Indoor positioning system using machine learning

Naser Helal Ali Alrowahi

MSc. Thesis

December 2024

A thesis submitted to Khalifa University of Science and Technology in accordance with the requirements of the degree of MSc. Electrical and Computer engineering in the Department of Electrical Engineering.

Indoor positioning system using machine learning

by

Naser Helal Ali Alrowahi

A thesis submitted in partial fulfilment of the requirements for the degree of

MSc. Electrical and Computer Engineering

at

Khalifa University

Dr. Nazar Ali (Main Advisor),
Associate Professor
Khalifa University

Dr. Ahmed Al Tunaiji (Co-Advisor)
Assistant Professor
Khalifa University

December 2024

Abstract

Naser Helal Alrowahi, "**Indoor positioning system Using Machine Learning**", M.Sc. Thesis, MSc.electrical and computer engineering, Department of Electrical and computer Engineering, Khalifa University of Science and Technology, United Arab Emirates, December 2024.

Indoor positioning has garnered significant attention from researchers and developers due to the rapid advancement of technology and artificial intelligence. However, this field still encounters numerous challenges and limitations, particularly in achieving a balance between efficiency and accuracy.

Indoor positioning systems significantly benefit various sectors, such as healthcare, smart homes, and smart transportation systems. Nonetheless, the intricate and ever-changing infrastructure of the indoor environment necessitates continuous improvement and development.

The indoor environment is rich in signals, including Bluetooth and WiFi, which can be detected using different techniques such as received signal strength (RSS) and channel state information (CSI). Each technique possesses unique features that can be extracted. Through the utilization of machine learning algorithms like support vector machine, decision tree, and other techniques, as well as more advanced methods such as deep learning, the system can extract features and establish relationships while accounting for noise and signal distortion to achieve indoor positioning and environment identification.

The accuracy of an indoor positioning system hinges on the quality and quantity of data, as well as how the data is processed and presented to enhance the feature extraction process. This paper will discuss different localization technologies and techniques, explore various artificial intelligence approaches, and examine different preprocessing techniques to enhance the output of indoor positioning systems.

Indexing Terms: Indoor Positioning Systems (IPS), Machine Learning Algorithms, Received Signal Strength (RSS), Channel State Information (CSI), Deep Learning for Indoor Localization.

Acknowledgment

I am thankful to my Creator, all mighty Allah, to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents, my wife, my little girl, or any other individual, was Your will, so indeed, none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout every department of my life.

I would also like to express special thanks to my Advisor **Dr. Nazar Ali** and Co-advisor **Dr. Ahmed Al Tunaiji** for their help throughout my thesis and for the courses which they taught me.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Declaration and Copyright

Declaration

I declare that the work in this thesis was carried out in accordance with the regulations of Khalifa University of Science and Technology. The work is entirely my own except where indicated by special reference in the text. Any views expressed in the thesis are those of the author and in no way represent those of Khalifa University of Science and Technology. No part of the thesis has been presented to any other university for any degree.

Author Name: Naser Helal Alrowahi

Author Signature: ______________________________

Date: 13/12/2024

Copyright ©

CONTENTS

List of Figures

List of Tables

List of Abbreviations

Artificial intelligence	AI
Internet of Things (IoT)	IoT
Line of Sight	LOS
No-Line of Sight	NLOS
Light Emitting Diode	LED
Angle of arrival	AOA
Decision Tree	DT
Support vector machine	SVM
K-nearest neighbor	KNN
Weighted K-Nearest Neighbor	WKNN
Deep Learning	DL
Artificial Neural Network	ANN
Convolutional neural network	CNN
Received Signal Strength	RSS
Deep Neural Network	DNN
Principal Component Analysis	PCA
T-distributed stochastic neighbour embedding	T-SNE
Root Mean Square Error	RMSE

Chapter 1: Introduction

This chapter overviews localization as a trending topic that attracts researchers and developers. It highlights the motivation for this research, the problem statement, and the methodology; at the end of this chapter, the thesis outline will be defined.

1.1 Motivation of research

Nowadays, people live a different life than they used to 10 years ago; the technology wheel is running so fast day by day, which impacts the lifestyle of humans and their needs. In 2024, there will be 7.1 billion smartphone users, representing approximately 4.4% more than the previous year [1]. Many people these days own smart devices such as tablets, laptops, smart watches, smart rings, etc. According to [2], 90% of people in the USA spend their time in the indoor environment, either at work, at home, or shopping.

As smart cities, the Internet of Things (IoT), and 5G continue to advance, the reliance on smart devices for daily tasks is increasing. Indoor localization, a key application of this trend, allows people to navigate and track objects within indoor spaces.

Defining or predicting the coordination of a user or a device in an outdoor or indoor area can be defined as outdoor/ indoor localization. For outdoor localization, the signal comes from the state in a space, which leverages the cellular network and depends on the direct signal from the state or cellular towers [3].

Indoor localization can use the same concept as outdoor localization; however, instead of having a satellite or cellular tower, there are signal emitters such as access points that provide connectivity to the users in that area, which can be considered a wireless-based technique; also there are many other techniques such as optical-based localization, which used the picture taken from that environment to estimate the location based on the visual information and by applying computational processing technique [4].

Indoor localization plays a significant role in enhancing many sectors, such as health care, where patients can be defined and tracked during pandemics and medicine can be served without having person-to-person interaction, reducing the risk to the medical crew. The technology can also be used to enhance the experience of people in a shopping mall where they can navigate to reach the product they are looking for, which will reduce the time needed for shopping, and parents can use it to track their kids inside the mall. This technology can support rescue operations for emergencies such as earthquakes or conflicts by tracking them from the smart devices they wear or have, reducing the time and saving more lives.

The indoor environment has unique dynamic circumstances, which may affect the technique for indoor localization.

Figure 1: Sectors for indoor localization application

1.2 Problem statement and objectives

Artificial intelligence (AI) is a trending topic used in many sectors to fill the gaps limited to classical technologies and applications. In other words, AI has been created to enhance and improve the capability of old technology applications in terms of complexity, behaviour, efficiency, etc.

Considering the conditions of the indoor environment, AI may reduce the application's complexity and increase the accuracy of location estimation. Work done in [5 to 8] will be a kick-off point.

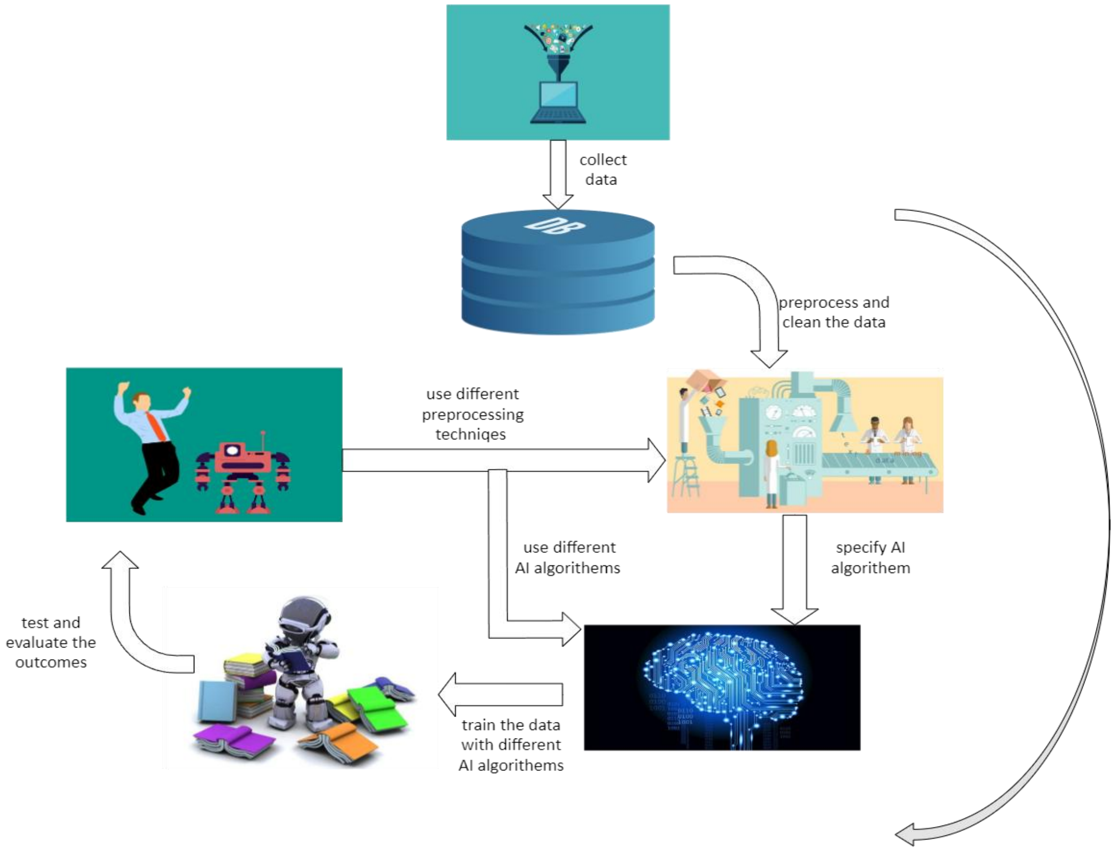

Figure 2: Process to build AI Applications

A comparison between different machine learning algorithms and different preprocessing techniques will be tested and discussed to reach a point where the application can be considered efficient, considering the condition of the indoor environment and whether there is a line of sight (LOS) or no-line of sight (NLOS).

