A Survey of Data Mining Techniques for Indoor Localization

Usman S. Toro
Department of Computer and Communications Engineering, Abubakar Tafawa Balewa University, Bauchi
E-mail: usmansalehtoro@gmail.com

Nasir A. Yakub, Aliyu B. Dala, Murtala A. Baba, Kabiru I. Jahun, Usman I. Bature and Abbas M. Hassan
Department of Computer and Communications Engineering, Abubakar Tafawa Balewa University, Bauchi
E-mail: aynasir@gmail.com, aliyudalabukar@gmail.com, limaminkobi@yahoo.com, kjahun@hotmail.com, usmanbature123@gmail.com, abbas_musahassan@yahoo.com

Abstract: The important need for suitable indoor positioning systems has recently seen an exponential rise with location-based services emerging in many sectors of human life. This has led to adopting techniques to mine location data to discover useful insights to improve the accuracy of the various indoor positioning systems. Although indoor positioning has been reviewed in some literary works, an in-depth survey of how data mining could improve the performance of indoor localization systems is still lacking. This paper surveys data mining techniques such as Naïve Bayes, Regression, K-Means, K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Random Forest (RF), Expectation Maximization (EM), Neural Networks (NN), and Deep Learning (DL) including how they were used to improve the accuracy of indoor positing systems using various supporting technologies such as WiFi, Bluetooth, Radio Frequency Identification (RFID), Visible Light Communication (VLC), and indoor localization techniques such as Received Signal Strength Index (RSSI), Channel State Information (CSI), fingerprinting, and Time of Flight (ToF). Additionally, we present some of the challenges of existing indoor positioning systems that employ data mining while highlighting areas of future research that could be exploited in addressing those challenges.

Index Terms: Data mining techniques, Indoor Localization techniques, Indoor Localization technologies

1. Introduction

Humans have always been interested in the question "where?" from time immemorial. Be it concerning humans, transportation, commerce, personal belongings e.t.c. As a result, a lot of research effort has been put over the years into locating objects and humans both indoor and outdoor. The advancement of technology has similarly improved techniques and methods employed in locating objects and humans.

Indoor localization involves the process of identifying the position of a user or device relative to another device or a reference point in an indoor environment [11]. Indoor localization has presented a particular challenge due to the nature of the indoor environment being surrounded by walls and having many items in between that could pose as obstacles including walking persons. Thereby making it almost impossible to have a clear Line of Sight (LoS) path. These obstacles also compound the non-line of sight (NLoS) problem due to the multipath effect they create. Another challenge is the dynamic nature of the indoor environment which requires a very robust indoor model that can capture changes in the number of obstacles, human positioning and other items that add to the multipath effect. For the sake of clarity, abbreviations used in this paper are summarized in appendix A.

Over the last two decades, a lot of research has been carried out on developing indoor localization systems due to its importance despite the challenges it poses. The recent rise to prominence of technologies such as Backscatter Communication (BackCom), Simultaneous Localization and Mapping (SLAM), and Internet of Things (IoT) which promise ubiquitous sensing and communication of billions of devices at low cost and higher precision makes indoor localization even more important. Application scenarios of indoor localization include the; detection of Radio Frequency Identification (RFID) tags, tracking robots in a factory, and monitoring of kids' whereabouts. These application scenarios give rise to large amounts of application-related data which can be leveraged together with existing indoor localization technologies (e.g. WiFi, Bluetooth Low Energy (BLE), Long Range Radio (LoRA), and RFID) and techniques (e.g. Received Signal Strength Indicator (RSSI), Time of Flight (ToF), Angle of Arrival

(AoA), and Channel State Information (CSI)) to have a better understanding and improve the performance of indoor localization systems [1,2,3,4,5].

