. While Global Positioning System (GPS) technology performs well outdoors, its signal worsens significantly within buildings, making it unreliable and often unusable for indoor positioning. Comparatively, alternative sensing methods—including WiFi—have been widely explored to enable accurate and reliable indoor location estimation.

Among these alternative methods, WiFi fingerprinting has become one of the most practical and cost-efficient approaches. Its practicality primarily lies in the fact that most buildings or indoor environments already having numerous WiFi access points (APs), and modern devices can easily record their received signal strength indicator (RSSI) values. A WiFi fingerprint is defined as the vector of RSSI measurements observed at a particular location. Machine learning models are able to learn the relationship between these fingerprints and spatial coordinates, allowing for the prediction of a device's location. With this machine learning approach, no special hardware is needed.

This project uses a publicly available RSSI dataset, UJIIndoorLoc dataset, and it is large and well-known dataset for WiFi-based indoor localization research. The dataset contains over 19,000 samples, collected across three buildings, multiple floors, with more than 520 different WiFi access points. Each sample includes RSSI readings (ranging from –104 dBm to 100, where 100 indicates "no signal"), along with the true building ID, floor number, longitude, and latitude. Given the size of the dataset, a dimensionality challenge is presented due to the high-dimensional structure including substantial noise and sparsity.

Machine learning is a reliable approach to solve the challenge of indoor WiFi localization because RSSI data can present complex, nonlinear patterns that can be affected by a multitude of factors. Whether it be distance, physical obstacles, or device variability, the RSSI signals will change as a result. Prior research has been completed and shown the effectiveness of K-Nearest Neighbors (KNN), Random Forests, and Gradient-Boosted Trees (e.g., XGBoost) in capturing these complex patterns accurately and efficiently when also combined with feature and dimensionality reduction techniques.

The primary objective of this project is to build a complete machine learning pipeline for indoor localization using the WiFi fingerprints. Specifically, this project predicts the building in which a fingerprint was recorded, the floor level, and the exact coordinates (longitude and latitude). The project consists of data preprocessing, dimensionality reduction, model comparison across classification and regression tasks, and also a quantitative analysis of overall performance. The results show that properly tuned traditional machine learning models are able to deliver accurate and efficient indoor localization.

II. Literature Review

A. Baseline Modeling

WiFi fingerprinting has been widely explored as a practical solution for indoor localization due to its low deployment cost and the availability of RSSI measurements on consumer devices. Early approaches, such as the RADAR system by Bahl and Padmanabhan, demonstrated the effectiveness of K-Nearest Neighbors (KNN) for matching WiFi fingerprints to known locations [2]. KNN remains a strong baseline due to its

simplicity, although it can struggle with the high-dimensional and noisy RSSI data.

B. Tree-Based Ensemble Models

To address the challenge of the noisy data, later research introduced tree-based ensemble methods, including Random Forests and Gradient Boosted Trees (e.g., XGBoost). These models are able to capture and learn nonlinear relationships and perform implicit feature selection, making them effective for environments with complex signals. More recent studies have also explored deep learning models, such as autoencoders and CNN-based RSSI embeddings.

In addition to classification-based approaches, regression-based localization has also been widely studied for estimating continuous spatial coordinates. These methods aim to directly predict longitude and latitude values rather than mapping fingerprints to discrete locations. In the past, tree-based ensemble models have performed particularly well for this task due to their robustness to noise and ability to capture the nonlinear signal relationships.

C. High-Dimensional Data

A common challenge in WiFi fingerprinting research is the high dimensionality and sparsity of RSSI features, as many access points are not detectable at all locations. Recent work has also explored RSSI indoor localization using alternative wireless technologies such as Bluetooth, emphasizing lightweight models suitable for real-time deployment on edge devices in obstructed indoor environments [3].

Prior studies have shown that dimensionality reduction techniques such as Principal Component Analysis (PCA) can significantly improve performance by reducing noise, removing redundant features, and improving overall computational efficiency. PCA has also been shown to enhance generalization performance, especially when used in combination with distance-based models such as KNN.

D. Deep Learning Approaches

While deep learning approaches, including autoencoders and convolutional neural networks, have demonstrated promising results in recent work, these methods often require significantly larger datasets and greater computational resources. As a result, classical machine learning methods remain highly successful and competitive, particularly in scenarios where interpretability, training efficiency, and robustness are prioritized.