1.3 Research challenges and possible approach

The indoor environment has unique circumstances that make applying positioning inside buildings challenging. However, nowadays, most indoor environments have access points that connect people to the internet. An access point emits a signal (transmitter) received by users' devices (receivers).

The signal travels through a medium to reach the receiver, consisting of obstacles of different sizes, materials, and shapes. These will affect and distort the signal in many ways; since the signal travels in an indoor environment (closed box), it will keep reflecting on the walls.

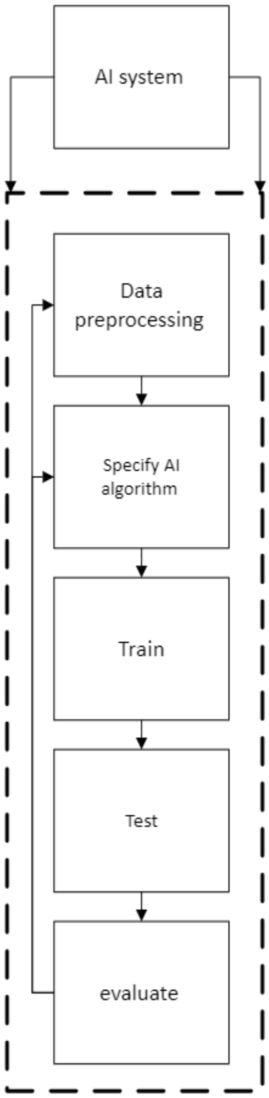

Figure 3: Possible approach

During this project, the signal behaviour will be studied and considered to minimize the effect of signal distortion, reducing the impact of the localization process.

Different indoor environments will be considered, which are classified as the following: high-cluttered, medium-cluttered, low-cluttered, and open space. Each environment will be divided into grids, and each grid point will have sweep iteration in different frequency steps. The data structure will be discussed later. The challenge here is the high amount of data and the reduction of the data to speed up the process without affecting system accuracy. Since the system will use an AI algorithm, discovering the optimal algorithm for this type of application is challenging.

1.4 Thesis outline

The paper is structured into six comprehensive chapters. In Chapter One, the introduction covers the motivation behind the research, the problem statement, the objectives, the research challenges, and potential approaches. Chapter two delves into an in-depth background and literature review of indoor technologies, techniques, and machine learning algorithms applicable to indoor positioning. Chapter three meticulously presents the dataset and its mathematical representation, along with an exploration of various preprocessing techniques. Chapter four elaborates on the methodology used to obtain and discuss the results. Chapter Five provides a thorough summary of the findings, while Chapter Six meticulously documents the references used in the paper.

Chapter 2: Overview of localization technologies and techniques

This chapter represents the background of indoor localization, exploring different technologies and techniques used for localization in indoor environments. The indoor localization techniques will be split into classic and machine learning which considered as AI approaches.

2.1 Localization technologies (signals)

Indoor environments can be rich with different signals, which are mainly used to provide connectivity between users. These signals can be used to provide location services by applying different techniques, which will be expressed later. [9] represent many indoor technologies, such as the following:

2.1.1 WIFI

WIFI is a technology that is widely used in the indoor environment to provide wireless local or global connectivity. Depending on the band and device, which gives privileges, WIFI can range from 100m to 1000m. Nowadays, most devices support WIFI technology. However, WIFI is mainly used to provide connectivity, not to support localization services; therefore, an algorithm that can be an extra layer must be applied to activate positioning services.

2.1.2 Bluetooth

Bluetooth is a wireless technology comprising a physical and Mac layer. It is an old technology; therefore, an improved version covers a higher range, a higher data rate, and optimized power consumption. Most devices support Bluetooth technology.

2.1.3 Zigbee

ZigBee is a wireless technology like Bluetooth. However, ZigBee provides a higher data rate transmission and has high power consumption. This technology is usually used when multi-hop routing and network organization are needed. The main challenge here is that few devices currently support this type of technology.

2.1.4 Visible light

Usually, this type of technology is used to create security for valuable items in museums; the way it works is by using a light emitting diode (LED), which works as a transmitter; on the other side, a receiver detects the signal that is coming from LED. This means this technology must satisfy line-of-sight (LOS) between transmitter and receiver, which can be challenging indoors.

2.2 Localization techniques

As mentioned earlier, the indoor environment is filled with various wireless technologies emitting different signals. These signals are primarily intended to establish connectivity between devices and the internet. Therefore, techniques and algorithms will be added to make the signal valuable for localization services.

2.2.1 Angle of arrival (AOA)

The "angle of arrival," also known as the directional of arrival, is used to define the location of an object by calculating the distance and angle at the receiving device [12]. This approach is primarily based on geometric principles to ascertain the location by intersecting the lines of bearing formed by radial lines to each receiving antenna. The principle of angle of arrival is like the concept of outdoor positioning using satellites. Therefore, at least two signal beams are required at the receiving antenna to observe the location. The accuracy of identifying the location can be improved by having more signal beams at the receiving antenna [13]. Most antennas lack the ability to compute the angle of the signal. Adding this capability may increase the device's cost. Moreover, as this technique resembles satellite localization principles, a line-of-sight scenario is required, which presents one of the indoor challenges and limitations [14].

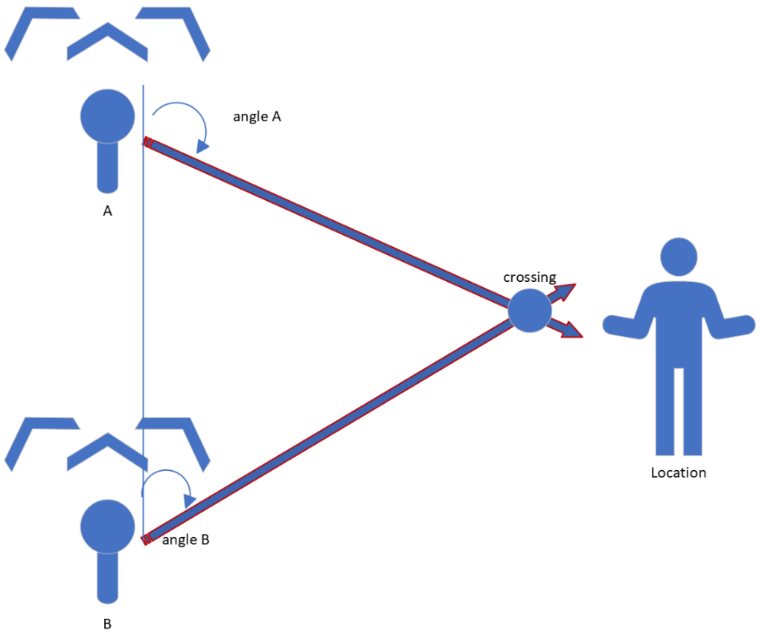

Figure 4: Angle of arrival

2.2.2 Time of flight

Measuring the time of flight or arrival involves determining how long it takes for a signal to reach the receiver based on the speed of light and time, which allows the system to calculate the distance. In indoor environments with multiple transmitters, the receiver's location is where the beams intersect, or conversely, the transmitter can detect the signal returning, representing the time of return. In this case, the time should be divided by two since the signal travels back and forth in the space [15].

Time-of-flight techniques require both the transmitter and receiver to be synchronized with the exact time source and time stamp that will be transmitted with the signal.

No line of sight (NLOS) will negatively affect the quality of the technique. Multipath will also affect the readings and calculation, increasing the range of approximate locations [15].