Data mining involves the application of statistical and machine learning tools to discover new and useful insights from large amounts of data. These insights can be used to improve processes in the domain where the data is generated or in a neighboring domain. As a result, data mining has shown the ability to improve processes in healthcare, transportation, security, agriculture, and education despite their dynamic nature and the large volume of domain-generated data [6,7,8,9,10]. Having witnessed the potential of data mining techniques, we aim to provide a comprehensive survey paper on utilizing such techniques in enhancing the performance of indoor localization system. The paper will serve as a good reference for new researchers into the field and also professionals wishing to get up-to-date with state-of-the-art on the subject. Though there exists some prior survey works on the subject [11,12,13,14,15,16], there is still some gap that needs to be filled. The strengths and limitations of the existing works discussed in section 2. A chart showing the sectors where data mining has improved and a list of data mining techniques are presented in Fig. 1 and Fig. 2 respectively.

The main contribution of this paper is providing an overview of indoor localization techniques, data mining techniques, and their use-cases in improving the performance of indoor localization systems. Additionally, the paper discusses the challenges and future research work related to employing data mining to improve the performance of indoor localization.

The rest of the paper is structured as follows; section two presents a survey of indoor localization and data mining, section three gives a primer on indoor localization techniques, section four describes the methodology adopted in the paper, section five gives an overview of data mining techniques, their use-cases in indoor localization, section six discusses the challenges and future work and finally, section seven concludes the paper.

Fig. 1. Data mining and various industrial sectors improved.

2. Related Work

There exist many recent surveys on indoor localization [11,12,13,14,15,16]. These surveys give an overview of indoor localization systems and their application in various scenarios.

In [11], an in-depth discussion on indoor localization was presented, including its types (device-based, monitor-based, and proximity- based), techniques and technologies for indoor localization, and a primer on application of indoor localization. Several works were evaluated based on; accuracy, scalability, latency, range, and cost. A discussion outlining challenges and proposed solutions of indoor localization was also given. But the data mining techniques discussion was limited to its mention concerning fingerprint/scene analysis technique. In [12], a survey of indoor localization systems based on WiFi Channel State Information (CSI) was presented. Therein, various techniques under geometric (angle, range and both) and fingerprint-based positioning systems were discussed. Also, the challenges and promises of CSI-based indoor localization were presented. But data mining techniques were only discussed concerning fingerprinting technique of WiFi-based (CSI) method of indoor localization. In [13], a detailed discussion on the technologies that enable Network localization, tracking, and Navigation was presented. 5G network-based localization was discussed highlighting its challenges and solutions based on Wireless Local Area Network (WLAN) and other technologies. Mobility estimation techniques using Global Positioning System (GPS), signal power measuring techniques, and Hand Over (HO) information were also outlined. Further, the challenge of mapping the indoor physical environment was also discussed highlighting solutions such as; processing raster images, floor plan, and SLAM. Despite the depth of solutions discussed, data mining techniques employed in the various solutions were mentioned under the different enabling technologies without a detailed discussion of the techniques. In [14], a

discussion of online fingerprinting-based indoor localization techniques was presented. Therein, a categorization and discussion of online fingerprinting into; SLAM-based and crowdsourcing-based was presented. Also, a summary of the various systems in terms of accuracy, robustness, processing time, and security was given. Although the online fingerprinting techniques have shown the ability to capture changes in the environment, a discussion of the data mining techniques involved in online fingerprinting was not presented. A discussion of indoor localization techniques for an emergency was presented in [15]. Due to the peculiarity of emergencies, a careful choice of supporting technology has to be made. Hence a discussion on the merits and challenges of some technologies was presented. Also, state-of-the-art positioning systems were reviewed and their performance was compared. Future challenges were also highlighted but, a discussion on data mining techniques associated with various emergencies for indoor positioning systems is not presented. In [16], a discussion of the various tools and techniques used to develop a Global Indoor Positioning System (GIPS) called KAILOS was presented. Owing to the versatility in the deployment of WLAN in various indoor settings, its fingerprinting capability is used for various environments. Unsupervised learning techniques were used to obtain radio maps of the indoor environments which are combined with a probabilistic method to perform indoor localization. Though a discussion of the tools (indoor layout, modeling, map construction, installation, and testing of system) for GIPS are discussed, the data mining techniques here were only discussed concerning indoor positioning systems based on WLAN fingerprinting. In [17], a discussion of data mining techniques for smartphone-based indoor localization was given. Therein, a detailed discussion of the steps involved in designing a smartphone-based localization system were discussed. They include; the collection of the indoor environment data, putting them into a database, using some data mining techniques to sort the data in the database according to some indices, and then applying the localization algorithms. Though some data mining concepts (Z-score and KNN) were discussed, they were only discussed concerning smartphone-based indoor localization systems. One of the most recent surveys of data mining techniques was presented in [18]. Therein, a survey of IoT-based indoor navigation systems was done. The various aspects of IoT that were improved using data mining were presented. A detailed discussion of indoor navigation and its various aspects (mobility, localization, navigation, communication, and planning) and the future challenges were presented. Further, data mining techniques applied to solve various indoor navigation challenges such as; accuracy, availability, scalability, cost, and privacy were outlined. Though the paper discussed the topic at length, the solutions provided by data mining techniques are restricted to IoT-based applications. The Table 1 of the paper presents the existing surveys on the topic. The overall idea of each survey and how recent it was published were shown. This information helped us in defining the scope of our survey and ensuring that existing works were not duplicated.