E. Model Comparison Conditions

Despite the increase in WiFi fingerprinting research, many studies are evaluating their models under different pre-processing pipelines, feature spaces, and performance metrics, making direct comparison quite difficult. Whether it be how much variance is being preserved in dimensionality reduction steps or the method of choice for dealing with missing values, differences in small but important tasks throughout modeling can directly hinder or improve the results. There remains value in model evaluations that apply consistent preprocessing and directly compare multiple machine learning approaches on the same benchmark dataset. This work contributes to that objective by providing a traceable yet experimental framework for both regression and classification on identical feature engineered and validation conditions.

III. Methodology

This methodology follows a structured pipeline consisting of data pre-processing, feature reduction, initial baseline model construction, generalizability assessments, final model training, and performance evaluation. Each stage was designed thoughtfully to address the challenges that come along with WiFi RSSI data. By applying consistent pre-processing steps and evaluating the models under identical conditions, the methods of this project allow for systematic comparison of the model performance for both the regression and classification tasks.

A. Dataset Overview

The dataset includes:

- Approximately 19,000 training samples
- 520 WiFi Access Point (WAP) RSSI features
- True Labels for BuildingID, Floor, Longitude, and Latitude

The UJIIndoorLoc dataset was selected due to its large size and diverse signals. It spans multiple buildings and floors and includes a large number of access points, making it representative of complex indoor WiFi environments. While challenging to train accurate machine learning models on, the dataset's high dimensionality and sparsity reflect realistic deployment conditions where many access points are not detectable at every location. All of these noted characteristics make the dataset a strong choice for evaluating both the robustness and generalization ability of different machine learning models.

B. Preprocessing

RSSI values of 100, which indicate no detected signal, were replaced with −105 to represent an extremely weak but still plausible signal. This change prevents models from incorrectly interpreting missing signals as large positive values and ensures consistency across all of the learning algorithms. All WAP features were standardized using StandardScaler to ensure consistent scaling across all of the access points. A variance threshold filter was applied to remove WAPs that never actually appeared in the dataset, reducing substantial noise and redundant features. Principal Component Analysis (PCA) was used for dimensionality reduction.

The order in which these pre-processing steps were applied is just as important as the steps themselves. The standardization step was applied prior to the dimensionality reduction step to ensure that all RSSI features contributed equally to distance based training and variance estimations. Without the standardization step, WAPs with larger absolute RSSI ranges may have disproportionately influenced model training. Variance thresholding was done before the PCA dimensionality reduction as an additional simplifying measure.

Multiple PCA configurations were tested by varying the number of retained components and analyzing the corresponding explained variance metric. Based on the analysis,

200 principal components were selected, preserving 89% of the total variance while significantly reducing the overall feature space. Choosing 200 principal components provided a strong trade-off between dimensionality reduction, computational efficiency, and predictive performance.

The preprocessing pipeline reduces noise, mitigates sparsity, and produces a much more compact feature representation that is proven to be effective for both distance-based and tree-based machine learning models.

C. Initial Baseline Modeling

Initial baseline modeling was done to ensure that the dataset contained predictive signals, and for initial exploratory purposes. K-Nearest Neighbors Regressor was used as an initial regression baseline approach, and Random Forest was used as the initial classification approach. The initial baseline models were assessed mainly for generalizability and proper functioning. In addition, cross-validation on initial baseline models was applied to assess the overall generalization abilities before the final modeling took place.

D. Final Models Evaluated

1) Classification Models:

- Random Forest Classifier
- K-Nearest Neighbors (KNN) Classifier
- XGBoost Classifier

2) Regression Models:

- K-Nearest Neighbors (KNN) Regressor
- Random Forest Regressor
- XGBoost Regressor

Classification models were used to predict BuildingID and Floor, representing discrete localization tasks. Regression models were used to predict continuous spatial coordinates, including Longitude, Latitude, and Floor Height.

The selected models represent a range of classic machine learning approaches, including distance-based methods (KNN) and tree-based ensemble techniques (Random Forest and XGBoost). This combination enables a systematic comparison of model behavior on the high-dimensional and noisy RSSI features.