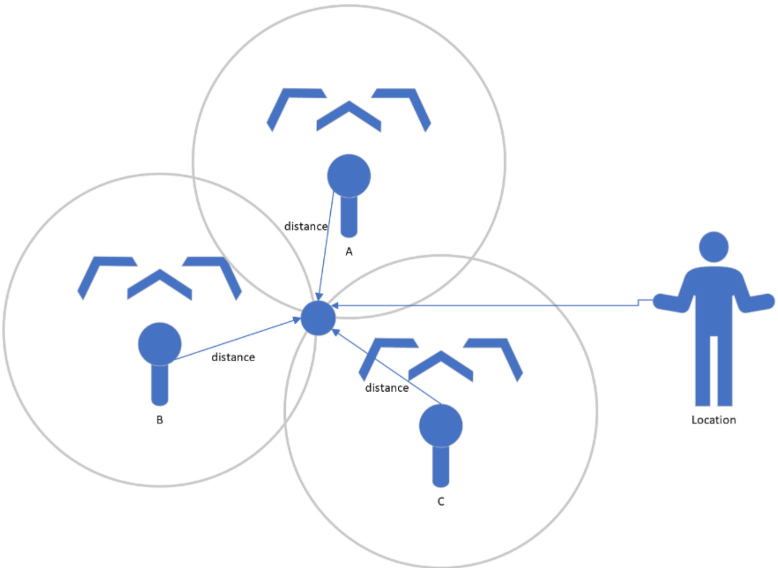

Figure 5: Time of arrival / flight

2.2.3 Fingerprinting

Fingerprinting is a technique that can be used in both machine learning and non-machine learning approaches. The main idea is to gather basic information about the environment, which can act as an offline database for location tracking. This information includes the strength of the received signal from various grid points in the area as well as the phase and the location based on the previous parameters. During the online stage, the system will compare the current signal reading with the stored database information to estimate the device's location. However, to define the location, the reading must be compared with all the database information, which can be computationally expensive in terms of power and time for large databases. The traditional approach does not support adaptive learning, which can impact the system's performance when the indoor environment changes [15].

2.3 Machine learning techniques

It is a technique where the leverage for AI technology is used to extract the features of the data as well as relationships in the offline stage, where the data is collected and processed. A machine learning algorithm extracts the relations of the data to link it with the location. Once the machine learning algorithms learn enough and show an optimistic testing accuracy, the application can be moved from offline to online, start measuring real-time data, and do the localization tasks. Depending on the application's complexity and data type, different types of AI algorithms and techniques can be used. In this section, different AI algorithms will be discussed.

2.3.1 Decision tree

A decision tree (DT) is an AI algorithm that works similarly to the (else, if) statement in classical programming. However, in DT, the feed-in data is the input and the ground truth; in indoor localization, the input data is signal measurements (real and imaginary), the ground truth is the environment type and the coordination (X, Y), the decision tree algorithm will define the conditions; however (if, else), the conditions and input have to be predefined to get the output.

DT is a machine learning algorithm that consists of three main nodes: root node, internal node, and leaf node, which will show the output result. The difference between this node is the incoming/ outcoming edges, representing the connections between nodes; the root node will have only an outgoing edge; the internal node, considered a decision node, will have an incoming and outgoing edge. The leaf node will represent the algorithm's output, with only an incoming edge [5].

The process of building up the DT is during learning. The DT will keep splitting; in other words, it will have more conditions to reduce the uncertainty parameter. DT can be built at different depths (simple, medium, complex), as it has been used in [5] for the classification process of defining the environment or the location of the signal.

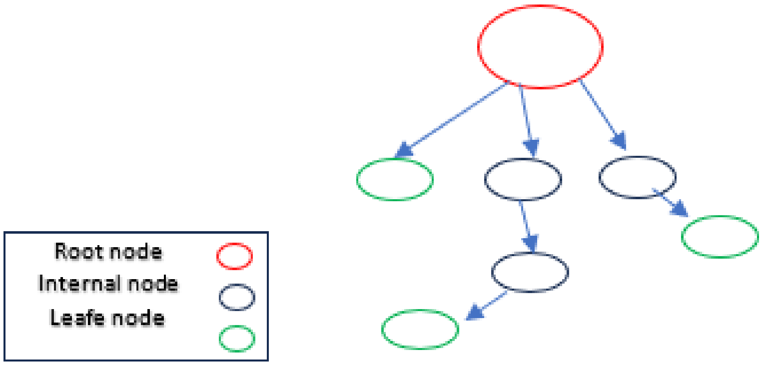

Figure 6 Structure of decision tree (DT).

2.3.2 Support vector machine

Support vector machine (SVM) is an AI algorithm used for classification and regression tasks; however, it has more advantages in the classification tasks. SVM can handle linear and non-linear relationships. In indoor localization tasks, SVM is mainly used to obtain floor levels using regression [10] with different Wi-Fi signals for each floor. The floor can be obtained by projecting all Wi-Fi signals into a space; the signal will be mixed in a one-dimensional space, and the process is done using a kernel.

The kernel is a machine learning algorithm that applies a linear classifier to non-linear data by projecting it into higher-dimensional space. Wi-Fi signals should be considered on different floors or other rooms, as mentioned in [11]. Data is projected in two-dimensional space to express more about it, and linear classifiers must be obtained using dot products. However, the data is non-linear (different floors/ different rooms), so mapping them into higher-dimensional space separates them between two environments. However, mapping features into higher dimensional space is expensive. Using the kernel in SVM will find the function representing higher dimensional mapping without performing the mapping.

SVM has many advantages, including speed, ease of construction, and strong generalization ability. However, as with many machine learning algorithms, the output accuracy will be low if the data is not accurate enough or selected carefully [9].

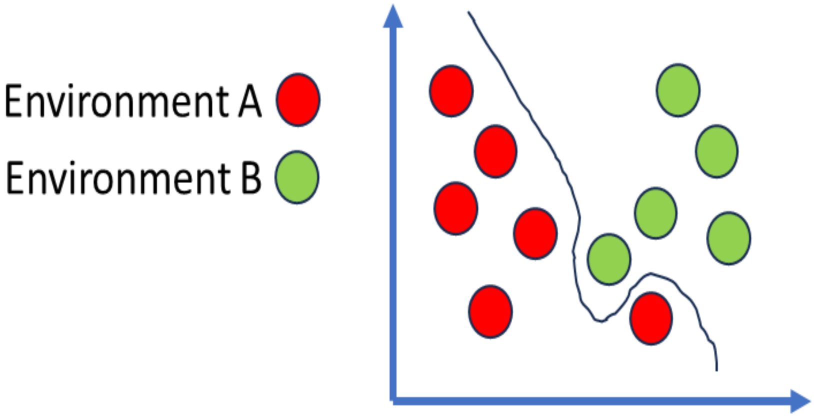

Figure 7: Support Vector Machine (SVM) for two environments

2.3.3 K-nearest neighbor (KNN)

KNN is a type of machine-learning algorithm that is used for classification purposes. KNN algorithm creates clusters of data in a space, where each one of the clusters considers a type of environment or coordination in a space. The classification process is performed based on the type of neighbor data. In indoor localization (environment identification), assuming there are four classes, which can be expressed as the following: 1) low clattered, 2) medium cluttered, 3) high cluttered, and 4) open space. When a new data sample is added to the space, the algorithm will predict which classes the sample is based on the number of neighbors that have been predefined in the algorithm (K value); based on that, if the K, for example, has been assigned to 5, then if there are three neighbors from the low cluttered environment and two from open space environment, the sample data will be assigned to the low cluttered environment [5].

However, taking only the number of neighbors cannot always be accurate. In some cases, based on the previous example, the two samples related to an open-space environment are closer than samples from a low-cluttered environment. Therefore, another version of the k-nearest environment considers the distance between the data sample and the neighbors, called weighted K-NN.

In the normal KNN, each neighbor contributes equally to the sample data; however, W-KNN depends on the Euclidean distance and the number of neighbors that can accurately compute and classify the sample data.

Based on [5], the main challenge of KNN and W-KNN is that they are not the best choices for large data sets. As the number of clusters increases, the algorithm's accuracy will be negatively affected.

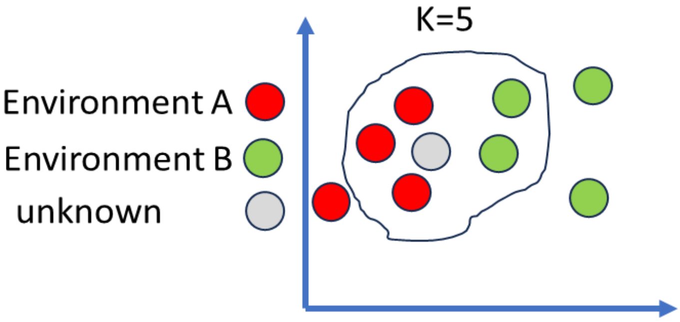

Figure 8: K-Nearest Neighbor (KNN) algorithm.

2.3.4 Deep learning

Machine learning algorithms extract features from the data, which learn the relationship between the input and the output. Further, as big as the number of data to learn from is as high as the accuracy and the extracted features, there are limitations regarding the classical machine learning algorithms when using an extensive data set; at some point, no accuracy improvement can be achieved. Therefore, deep learning algorithms such as ANN and CNN can be used to overcome these limitations.

A neural network consists of three main blocks: 1) input layer, 2) hidden layers, and 3) output layers. Each layer consists of several neurons responsible for extracting features and performing the learning process during forward and backward propagation.

A convolutional neural network (CNN) is like an NN; however, it has a filter that reduces the number of learning parameters; in other words, the learning parameters in CNN equal the filter size.