Table 1. Existing indoor localization surveys and their scope on data mining techniques

Reference	Year	Scope of discussion of data mining in relation to indoor localization
[11]	2019	data mining techniques (SVM, ANN, and KNN) were discussed in relation to fingerprinting technique
[12]	2019	data mining techniques (SVM, RF, Naive Bayes, KNN, and DL) were discussed concerning CSI-based WiFi fingerprinting technique
[13]	2018	data mining techniques (KNN and its variants) were discussed for the various network localization techniques
[14]	2018	data mining techniques for online fingerprinting were not discussed
[15]	2017	data mining techniques for emergency positioning were not discussed
[16]	2017	data mining techniques (KNN and EM) were discussed for WLAN fingerprinting
[17]	2012	data mining techniques (KNN) were discussed in relation to smartphone-based positioning
[18]	2019	discussed data mining techniques for indoor navigation systems at length. But restricted discussion to IoT applications

3. Primer on Indoor Localization Techniques

This section summarizes the techniques that have been used for indoor localization.

Since the devices that are used for indoor localization can receive and transmit signals based on various technologies (e.g. WiFi, Cellular, Bluetooth, RFID), indoor localization techniques have been developed based on signals of such technologies. The techniques are classified into; centralized and distributed [18]. The choice of technique to use depends on the structure of the indoor environment [15], the technology associated with the device, and the type of location-based application. Generally, the techniques are classified as follows;

A. Received Signal Strength Indicator (RSSI)

This technique employs received signal strength (RSS) from a device to estimate its location. Since the strength of the received signal is obtained in relative measurements decibel-milliwatts (dBm), the absolute location could only be obtained using a propagation model. Localizing a device using RSS with reference to many nodes requires N-point lateration. This technique provides a simple way of estimating the location of a user but suffers from inaccuracies due to the multipath effect. Increasing the complexity of algorithms could improve the accuracy [11].

B. Channel State Information (CSI)

Many existing wireless technologies are frequency selective (i.e. different frequencies exhibit different amplitude and phase behavior). Furthermore, multiple antenna systems show different channel frequency responses between

transmitter-receiver pairs. The fine variations are simply aggregated as received signal strength in the RSSI technique leading to inaccuracies in estimation. Currently, systems like IEEE802.11 Network Interface Cards (NICs) provide CSI which consists of Channel Impulse Response (CIR) enabling a finer granularity of information on channel behavior than the RSS. The CSI is more stable and provides higher accuracy of indoor location [11].

C. Fingerprint

This involves obtaining RSS or CSI measurements of an environment during an offline phase to have a prior idea of the dynamics of the environment (feature/fingerprint). During an online phase, the prior feature is matched with the new feature to obtain the actual location of a user or device. The offline fingerprint is matched with the online fingerprint using probabilistic or data mining techniques [11]. Though this technique is quite accurate, it requires constant updating of the fingerprint.

D. Phase of Arrival (PoA)

This technique explores the Phase of arrival of a signal at individual elements of a receiver antenna array to estimate the distance between transmitter and receiver. Under the assumption that signals are pure sine waves with the same frequency and zero offset, the phase of arrival can be calculated. This technique suffers in the absence of LoS. Its accuracy can be improvedby combining it with other techniques [11].