KNN was selected as an initial baseline due to its widespread use in WiFi fingerprinting research and given that it has an intuitive distance based interpretation. Random Forest models were chosen for their ability to handle high-dimensional data and its robustness to noise. Ensemble averaging also aids in resisting overfitting which can be a large concern. Lastly, XGBoost was chosen due to its ability to model complex nonlinear relationships effectively. At a holistic perspective, all of these models provide a balance comparison between interpretability and predictive power for both the regression and classification tasks.

All of the models were trained on the preprocessed training data and evaluated on the isolated validation dataset.

E. Evaluation Metrics

Model performance was evaluated using task-appropriate metrics for both classification and regression problems. Accuracy was used to evaluate BuildingID and Floor classification, as these tasks involve discrete label prediction and class distributions were not severely imbalanced. Mean Absolute Error (MAE) was used for Longitude, Latitude, Floor regression.

Accuracy was selected as a main evaluation metric for classification tasks because correct identification of building and floor represents discrete localization success and is easily interpretable in practical deployment scenarios. Precision, recall, and F-1 scores were also examined in baseline modeling to ensure accuracy results could be trusted. MAE was selected to evaluate the regression tasks because it provides a direct measure of the average localization error and ensures that all errors are treated equally. This is important because both small and moderate positioning errors are of the same relevance for this project. In comparison to squared-error metrics, MAE is less sensitive to outliers and allows for intuitive interpretation in the context of the spatial localization problem.

Cross-validation was used as a reliability check to assess the initial exploratory baseline model stability on the training data. The primary and final model performance results reported in this paper are based on evaluation on a held-out validation set to avoid optimistic bias.

IV. Final Model Results

A. Classification Performance

TABLE I. BUILDING CLASSIFICATION ACCURACY

Model	Accuracy
Random Forest	100%
XGBoost	99.7%
KNN	99.1%

TABLE II. FLOOR CLASSIFICATION ACCURACY

Model	Accuracy
XGBoost	87.9%
Random Forest	84.6%
KNN	82.3%

All evaluated models achieved high accuracy in predicting the BuildingID. Random Forest achieved perfect classification accuracy, while XGBoost and KNN also exceeded 99%. These near-perfect results indicate that WiFi signal fingerprints within the dataset are highly distinct across different buildings, resulting in strong separability in the feature space.

The classification models demonstrated highly consistent performance across the validation dataset, indicating strong model generalizability. The negligible performance gap between Random Forest, XGBoost, and KNN suggests that the building classification is largely influenced by the presence or absence of distinct access points instead of model complexity.

Floor classification proved to be more challenging than building classification. While XGBoost achieved the highest accuracy at 87.9%, Random Forest performed second best with an accuracy of 84.6%, and KNN performed the worst with an accuracy of 82.3%. Overall, performance was lower across all models compared to the building task. This reduction is expected due to increased signal similarity between vertically adjacent floors and larger RSSI variability. Additionally, floor prediction displayed larger sensitivity to model choice than building classification. Despite these challenges, the results still demonstrate satisfactory floor-level localization performance using the WiFi fingerprint data.

B. Regression Performance (Coordinates)

TABLE III. REGRESSION PERFROMANCE (MAE)

Model	MAE Longitude	MAE Latitude	MAE Floor
KNN	7.45	7.35	0.22
Random Forest	6.46	5.91	0.36
XGBoost	6.72	5.96	0.24

Random Forest achieved the lowest error for longitude and latitude prediction, indicating the strongest overall performance for spatial coordinate estimation. XGBoost produced comparable results, while KNN exhibited higher coordinate errors but achieved the lowest floor regression error.

While individual predictions may exhibit larger errors, the average localization error remained within 6–7 coordinate units and approximately 0.22–0.36 floors across models, demonstrating strong localization performance given the complexity and noise inherent in WiFi RSSI signals.

Scatter plots of XGBoost predicted versus true coordinates further confirm a strong linear relationship between model predictions and ground-truth values, indicating effective learning of spatial patterns. The concentration of points along the diagonal in Fig. 1 and Fig. 2 demonstrates strong accuracy between predicted and ground-truth coordinates. The linear alignment confirms that the regression models successfully learned the underlying relationships and structure of the WiFi fingerprint data. While Random Forest achieved the lowest error, XGBoost offers a strong balance between accuracy and robustness for the multi output localization task.