In [9], CNN has been using indoor localization, mapping the location with the received signal strength RSS for classification purposes.

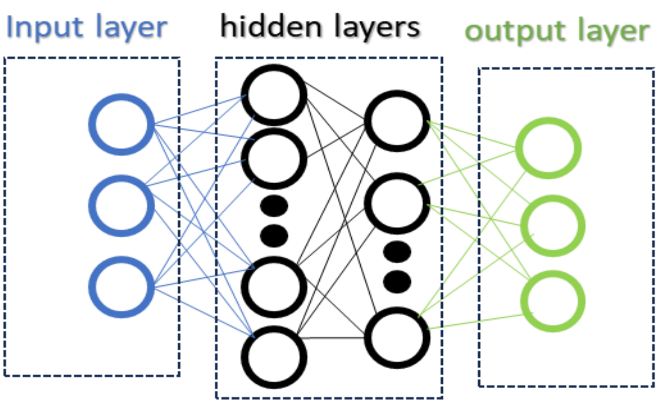

Figure 9: Structure of Deep Neural Network (DNN)

2.4 Localization architecture

Different types of localization architectures are utilized for various applications and purposes. Indoor localization typically involves two core components: a transmitter (access point) and a receiver (smartphone or smart wearable device). Tracking system, the access point passively detects the smartphone signal and determines the device's location based on the known locations of the access points. The second approach involves a navigation task where the smartphone detects signals from different access points. As a result, localization architecture can work in two directions [16].

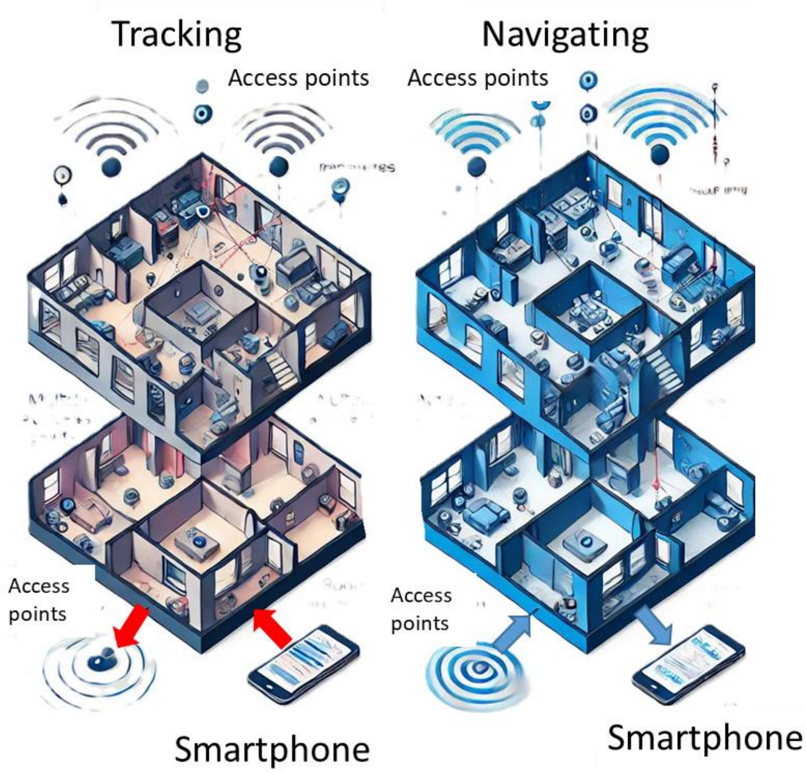

Figure 10: Localization architecture

Chapter 3: Data preparation

This chapter introduces the data used to train and test the AI engine to achieve indoor environment identification as well as positioning tasks and data preprocessing techniques.

3.1 Data-set

The dataset comprises channel coherence function readings gathered from four different environments at Khalifa University Abu Dhabi: 1) high cluttered, 2) medium cluttered, 3) low cluttered, and 4) open space. Each environment includes two types of readings: 1) static mode and 2) moving mode. In the static mode, there is always a line of sight, while in the moving mode, there is no line of sight.

Each environment has been divided into 13X13 squares, starting from (0,0). Each square in the lower left corner represents a grid point, with 200 sweeps (readings) and 101 frequency steps, ranging from 2.4 GHz to 2.5 GHz, with steps of 0.1 GHz.

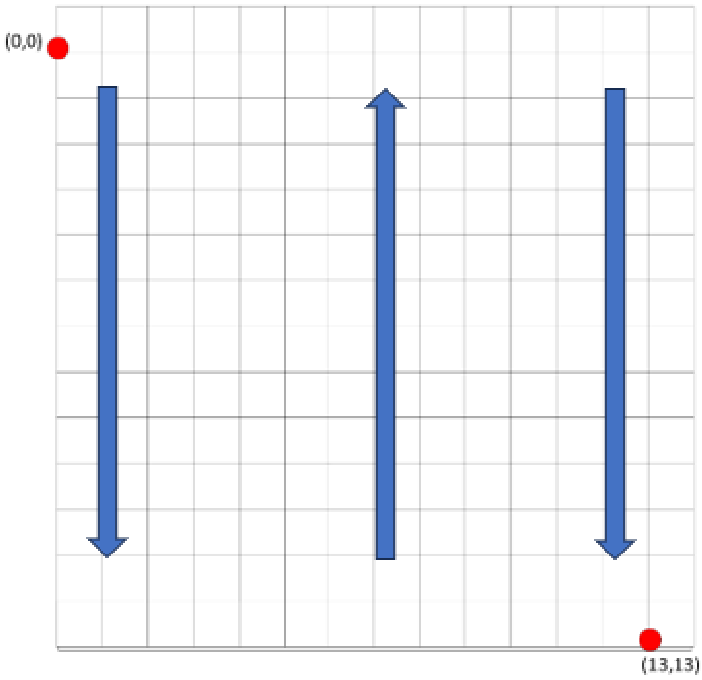

Figure 11: Each environment has been divided into (13,13) grid. The arrows show the sequence of the grid.

3.2 Mathematical representation

The data has been meticulously gathered and organized into four distinct environments based on the cluttering level. Each of these environments is composed of 101 frequency steps, with each step replicated 200 times at the Abu Dhabi campus. Throughout this comprehensive process, each environment has been divided into 13 x 13 grid squares starting from (0,0), which can be denoted as (X, Y) [5].

The represented data will be considered a 4-dimensional tensor and can be denoted as (M). The four dimensions can be detailed as follows:

- Nx: number of x-axis points
- Ny: number of y-axis points
- Nf: number of frequency steps
- Nr: number of repetitions or sweeps in each environment

The following equation can express the mathematical representation:

$$M \in \mathbb{R}^{Nx \times Ny \times Nf \times Nr}$$

Each tensor element is condensed to a single measurement at a specific point in the grid. For example, the element M(x,y,f,r) represents the measurement at row x, column y, frequency step f, and repetition r.

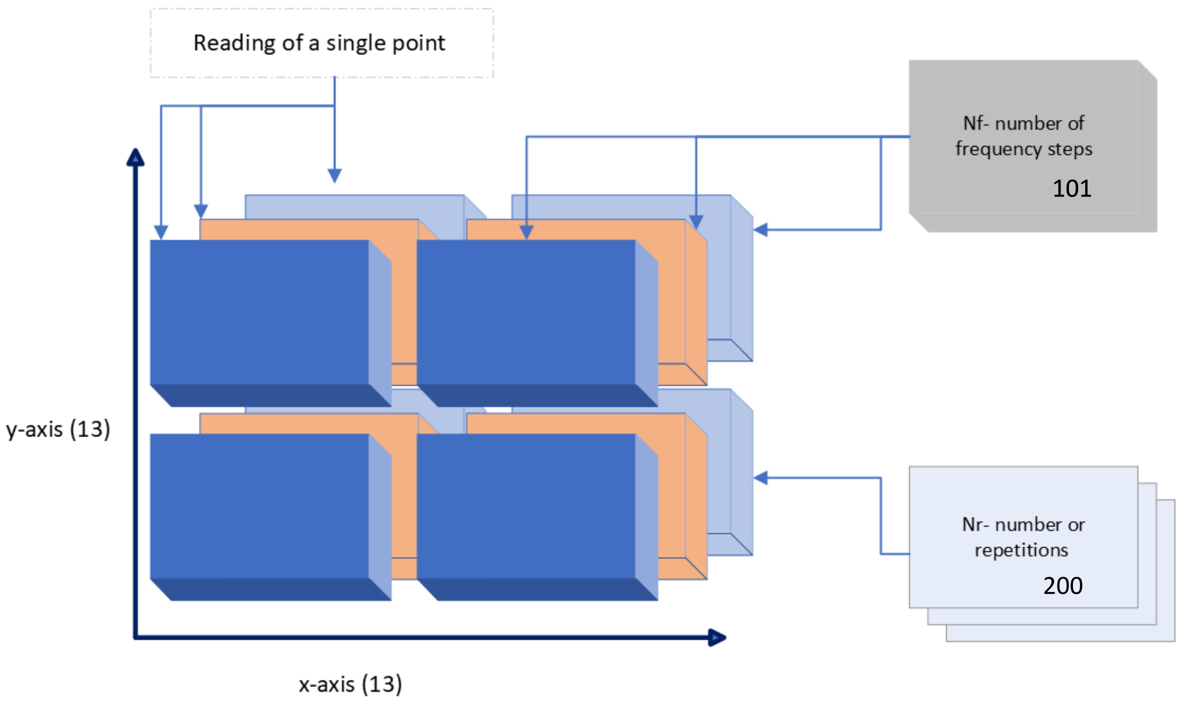

Figure 12: 4-Dimensions tensor

3.3 Data preprocessing

Data Preprocessing is an essential initial step before analysing data. During this stage, the data is transformed and cleaned to prepare it for further processing. The primary goal of data preprocessing is to enhance the quality of the data, improve the system's performance for machine learning or statistical modelling, and increase efficiency in terms of processing and computing power. This process significantly boosts the system's performance and efficiency, giving confidence in its effectiveness.