E. Angle of Arrival (AoA)

This technique uses an antenna array at the receiver to estimate the angle at which the signal was transmitted. This is done by calculating the time difference of arrival at individual elements of the antenna array. Though this technique provides high accuracy, its performance deteriorates with an increase in transmitter-receiver distance. And it requires complex processing algorithms [11].

F. Time of Flight (ToF)

In this technique, the time a signal takes to propagates from a transmitter to a receiver is used to estimate the distance, provided its speed is known. In situations where there are multiple transmitters, N-point lateration can be used to calculate the distance. This technique requires strict synchronization between transmitter and receiver and sometimes time stamps on signals transmitted to reduce location estimation errors. The technique suffers from low bandwidth signals and inconsistency of sampling rate with signal arrival. These problems are addressed by using higher bandwidth and frequency domain super-resolution algorithms [11]. Another challenge of the ToF technique is, absence of LoS causes a signal to take paths that are longer than the actual distance.

G. Time Difference of Arrival (TDoA)

Here, the time difference between arrivals of signals from different transmitters is used to estimate the location of an object. This technique faces similar challenges to ToF. But here, synchronization is only necessary between all the transmitters [11].

H. Return Time of Flight (RToF)

This technique measures the distance by taking the ToF of one round trip (which is the ToF where a signal is transmitted and a receiver process and sends it back to the transmitter). This technique requires moderate synchronization. The RToF technique faces similar challenges to ToF. Moreover, the processing time of a signal at the receiver might increase the round-trip time in short-range systems thereby causing estimation inaccuracy [11].

4. Methodology

Considering that this paper is survey of existing works aimed at identifying strengths and weaknesses in a certain application scenario, we employ an unconventional methodology. First, we review existing survey papers on the subject and identify their scope. We then describe the scope of our survey paper outside those of the existing surveys. Further, we collected papers in the last ten (10) years on the subject and summarized them based on the data mining techniques and indoor localization methods they adopt. Lastly, we tried to identify existing gaps in the subject and suggest ways that those gaps could be filled.

5. Overview of Data mining techniques and their use-cases in Indoor Localization

This section introduces the various data mining techniques used to improve the performance of indoor localization systems, which are shown in Fig. 2.
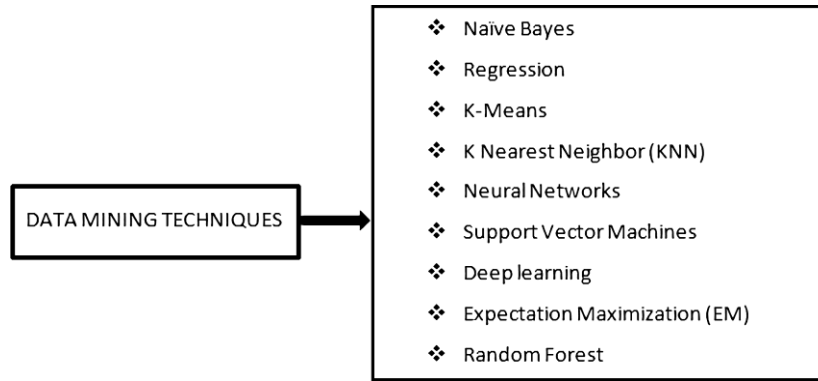
Fig. 2. Data mining techniques.

A. Naïve Bayes

The Naïve Bayes is a classification algorithm based on the Bayes' theorem. The Bayes model used in the algorithm is called "naïve" because of the assumption of conditional independence. This assumption means the presence of a feature in a class is not dependent on other features. The naive assumption is shown in (1) [19].

$$P(x_1 = a_1, ..., x_d = a_d | C = c) = \prod_{j=1}^{d} P(x_j = a_j | c) \tag{1}$$