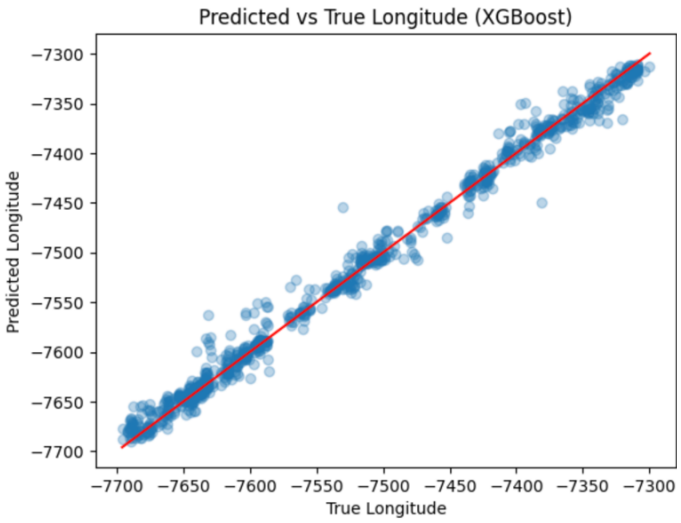

Fig. 1. Predicted versus true longitude values for the XGBoost regression model. The red diagonal line represents perfect prediction (y = x). (*figure caption*)

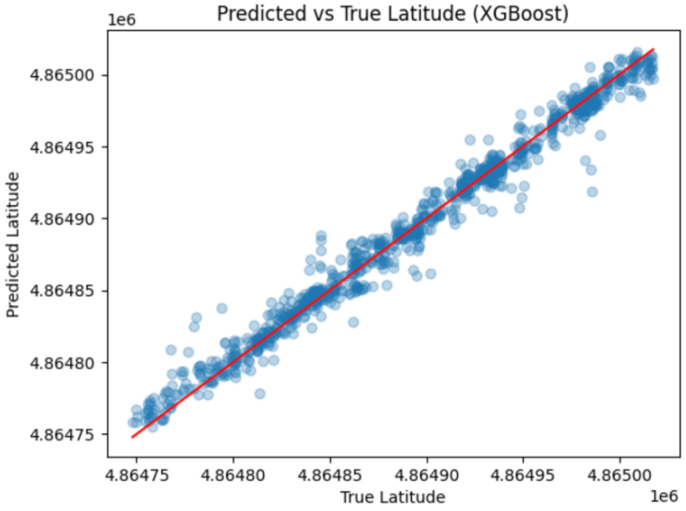

Fig. 2. Predicted versus true latitude values for the XGBoost regression model. The red diagonal line represents perfect prediction (y = x). (*figure caption*)

V. DISCUSSION

A. Strong Performance Across All Models

The results demonstrate that WiFi RSSI fingerprints provide strong predictive power for indoor localization tasks, particularly at the building level. The near-perfect building classification accuracy observed across all models can be attributed to the highly distinct access point distributions present in each building. Because different buildings contain largely non-overlapping sets of WiFi access points, the resulting RSSI fingerprints are highly separable in feature space, making building identification a relatively straightforward classification task. As a result, even relatively simple models such as KNN are able to achieve a strong accuracy score while ensemble based methods such as Random Forest and XGBoost further reinforce

the stable classification through averaging and their ability to handle nonlinear decision boundaries.

Random Forest achieving 100% accuracy requires attention and presents a possible red flag because 100% accuracy can often point to overfitting, data leakage, or unusually separable classification tasks. Overfitting is unlikely given that all models performed similarly exceptional in accuracy. Data leakage was not found as pre-processing was fit only on training data, and the validation data was transformed using the learned parameters. It appears that the underlying cause may be rooted in the concept of building classification being inherently easy within the UJIIndoorLoc dataset. The decision boundary may be as easy boundary to learn.

In contrast, floor classification proved to be more challenging. Although performance remained strong, accuracy was consistently lower than that of building classification across all of the models. This outcome is expected due to the physical properties of WiFi signal propagation, which exhibits weaker vertical gradients compared to horizontal separation. Signals from access points can often penetrate multiple floors, resulting in overlapping RSSI patterns between vertically adjacent floors. Additionally, the dataset exhibits moderate class imbalance at the floor level, particularly higher floors have fewer samples, which further contributes to classification difficulty. Among all of the evaluated models, XGBoost achieved the highest floor accuracy, suggesting that the gradient boosted ensembles may be better suited for capturing the subtle complexities and variations associated with vertical positioning.