Data preprocessing does not always mean reducing the number of features. It also means preparing the data so that an AI engine can obtain useful relations and learn from it.

3.3.1 Principal component analysis

Principal component analysis (PCA) is a widely used preprocessing technique for reducing data dimensionality, simplifying training, and aiding in data visualization. PCA effectively transforms higher-dimensional data into lower-dimensional data while retaining the essential features. However, it's important to note that this technique involves a trade-off between simplicity and accuracy [17].

To conduct principal component analysis, several steps must be taken to ensure data quality and preserve essential features. First, the data should be standardized to ensure it is all on the same scale, preventing a large variance range from dominating a smaller one. This involves transforming the data to a comparable scale by subtracting the mean from each value and dividing the result by the standard deviation. Once the data is standardized, the next step involves computing the covariance matrix. This matrix helps to understand how the variables in the input data vary in relation to each other and to identify redundant information. The covariance matrix is symmetric; positive numbers indicate correlated relations, while negative numbers indicate an inverse correlation. By analysing the covariance matrix using eigenvectors and eigenvalues, uncorrelated components can be determined, also known as principal components. The eigenvector represents the direction of the principal components, while the corresponding eigenvalue is the coefficient assigned to the eigenvector. Once these components with the most valuable information are identified, the feature vector is derived, representing the final shape of the principal component analysis and displaying only the selected components. This ultimately leads to dimensionality reduction by choosing only those principal components that contain the most valuable information.

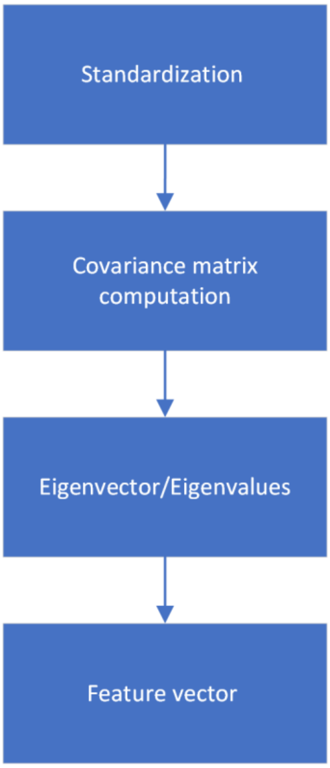

Figure 13: Principle component analysis process chart

3.3.2 T-SNE

T-distributed stochastic neighbour embedding (t-SNE) is a widely used preprocessing technique for visualizing data. It involves transforming higher-dimensional data into lower-dimensional data while preserving the local structure of the data. This method maintains the relationships between neighboring data points, making it suitable for clustering and visualizing complex data [18].

To achieve t-SNE, several steps are involved. First, the algorithm constructs a probability distribution that measures the pairwise similarity between samples in dimensional space. This is done by defining the probability that comes from the average distance between samples. Higher probability represents high similarity and lower distance between neighbors. The key feature is using a Gaussian distribution to model similarities in higher dimensions.

Next, t-SNE maps the data to a lower-dimensional space by creating the same probability distribution that reflects the pairwise similarity in the high-dimensional space. Kullback-Leibler divergence is used to minimize the distribution difference between the original and lower-dimensional space. Kullback-Leibler divergence measures how much data is lost to represent the data in the lower-dimensional space; the aim is to minimize the difference between the real distribution in higher-dimensional space and the approximate distribution in the lower-dimensional space.

The technique's main advantage is its capture of non-linear relationships and complex data relations. However, with a large amount of data, the computation can be expensive and does not always retain global relations between different clusters.

Chapter 4: Approach

In this section, the outcomes of implementing a variety of algorithms will be analyzed for indoor identification as the initial stage, followed by a detailed examination of the localization task as the next stage. An in-depth review of the input data used for training the modules, including discussions on data sources, preprocessing techniques, and feature selection.

4.1 Indoor identification

To achieve an accurate indoor localization system, a machine learning algorithm must be employed to conduct precise indoor identification. This algorithm will ascertain the optimal approach for identifying indoor environments, including open, highly cluttered, medium, and low-cluttered areas.

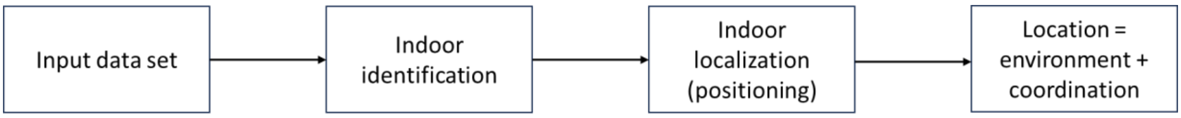

Figure 14: Indoor localization process

The following table shows the accuracy and algorithm types for performing indoor identification tasks; the results will be discussed later in section (4.5). All algorithms have been repeated ten times, and the average has been written.

Table 1: Indoor identification accuracy

Algorithms	DT4	DT20	DT60	DT100	KNN5	WKNN5	KNN10	WKNN10	CNN
Accuracy	57.4%	69.9%	70.3%	66.7%	77.28%	75.8%	76.5%	77.8%	91.4%

4.2 Formatting the data

In section (3), it is noted that the data encompasses channel transfer function readings, featuring real and imaginary numbers corresponding to magnitude and phase readings for individual data points. The frequency range spans from 2.4 GHz to 2.5 GHz, with a step size of 0.01 GHz. Each step involves 200 sweeps across a 13 by 13 grid.

To facilitate improved feature extraction, the input vector has been reshaped to a two by 101 matrix, while the training sample consists of a 196 by 200 matrix. This reshaping enables easier feature extraction for the algorithm, allowing related data with the same frequency bandwidth to be presented jointly. For each grid point, real and imaginary values for all frequency steps are concatenated into a two by 101 matrix. This approach ensures that various bandwidth utilizations can be covered during training, thus enabling the study of the system's behavior across different frequency ranges for the same grid point.

4.3 Indoor positioning

The process of indoor localization involves training a separate module for each specific environment. The input vector is 2 by 101, representing real and imaginary numbers in sequence across the entire bandwidth for each coordinate.

Indoor positioning is classified as a regression task, where the module is tasked with predicting a continuous value, specifically coordinates, based on extracted features and relationships from the training process. The algorithm's performance will be evaluated using Root Mean Square Error (RMSE), a metric that quantifies the disparity between predicted and actual values for regression models. RMSE computes the square root of the average of the squared variances between ground truth and predicted values, as depicted in the following equation, where n denotes the number of samples [19]. Table 2 represents the performance of positioning tasks with different algorithms. All algorithms were repeated ten times, and the average value was written.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{real} - y_{predicted})\ \hat{}2}$$

Table 2 Indoor positioning performance

Algorithms	DT	KNN5	WKNN10	CNN	Environment
RMSE	0.43m	0.49m	0.26m	0.26m	Medium Clattered
	0.53m	0.76m	0.32m	0.32m	High cluttered
	0.25m	0.5m	0.23m	0.23m	Low clattered
	0.7m	0.81m	0.39m	0.39m	Open space
Processing time	40s	96s	120s	720s	

4.4 Data preprocessing

Data preprocessing refers to manipulating, rearranging, and cleaning data before it is used for analysis or model training. It is a critical step in data analysis and machine learning in addition to data visualization. The purpose of data preprocessing is to ensure that the data is in a format that is suitable for the specific analysis or model being used. This can involve tasks such as handling missing data, normalizing the data, and removing irrelevant features. By performing data preprocessing, the quality of the data is improved, making it more suitable for training machine learning models and obtaining meaningful insights from the data. In the context of the given text, the dataset contains two main features: real and imaginary, which respectively represent the magnitude and phase of the signal over a 1GHz bandwidth. This information is crucial for understanding and processing the data effectively.