Naive Bayes use-cases: In [20], the conditional independence of RSS was exploited by implementing Naïve Bayes as the classification algorithm for indoor fingerprinting since it also assumes the independence of events. Despite that, the incompleteness of data taken during offline fingerprinting causes Zero Probability (ZP) which significantly reduces localization accuracy. The paper presents an improved version of Naïve Bayes learning 'INBS' which can handle ZP. INPS achieved an average localization error of 5.048 m. In [21], an indoor navigation system was proposed. Since navigation consists of localization and path planning, the system implements both, thereby allowing a user to specify a multi-destination path. Initial locations are obtained from WiFi fingerprint and the path is planned using 2-opt and A* algorithms. The system accurately updates the position of a user on the path using Naïve Bayes considering the prior location. The system shows a median accuracy of 1.86 m-2.54 m depending on the grid model where it's deployed. In [22], a Bluetooth Low Energy (BLE) based indoor localization system was presented. It uses the fingerprint obtained from BLE for the training and testing of the system. Since RSS values of BLE fluctuate, the indoor area is divided into classes rather than exact locations. The system shows accurate detection of classes where objects are placed. In [23] an indoor 3D-positioning system based on a hybrid (Naïve Bayes and KNN) was presented. The system uses RSS values obtained from WiFi Access Point (AP) and applies the hybrid algorithm to detect the 3D location of unknown objects. It achieved a mean error of 1.80 m. Contrary to the adoption of RSS values from all access points and taking their weights to form fingerprints of an indoor environment, [24] proposed a technique (LocalRelief-C) for identifying and eliminating the APs that are just causing interference to the RSS fingerprints. In the online phase, APs are discriminated against and put into clusters. Then Hidden Naïve Bayes is used to detect the location of an unknown object. In [25], a passive indoor localization system based on CSI was designed. The system uses the CSI to provide more stability and accuracy compared to RSS. The system uses the CSI obtained during the offline phase and a Naïve Bayes classifier to train the system. A confidence level is also introduced to improve the predictions of the classifier. Results show 86% accuracy and at least 15% improvement over existing techniques.

B. K Nearest Neighbor (KNN)

The KNN algorithm is a supervised data mining technique that is mostly used for regression and classification. It operates on the principle "birds of the same feather flock together" i.e., an object will belong to the same class with its K nearest neighbors. The choice of K determines the number of objects around the one to be classified and also determines the correctness of the results. A popular metric for determining how near objects are the Euclidean distances [19].

KNN use-cases: In [26], an effort to improve the performance of WiFi fingerprinting-based indoor localization was presented. Therein KNN was used to improve the accuracy of predicted location by considering the dynamic nature (obstacles, walls etc.) of the indoor environment. Results show a mean error of 1.7928 m. Contrary to many existing approaches that use CSI to perform fine-grained localization, a distributed CSI approach was used in [27]. The received CSI from multiple antenna is decomposed into subcarrier CSI to exploit channel and spatial diversity during the training phase. During the localization phase, a test CSI vector is used together with KNN (based on a similarity measure) to calculate the unknown position of an object. Results of this algorithm show a median accuracy of 1 m. In [28], improved KNN was used to enhance the performance of an Ultra-High Frequency (UHF) RFID-based

indoor localization system. In the system, the RSSI values of the environment are obtained then classical KNN is used to obtain the target's location with location identification based on dynamic active RFID calibration (LANDMARC) algorithm. Based on the K-values obtained, an improved algorithm is used to refine the obtained location of the target. Results show an improved accuracy of 0.5 m in a 10 x 10 m room with 80 tags. Similarly, an improved KNN is used to enhance the performance of an RSS fingerprint-based indoor localization system in [29]. The system obtains indoor fingerprints in the offline phase. It then uses the online values of RSS to calculate the distance to the target by employing an improved KNN which considers the non-Gaussian distribution of measurement noise. Results show an average accuracy of 1.5 m when 15 APs are deployed. In [30], the performance of weighted KNN based WiFi localization systems was improved by combining the weighted KNN with a clustering technique (K-means). This is done to select the best possible candidates whose positions will be weighted during the localization. The system shows an average error of 1.6 m. In [31], iterative weighted KNN (IW-KNN) was proposed to improve the performance of weighted KNN. But here the system uses RSS fingerprints of BLE for indoor localization. After obtaining the fingerprint database, the neighboring iBeacons of the target are iteratively selected and used to compute the location of the target. Results show an improvement of 1.5 m on mean estimation error. UILoc [32] uses Pedestrian Dead Reckoning (PDR), WiFi, and iBeacons to design an indoor localization system. It does not require offline measurements to obtain finger printing database. It uses PDR to obtain the initial position then uses iBeacons with a fusion algorithm to obtain accurate values of the target location. Results show an average localization error of 1.1 m. In [33] an indoor positioning system that uses KNN and ANN was used to design an improved system. The system achieves an average accuracy of 0.9 m. In [34], AM radio signals were used to generate RSS fingerprints for indoor localization. Results of KNN and its variants show an error of 2.76 m when K=1. To achieve high accuracy of smartphone-based indoor localization, [35] proposed a system that integrates CSI and Magnetic Field strength (MFS) data to generate a fingerprint of an environment. KNN was employed to compare the fused fingerprint values during location estimation. The system achieved 0.5 m accuracy.