Regression results for longitude, latitude, and floor height indicate low average localization error despite the indoor WiFi signals noise and variability. Achieving MAE values of approximately 6–7 coordinate units and up to 0.36 floors demonstrates that the models were able to learn the spatial relationships from the high-dimensional RSSI data. While individual predictions may occasionally exhibit larger errors, the overall error distribution remains closely concentrated, reflecting strong average localization performance.

Differences in regression performance across models highlight some notable trade-offs. Random Forest achieved the lowest error for its latitude and longitude prediction, likely due to its ability to reduce variance through ensemble averaging. XGBoost produced comparable results while offering improved modeling of the nonlinear interactions. These findings emphasize that model selection should not just consider the target variable but also search for the desired balance between bias and variance.

B. Importance of Principal Component Analysis

Dimensionality reduction using Principal Component Analysis (PCA) played a critical role in achieving these results. By reducing the feature space from hundreds of noisy and sparse RSSI dimensions to a smaller set of principal components while preserving approximately 89% of the variance, PCA was able to mitigate the dimensionality curse that is often inherent to WiFi fingerprinting. RSSI measurements can be highly correlated across access points due to environmental structure and signal propagation effects, and PCA is able to effectively capture any correlations while minimizing redundant and noisy dimensions.

The reduction to 200 principal components improved model generalization and reduced overfitting, particularly for distance-based methods such as KNN. The tree-based ensemble models, Random Forest and XGBoost, also benefited from this reduced feature space and noise, enabling robust learning of nonlinear signal relationships.

C. Evaluation Trade-offs and Practical Considerations

While accuracy was used as the primary evaluation metric for the classification tasks, it represents only one part of performance in a practical indoor localization system. In real-world applications, trade-offs between accuracy, computational complexity, scalability, and efficiency all must be carefully examined depending on the context of the application. For example, robotics or real-time navigation systems may prioritize lightweight models and minimal lag, while emergency response application systems would most likely require higher precision at the expense of a higher computational cost [4]. While one trade-off is not universally optimal, the appropriate performance metric balance and choices depend on the context of the application.

Accuracy was selected in this work to establish a consistent baseline comparison across models under identical conditions. Future evaluations should expand on this baseline approach, incorporating additional metrics such as precision and recall to better reflect operational constraints and requirements of a real indoor localization environment.

The use of cross-validation further supports the reliability of the reported results. There was consistent performance across the folds, indicating that the baseline models were able to generalize strongly across different subsets of the RSSI fingerprint data. While cross-validation provided additional insight into model consistency, final model performance and comparisons are reported using the separate validation dataset. As a result, validation-set performance is emphasized as the most reliable indicator of real-world generalization, ensuring that metrics reflect generalization to unseen data.

D. Simple Architecture Performing Well

Overall, this discussion exposes that pre-processing, dimensionality reduction, and model selection were all essential steps to perform effective indoor localization. The strong model performance highlights that, while modern deep learning approaches are growing in popularity and do have their designated use cases, complex architectures are not always necessary to obtain comparable and strong results.

VI. Future Work

A. Estimating Physical Distances and Numerical Conversion

Several opportunities present themselves to improve the framework completed for this project. One important direction for future work involves converting predicted longitude and latitude coordinates into estimated physical distances such as meters. A conversion like this would allow localization error to be interpreted in real-world spatial terms, making the results more intuitive for practical deployment. It would also allow for direct comparison with many other indoor positioning systems that use the same units. In the long run, the conversion may also facilitate the integration with indoor navigation applications.

This conversion would require mapping longitude and latitude to a local coordinate system using known building reference points or a floor plan. One challenge in this additional process may be ensuring consistency across buildings that have different coordinate origins or floor plans, highlighting the importance of standardized data in indoor localization research.

B. Exploring Advanced Learning Methods

In addition to coordinate transformation, more advanced learning methods could be explored to continue to improve localization accuracy. In particular, neural networks or transformer-based models could be used to learn compact RSSI embeddings that better capture nonlinear signal relationships. These models may be especially effective in environments with dense WiFi infrastructure, where interactions between multiple signals are more common and complicated. Deep learning approaches could also utilize representation learning to reduce the reliance on the manual feature engineering and PCA steps.