Principle component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are frequently employed in data analysis to reduce the dimensionality of high-dimensional feature spaces. PCA focuses on finding the directions of maximum variance in the data and projects the data onto a new coordinate system. At the same time, t-SNE is particularly effective at visualizing high-dimensional data in two or three dimensions, preserving the local structure of the data points. Both methods are valuable for gaining insights into the relationships and variations within the data while facilitating visualization and interpretation [18][21].

Given that the dataset only contains two features, an alternative approach will be explored such as reducing the bandwidth or frequency step of the data used for training the model. In other words, the number of frequency points will be reduced. This approach will help analysing and understanding the impact of reducing the bandwidth on the model's accuracy for classification tasks related to environment identification, as well as its performance for regression tasks related to indoor positioning.

Various scenarios will be examined using the entire bandwidth, as outlined in tables (1) and (2). The second scenario involves utilizing only 70% of the bandwidth, as shown in Tables (3) and (5) , and additionally, a scenario utilizing 20% of the bandwidth has also been analysed, as shown in Tables (4) and (6).

Table 3: Indoor identification accuracy using 70% of the bandwidth

Algorithms	DT4	DT60	KKNN5	WKNN10	CNN
Accuracy	54%	67.8%	76.9%	76%	91.4%

Table 4: Indoor identification accuracy using 20% of the bandwidth

Algorithms	DT4	DT60	KKNN5	WKNN10	CNN
Accuracy	46%	57%	64.5%	64%	91.4%

Table 5: Indoor positioning performance using 70% of the bandwidth

Algorithms	DT	KNN5	WKNN10	Environment
RMSE	0.42m	0.56m	0.5m	Medium Clattered
	0.74m	1.1m	0.8m	High cluttered
	0.3m	0.55m	0.35m	Low clattered
	0.8m	0.9m	0.8m	Open space

Table 6: Indoor positioning performance using 20% of the bandwidth

Algorithms	DT	KNN5	WKNN10	Environment
RMSE	0.5m	0.6m	0.58m	Medium Clattered
	1m	1.2m	0.95m	High cluttered
	0.5m	0.9m	0.7m	Low clattered
	1.25m	1.31m	1.2m	Open space

Chapter 4: Discussion

The model has been evaluated using various machine learning algorithms across different environments, clustered into high-cluttered, medium-cluttered, low-cluttered, and open-space environments, and tested across different bandwidth utilization with a training size of 70% and testing size of 30%. The study focused on decision trees (DT), K-nearest neighbors (KNN), and deep learning algorithms, which are convolution neural networks (CNN), that have been proposed in [8]. Since the main goal is indoor positioning, the performance has been assisted with two main tasks: environmental identification and indoor positioning.

Before predicting the location, the environment should be identified, which can be considered a classification task. Table (1) represents the accuracy of each algorithm by utilizing the full bandwidth. The accuracy parameter is calculated based on the ratio of correct predictions to the number of total predictions made by the model, which can be presented as the following equation [22].

$$Accuracy = \frac{No. of\ Correct\ Predictions}{Total\ Predictions} \times 100$$

The Design Trees (DT) algorithm displayed varying levels of result accuracy based on the depth of the tree. It was observed that the accuracy of the algorithm improved as the depth of the tree increased, with the accuracy increasing from 57.4% at a depth of 4 to 70.3% at a depth of 60, representing a significant 22.47% improvement. However, it is important to note that this improvement is not necessarily linear. This was evident when, at a depth of 100, the accuracy dropped to 66.7%, a decrease of 5.12% from the accuracy at a depth of 60. This demonstrates that the relationship between depth and accuracy is not always straightforward and may not always lead to continuous improvement [22].

K-nearest neighbor (KNN) has shown an impressive result as a minimum of 75.8% up to 77.8% where different numbers of neighbors have been tested in addition to weighted K-nearest neighbor (WKNN), by increasing the number of neighbors; the accuracy didn’t really change that much, by going from K=5 to K=10 the result went from 77.28% to 76.5%, however by adding weight (distance) to the algorithm the result went from 75.8% to 77.8% respectively. The idea behind it is that using a smaller number of neighbors provides better results when the local pattern is relevant, which may provide better accuracy; however, when increasing the number of neighbors, the point may include a non-relevant pattern, which reduces the accuracy and also smooth the boundaries between clusters. By adding weight to the (KNN), the closer the neighbors are, the more relevant the data points are to the algorithm, which makes the algorithm focus more on the nearest neighbors, and by increasing the number of neighbors with respect to distance, the data point will have access to more information allowing the system to make a reliable prediction especially when neighbors are not equally relevant [23].

Using a convolutional network provides the most accurate result, with an accuracy of 91.34%, which supports what has been mentioned in [8]. CNN can capture local dependence in RF signal data, which allows it to extract meaningful patterns. These patterns help identify the environment by capturing the spatial structure of the signal and provide more accurate results than classical machine learning.

As mentioned in section (4.4), data preprocessing will be done by minimizing the utilization of the bandwidth; Table (3) and Table (4) represent the accuracy after reducing the utilization of the bandwidth.

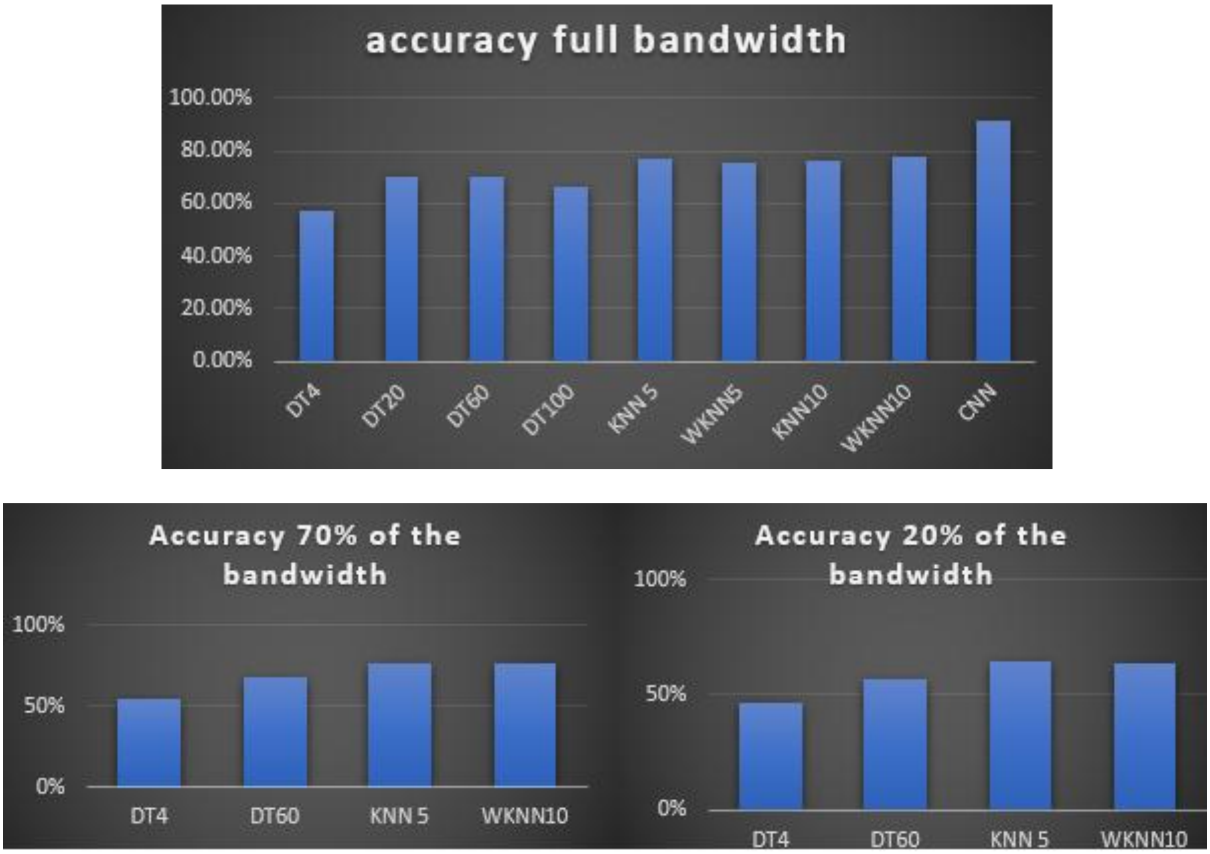

Figure 15: Shows the effect of reducing the frequency points for environment identification. This process has been done on classical machine learning algorithms.