C. Regression

Regression is also a supervised data mining technique that is used to establish a causal relationship between a dependent and independent variable. The most common ones (linear and logistic regression) can be used to predict a numerical outcome and for classification respectively [19].
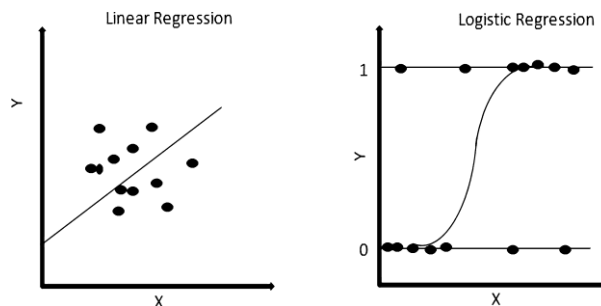

Fig. 3. Linear and Logistic Regression.

Regression use-cases: In [36], an AoA estimation method for correct estimation of Direction of flight was implemented using BLE. In contrast to existing techniques that employ parametric or computationally expensive methods, the classical multiple signal classification (MUSIC) algorithm was used with regression-based models for the AoA based localization algorithm which is more accurate and energy-efficient. Results show a mean error of 2^0 with 30 dB Signal to Noise Ratio (SNR) and 0-49^0 elevation. One-to-all regularized logistic regression classifier (ORRLC) based indoor localization was presented in [37]. The algorithm uses an iterative optimization process to select a location from an M-class of possible locations obtained using WiFi RSS. ORRLC shows a location estimation accuracy of 95.8%. In contrast to existing WiFi-based localization systems that use CSI from multiple links, [38] designed and implemented a robust sub-meter level system that utilizes a single WiFi link thereby reducing computational overhead. It implements the algorithm using logistic regression-based deep learning and Cramer-Rao lower bound assisted training. While the preceding systems have used regression to improve the accuracy of the localization algorithm, [39] used Gaussian-based regression to augment the fingerprint of an environment by using minimal information. Then KNN is used to obtain the accurate position based on the augmented fingerprint. The results show an improved accuracy by obtaining a localization error of 5.12 m. In indoor localization systems where RSS measurements are obtained by crowdsourcing, the variation of RSS becomes worse due to the difference in devices used to obtain the measurements. The system in [40] uses linear regression to resolve these RSS variations. The resolved RSS values are then used to form a radio map for locating targets. Results show an improved average localization error of 2.31 m. Similarly, variation in RSS values caused by heterogeneous orientation of IoT devices was addressed in [41] using a multistage linear regression and Gaussian mixture model. Results show an

improved accuracy of 1.35 m when detecting IoT devices using WiFi signals. In [42], a regression technique was used to improve the performance of a VLC-based indoor localization. The system combines regression with linear and non-linear least square estimators to improve the accuracy of estimation. The results show an average error of 0.22 m.

D. Support Vector Machines (SVM)

Support Vector Machines (SVMs) are naturally defined for binary classification of a single class and multi-class numeric data. The classification is done by finding the best hyperplanes that separate the boundary between the classes of n-dimensional plotted data. Where the data are not easily separated using hyperplanes, the SVM employs kernel tricks to transform the data into another dimensional space before separating them [19].
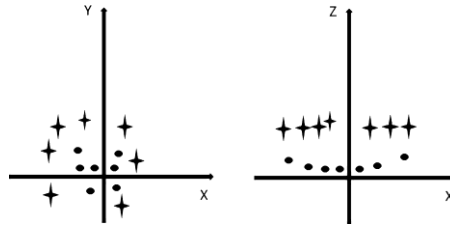

Fig. 4. Data transformation in SVM.