Choosing either advanced learning methods or classical machine learning methods is not required, as a hybrid approach may offer the most ideal balance between performance and efficiency. Both levels of learning offer possible project improvements.

C. Class Imbalance Exploration

While floor classification performance was already strong, additional improvements may be achieved by explicitly addressing class imbalance within the dataset. Applying class weighting or cost-sensitive learning techniques could help reduce bias toward more frequently observed floors and improve the model's performance on underrepresented floors. Other strategies such as hierarchical classification or multi-task learning may also improve the vertical localization accuracy by jointly modeling the building, floor, and coordinate predictions.

Future analysis can investigate the impact of dataset imbalances on misclassification patterns, particularly between adjacent floors. Evaluating confusion matrices and per-class recall can provide a deeper insight into which floors are being frequently wrongly predicted and why. These additions may improve vertical localization reliability.

D. Additional External Validation

Lastly, validating the trained models on additional external datasets or unseen buildings would provide a stronger assessment of generalization capability across different indoor environments. Although the models performed well on the dataset selected for this project, testing on additional environments would provide stronger evidence of model generalizability for different building layouts, access point configurations, and signal conditions. These factors commonly occur in real-world deployments but are difficult to capture within a single dataset. Additional validation is necessary to assess real-world application abilities, and demonstrating consistent performance across a diverse set of environments will significantly strengthen the case for the proposed framework in this project.

VII. Conclusion

This project successfully implemented a complete WiFi-based indoor localization system using classical machine learning techniques. The results showcase accurate performance for building classification, strong accuracy for floor classification, and a low average error for coordinate regression, highlighting the effectiveness of WiFi fingerprinting data for indoor navigation tasks. The high classification accuracy achieved, particularly at the building level, indicates that WiFi signal fingerprints contain predictive information capable of supporting reliable indoor positioning tasks.

The findings confirm that, when paired with appropriate preprocessing, dimensionality reduction, and model selection, machine learning models can reliably learn the spatial patterns from high-dimensional and noisy RSSI data. Replacing missing signal values with physically plausible representations, applying feature scaling, and reducing dimensionality through Principal Component Analysis (PCA) played a critical role in improving both computational costs and model generalization. In particular, tree-based ensemble methods such as Random Forests and XGBoost demonstrated strong robustness to signal variability and noise, while PCA helped mitigate the curse of dimensionality inherent in WiFi fingerprinting datasets.

Beyond the overall performance metrics, this work emphasizes that classical machine learning approaches remain highly competitive for indoor localization problems. Despite the growing interest in deep learning methods, the models used in this project were able to achieve strong results with low computational cost, easier interpretability, and reduced training complexity. The approach taken within this project may be suitable for real-world applications where reliability and efficiency are both important considerations.

The comparison and evaluation of multiple models provides insight into the strengths and limitations of different machine learning approaches for indoor localization. Importantly, the results suggest that no single model universally dominates across all localization tasks; instead, model selection should consider the specific target variables, performance trade-offs, and any deployment constraints.

Overall, the proposed framework of this project can be adapted for more advanced models, additional datasets, and also real-world navigation applications, supporting future projects in WiFi-based positioning systems. By systematically evaluating both classification and regression tasks under consistent preprocessing and validation conditions, this work contributes a practical and reproducible pipeline that supports continued development in WiFi-based positioning systems.

Acknowledgment

The author would like to thank the course instructor for their guidance and support throughout the semester, which greatly contributed to the development and completion of this project. The author also acknowledges the creators of the UJIIndoorLoc dataset for making the data publicly available for research purposes.

References

[1] G. J. Torres-Sospedra et al., "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization," in Proc. Int. Conf. Indoor Positioning and Indoor Navigation (IPIN), 2014.

[2] P. Bahl and V. N. Padmanabhan, “RADAR: An In-Building RF-Based User Location and Tracking System,” in Proc. IEEE INFOCOM, 2000, pp. 775–784.

[3] K. Ok, G. Heo, S. Bae, C. Lee, and S.-H. Ahn, “Bluetooth RSSI-based lightweight indoor robot localization for edge devices in obstructed environments,” IEEE Access, vol. 13, pp. 196236–196246, 2025, doi: 10.1109/ACCESS.2025.3632470.

[4] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.), vol. 37, no. 6, pp. 1067–1080, Nov. 2007, doi: 10.1109/TSMCC.2007.905750.