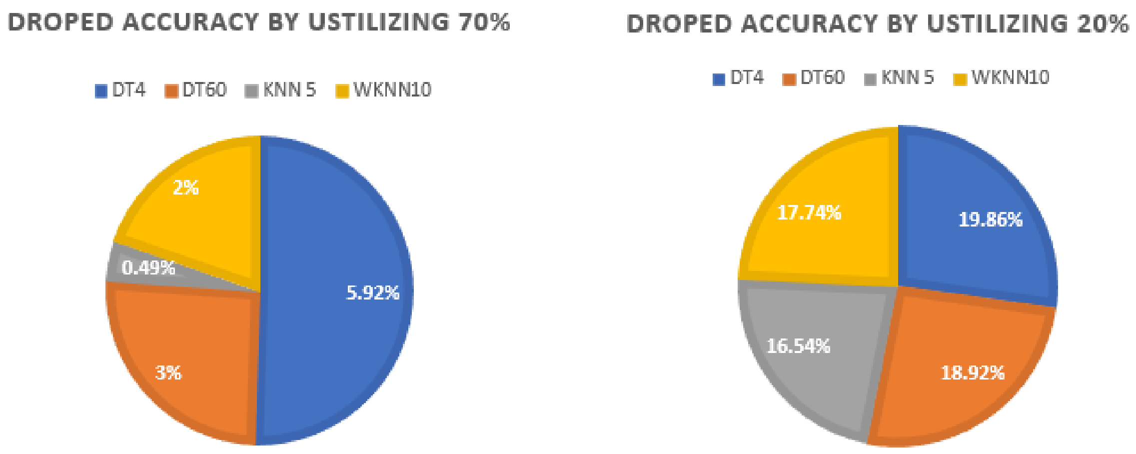

Figure 16: Shows the accuracy drop percentage by reducing the frequency points.

The number of frequency points and the bandwidth utilization have a proportional relationship with accuracy. Reducing these factors will decrease accuracy, but it's important to understand the extent of this impact and how it affects the model's performance compared to using the full bandwidth.

According to Figure 16, utilizing 70% of the bandwidth led to a decrease in accuracy ranging from 5.92% to 0.49%, while utilizing 20% of the bandwidth resulted in a decrease in accuracy ranging from 19.86% to 16.54% compared to using the full bandwidth. Therefore, there's a trade-off between bandwidth utilization and model performance, considering computing power, time, and accuracy.

When utilizing 70% of the bandwidth, decision trees (DT) had a larger effect which demonstrates the algorithms are sensitive to data reduction, while K-nearest neighbors (KNN) showed high resilience when data was reduced as well and KNN can be considered as robust algorithm in a scenario where computational power and bandwidth are constrained specially in a classification task.

In the indoor positioning model, system performance is assessed using the root mean square error, as outlined in section 4.3. Different machine learning algorithms have been evaluated, focusing on a regression task to predict continuous values that will represent the coordination.

Table (2) displays the RMSE results in meters, with the RMSE value being scaled in meters by multiplying by 12.5, representing the spacing between grid points. The results vary based on the algorithm and the condition of the environment in terms of the cluttering level. The best overall performance was observed in a low-cluttered environment with minimal noise in the data, where the algorithms successfully captured the relationship between the data points.

Specifically, the decision tree (DT) showed an RMSE of 0.25m, while KNN, with a value of 5, resulted in 0.5m, and WKNN, with a value of 10, resulted in 0.56m. A deep learning algorithm with the same structure as mentioned in [8] has been used here; however, it was used for the regression task, with a result of 0.23m.

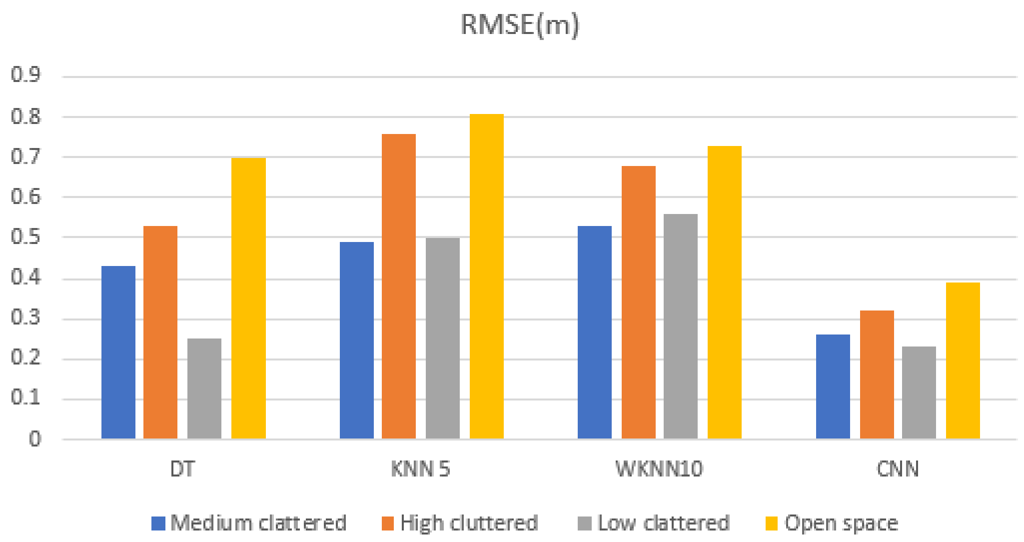

Figure 17: performance of different algorithms by utilizing the full bandwidth in different environment conditions.

Figure 17 shows the diversity of RMSE values with different algorithms as well as different environmental conditions; regarding the figure above, an open space environment has less performance accuracy, which means a considerable RMSE value, which happened due to a lack of reflection, which some sometimes support the model to identify relations and extract features from the data according to [24] the author represent how cluttered environment can lead to reflection which supports the indoor positioning system by providing additional signal path that enhances the ability of indoor positioning. What can be noticed in the above graph is that the predictions of location are worse when having open space as well as a highly cluttered environment. For a highly cluttered environment, according to [15], the multipath can reach a point of extreme, making the signal very hard to process due to noise, making it difficult for the indoor positioning system to identify the relation between data.

The impact of reducing frequency points on indoor positioning systems has been examined using various algorithms to assess the effect on RMSE.

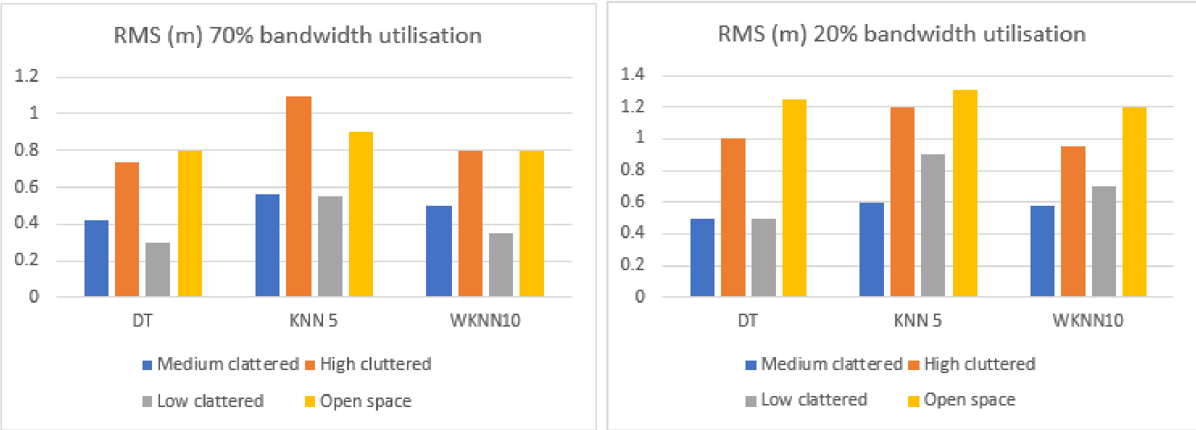

Figure 18: performance of different algorithms by utilizing part of the bandwidth in different environment conditions.

Reducing the bandwidth utilization to 70% has a less significant impact on the decision tree. In a medium-cluttered environment, using a decision tree decreased from 0.43 to 0.42 meters, indicating an improvement. However, the distance increased from 0.7 to 0.8 meters in an open space environment. The high-cluttered environment had a more significant impact, with the RMSE rising from 0.53 to 0.74 meters, representing a 39.62% increase.

On the other hand, using a weighted KNN of ten neighbors in the medium-cluttered environment resulted in a decrease from 0.53 to 0.5 meters. In an open space, it increased from 0.7 to 0.8 meters. In the low-cluttered environment, there was a decrease from 0.56 to 0.35 meters, indicating an improvement of 37.5%. In a highly cluttered environment, the distance increased from 0.68 to 0.8 meters, representing a 17.65% increase.