SVM use-cases: In [43], an indoor localization system for underground mines was presented. The system uses ZigBee fingerprints and an SVM classifier to determine the location of targets. The system achieved a median accuracy of 1 m. In [44], an indoor device-free localization system based on WiFi was presented. The system uses SVM to map the non-linearities between CSI of link and device location. But before that, it removes the noise in the CSI and performs principal component analysis (PCA) to get the most important parts of the CSI. The system achieves a mean localization error of 1.22 m. Similarly, an indoor device-free localization algorithm (MaLDIP) was presented in [45]. MaLDIP exploits the channel and spatial diversity of Multiple Input Multiple Output (MIMO) CSI to estimate the target location. To avoid the effect of multipath, a novel subcarrier selection is carried out to obtain channels that are most relevant to the target location, and SVM is applied to estimate the target location based on cells. Results show a 98% accuracy of cell detection. In [46], SVM was used to improve the performance of an affinity-based indoor location system. Due to the unpredictability of RSS, an AP selection algorithm was used to select the most reliable APs to reduce input feature components. An affinity propagation-based clustering was used to reduce the computational complexity of the localization system. Results show an average RMS error of 1.8 m. A novel fuzzy Least Squares Support Vector Machine (LS-SVM) based indoor localization system was implemented in [47]. RSS values of known locations are obtained during the offline training. In the online phase, the LS-SVM is used to obtain the target location using fuzzy membership functions of samples. Results show an improved average estimation error of 2.56 m over the traditional method. Though ultra-wideband (UWB) based indoor localization has proved to be one of the most accurate in estimating target location, it still suffers in the absence of LoS Signal. This situation is common in indoor environments. In [48], SVM is used with Linear Fischer's discriminant in order to identify the LoS signal using CIR. Results show an almost 100% detection of the LoS signal. In [49], an RFID-based indoor location system was implemented. The system uses non-linear SVM to map RSS measurement with target distance in the offline phase. To minimize the cost function of mapping, particle swarm optimization (PSO) is used. Results show an improved localization error of 0.12 m.

E. K-Means

This is an unsupervised data mining technique that is used to group data into clusters. The clusters are centroid-based to make each cluster have a small distance between the elements of the cluster and the centroid. Therefore, elements of a cluster will have similar patterns or properties. The number of clusters is predefined by choosing the K value. A small average intra-cluster distance and a high inter-cluster distance will produce better clustering results [19].

K-Means use-cases: Though not widely utilized as radio signals, the visible light signal has also been used for indoor localization. In [50], a Visible Light Positioning (VLP) system which emulates the RF fingerprinting technique was presented. In the offline stage, two LED transmitters are used to obtain a sparse fingerprint of the environment. Then bilinear interpolation is used to form dense fingerprints which are combined with hierarchical k-means to detect the position of the target in the online stage. Average accuracy of 0.31 m was achieved in a 4.3x4x4 m^3 indoor environment. To address estimation errors caused by changes in the intensity of received light signal, [51] used linear regression in combination with k-means to further improve the LED fingerprint by using linear regression on the formed clusters. Results show an average accuracy of 40 cm in a 5x5x4 m^3 indoor environment. In [52], a k-means-based WiFi localization system that improves the clustering process was presented. In contrast to traditional k-means which forms clusters from the fingerprints based on a single attribute (ToA, TDoA, Phase etc.), a multi-attribute algorithm was used to form clusters based on the best attribute that proved optimum inter and intra-cluster