By reducing the bandwidth further to reach bandwidth utilization of 20%, the figure above shows that the decision tree has less impact when reducing the frequency points in a regression module, taking into consideration the environmental condition as well as the performance needed, which represent the ability of decision tree to capture the relations by splitting which can still be effective even when reducing data. However, utilizing only 20% of the bandwidth reduces all algorithms' performance. The decision tree shows an excellent ability to capture features, especially for regression tasks that represent the ability to use key decision points during splitting. However, KNN algorithms were stable when reducing the data up to 70%, and when the frequency points were reduced to 20%, the RMSE degraded, which shows the algorithm's dependency on a large amount of data to achieve regression tasks.

Chapter 5: Conclusion

This paper presents a highly accurate indoor positioning system that contains classification and regression tasks, which are used to identify the environment and predict the coordinates. The paper discussed different indoor technologies and techniques that can be used to achieve the task, as well as different machine learning algorithms such as decision trees (DT), weighted and non-weighted K-nearest neighbors (KNN) with different hyperparameters tuning, interim of depth, and a number of neighbors. the performance of each algorithm has been examined, which shows the ability of each algorithm to achieve classification and regression tasks, K-nearest neighbor has better performance for environment identification. On the other hand, the decision tree can capture more features and produce better performance in indoor positioning (regression task). Data preprocessing has been applied by reducing the utilization of the bandwidth since there are only two features, and the effect of lowering frequency points has been examined and tested to show the ability of algorithms to capture features and relations to achieve environment identification as well as indoor positioning. The paper shows the effect of the environmental conditions in terms of cluttering level and how that affects the performance, especially when reducing bandwidth utilization. The best performance of the system was low and medium cluttered environment. As well as Deep learning algorithms such as convolutional neural networks have been tested and showed the best result in terms of environment classification with an accuracy of 91.4% and an RMSE value for regression tasks in a low cluttered environment of 0.23 meters. Which proves the ability to capture more features by utilizing the full bandwidth.

Indoor positioning systems involve intricate tasks that are influenced by numerous factors. These factors include system efficiency in terms of running time and computing power, environmental conditions such as cluttering level, algorithm performance, and the desired accuracy and predictions. Achieving optimal performance in indoor positioning systems requires carefully balancing these various elements to best suit the environment's specific requirements and constraints.

Chapter 6: Future Work

This research has shown encouraging results in indoor positioning, focusing on two main tasks: classification and regression. The classification task identifies different indoor environments, enhancing our understanding of their functions, while the regression task predicts accurate coordinates to improve navigation.

However, there are still areas for improvement. Enhancing classification algorithms could boost environmental identification accuracy, and developing advanced models for coordinate prediction may increase location tracking reliability in complex settings. Addressing these aspects will enhance the effectiveness of this research and contribute to advancements in indoor positioning technologies.

Enhancing algorithm performance can be achieved through the development of hybrid algorithms that integrate classical machine learning techniques with modern deep learning approaches. By doing so, we can take advantage of each methodology's unique strengths, leading to more robust and accurate models.

For instance, classical machine learning algorithms often excel in scenarios with smaller datasets and provide interpretable results, while deep learning models excel in handling large and complex datasets with high-dimensional features. By combining these two approaches, we can create systems that not only perform better but also offer insights into their decision-making processes.

Furthermore, it is essential to investigate reinforcement learning, which focuses on training algorithms to make decisions in dynamic environments through trial and error. This adaptive capability allows systems to learn and modify their behavior in entirely new situations, making them more flexible and efficient. By incorporating reinforcement learning into our hybrid frameworks, we can enhance the algorithms' ability to navigate, learn, and thrive in changing environments, ultimately improving their effectiveness in real-world applications.

Additionally, it is important to examine the influence of emerging wireless technologies, particularly those that operate at very high frequencies and offer high bandwidth capabilities. These advancements can significantly alter the overall system performance, introducing new levels of complexity in both design and implementation. It would be beneficial to analyze how these technologies affect the accuracy of data transmission and reception, as well as the potential challenges they pose in terms of interference, signal integrity, and system management. Understanding these factors will provide valuable insights into optimizing the performance of wireless communication systems in the future.

Numerous evaluation metrics can be incorporated into future work to assess the system's performance more comprehensively. Among these metrics are the F1 score, mean specificity, and mean recall. Each serves a vital purpose, depending on the specific domain in which the system is applied.

For instance, in the healthcare sector, the F1 score, which balances precision and recall, can be crucial when diagnosing conditions to minimize false positives and false negatives. Mean specificity, which measures the true negative rate, is particularly important in industries where distinguishing non-cases is essential for patient safety. Similarly, mean recall is vital when ensuring that the system identifies as many relevant cases as possible, especially in scenarios like disease outbreak detection.

In industrial applications, these metrics can help optimize processes by accurately identifying defects or anomalies in production, thereby enhancing overall efficiency. Lastly, in areas focused on quality of life, such as smart home technologies or urban planning, the choice of metrics will significantly influence user satisfaction and operational effectiveness. Thus, selecting the appropriate evaluation metrics is crucial and can substantially impact the system's effectiveness in achieving its intended purpose across various domains.

References

[1] GilPress. (2024, January 31). How many people own smartphones? (2024-2029). What'S the Big Data? https://whatsthebigdata.com/smartphone-stats/

[2] Stamatakis, J. (n.d.). 5 Indoor air quality Statistics You should know about. https://blog.attuneiot.com/5-indoor-air-quality-statistics-you-should-know-about

[3] Aughey, R.J., 2011. Applications of GPS technologies to field sports. International journal of sports physiology and performance, 6(3), pp.295-310.

[4] Q. Niu, M. Li, S. He, C. Gao, S.-H. Gary Chan, and X. Luo, ‘‘Resourceefficient and automated image-based indoor localization,’’ ACM Trans. Sensor Netw., vol. 15, no. 2, pp. 1–31, Apr. 2019.

[5] MI AlHajri, NT Ali, RM Shubair, “Classification of Indoor Environments for IoT Applications: A Machine Learning Approach” IEEE Antennas and Wireless Propagation Letter, Sept. 2018.

[6] M. I AlHajri, N. T Ali, R. M Shubair, “A machine learning approach for the classification of indoor environments using RF signatures” IEEE Global Conference on Signal and Information Processing(GlobalSip 2018). November 2018, California USA.

[7] MI AlHajri, NT Ali, RM Shubair, “Indoor Localization for IoT Using Adaptive Feature Selection: A Cascaded Machine Learning Approach” IEEE Antennas and Wireless Propagation Letter, May 2019.

[8] Z Chen, MI AlHajri, M Wu, NT Ali, R Shubair, “A Novel Real-Time Deep Learning Approach for Indoor Localization Based on RF Environment Identification” IEEE Sensors Letters, Vol 4, June 2020.

[9] S.Shang. And L Wang. (2022) “Overview of WiFi fingerprinting-based indoor positioning”, IET Communications, 16(7), pp. 725–733. doi:10.1049/cmu2.12386.

[10] S Zhang., J Guo., W Wang, Hu, J.: “Indoor 2.5D positioning of WiFi based on SVM. 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)”, 1–7 (2018).

[11] Y Rezgui, Y Pei, L.,X Chen, F Wen, C Han.: “An efficient normalized rank based SVM for room level indoor WiFi localization with diverse devices”. Mobile Inf. Syst. 2017, 1–19 (2017)

[12] S. Wielandt, B. Thoen, and L. De Strycker, ‘‘Experimental evaluation of a single anchor multipath assisted indoor angle of arrival localization system in the 2.4 GHz and 5 GHz band,’’

in Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN), Nantes, France, Sep. 2018, pp. 1–7.

[13] B. Yimwadsana, V. Serey, and S. Sanghlao, ‘‘Performance analysis of an AoA-based Wi-Fi indoor positioning system,’’ in Proc. 19th Int. Symp. Commun. Inf. Technol. (ISCIT), Ho Chi Minh City, Vietnam, Sep. 2019, pp. 36–41

[14] Arbula, D. and Ljubic, S., 2020. Indoor localization based on the infrared angle of arrival sensor network. Sensors, 20(21), p.6278.

[15] Zafari, F., Gkelias, A. and Leung, K.K., 2019. A survey of indoor localization systems and technologies. IEEE Communications Surveys & Tutorials, 21(3), pp.2568-2599.

[16] Mautz, R. (2012). Indoor positioning technologies. Habilitation thesis, ETH Zurich. DOI: 10.3929/ethz-a-007313554

[17] Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. International Journal of Innovative Science and Research Technology, 8(5), 2028-2032.

[18] van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[19] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)– Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

[20] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann.

[21] Jolliffe, I. T. (2002). Principal Component Analysis. 2nd ed. Springer Series in Statistics.

[22] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media. ISBN: 978-1492032649.

[23] Cover, T., & Hart, P. (1967). "Nearest neighbor pattern classification." IEEE Transactions on Information Theory, 13(1), 21-27.

[24] Mendrzik, R., Dammann, A., & Jost, T. (2016). Exploiting multipath propagation for improved indoor localization in wireless networks. In IEEE International Conference on Communications Workshops (ICC).