distance. An accuracy of fewer than 3 m was obtained in 80% of estimations. In [53], bisecting k-means was used to improve the accuracy of WiFi fingerprints. The fingerprints are divided in k-clusters based on the number of Reference Points (RPs) in the offline phase. The location of the user is then estimated in the online phase using the bisected fingerprints. Results show an average accuracy of 1.51 m. In [54], the accuracy of the k-means based indoor location system was improved by introducing a Kalman filter to form a joint algorithm. The algorithm firstly divides the original fingerprint database into K clusters with the k-means clustering algorithm and Gaussian mixture model algorithm. Then the Kalman filtering algorithm is used to process the collected testing signals. Average accuracy of 1.65 m is achieved. K-means was used in [55] to improve the accuracy of an UWB-based indoor localization system. The number of clusters to be used was optimally obtained from the silhouette coefficient. Then, a Kalman filter was used to remove noise from the measured signals before estimating the position of a target. The system achieved a mean error of 9.27 cm. Considering the complexity of generating RSS fingerprints for multi-floor indoor environments, [56] used k-means to reduce the complexity and thereby reducing the time to find the correct floor. The corresponding cluster heads (CHs) of each floor and their labels are sent to the fingerprint database which reduces complexity while maintaining accuracy. Results show fast floor detection, low size of data, and high probability of correctness.

F. Random Forest (RF)

This is one of the most effective data mining techniques. It is also referred to as (Classification and Regression Trees (CART)). The RF algorithm builds an ensemble of trees built by selecting a section of the total variables in an object and a sample of the total objects (bootstrapping) to build a decision tree. This process is repeated several times and the results are averaged. The result improves the accuracy of prediction [19].

RF use-cases: The non-identification of the NLoS signal component of the received signal at the receiver causes a large amount of error in target location detection. Hence [57] designed an RF-based method for detecting the NLoS component of a received signal. In the proposed method, RF is used both for classification and regression to build cooperative supervised localization models which are trained offline using WiFi RSS to obtain the target location. The method is improved to provide a non-prior (non-offline trained) estimation of target location using Gaussian Process Latent Variable Model (GPLVM). Results show a robot location detection with a mean localization accuracy of 36 cm. Similarly, RF was used to identify the LoS signal in [58]. Extracted features from CIR are used together with RF to estimate target location. Results obtained were compared with that of LS-SVM and they showed 97% and 95% accurate detection NLoS and LoS respectively. An RF classifier-based indoor localization system was designed in [59]. The system uses RSS measurements of all APs in order to form a radio map of the indoor environment. Then RF is used for target location estimation. The performance is compared with KNN and JRip classifiers. Results showed a 91% accurate estimation of the target location. Considering the time-consuming nature of obtaining fingerprints and susceptibility of a single of fingerprint (SIOF) based indoor localization systems to NLoS signals and multipath effect, [60] proposed a system that uses group of fingerprints (GOOF) to improve accuracy and reduce the time it takes to build fingerprints. RF-based GOOF is used in the offline phase to form the fingerprint and a sliding window aided mode-based (SWIM) algorithm is used to balance the localization accuracy and time. SWIM showed 100% prediction accuracy at an SNR of 15 dB. Considering the high cost and power consumption of deploying Radio Frequency devices for indoor localization, [61] designed a Software Defined Network (SDN) enabled indoor positioning system which does not require specialized hardware. The system uses RF-based cross-validation to train itself and estimate the target location. Results show an accuracy of 98.3% detection of IoT devices using k-fold cross-validation. In [62], a novel Hybrid indoor localization system "HybLoc" was presented. The system uses Gaussian Mixture Model (GMM)-based soft clustering and Random Decision Forest (RDF) ensembles to locate the target (room-level and latitude-longitude prediction). GMM-based soft clustering allows finding natural data subsets to help cascaded classifiers better learn underlying data dynamics while the RDF ensembles enhance the capabilities of decision trees providing better generalization. HybLoc showed an accuracy of 85% (room level) and 6.2 m (latitude-longitude). In [63], a WiFi-based indoor location recognition system that uses wearables to improve its accuracy was presented. It uses both RSS and Basic Service Set Identifier (BSSID) to solve the problem of various positions with similar RSS. By performing the two-stage filtering and implementation of RF-based location awareness, the accuracy and execution time of the position recognition system is improved. Results show a location recognition accuracy of 97%.

G. Expectation Maximization (EM) Algorithm

The EM algorithm is a data mining technique that is used to handle data with missing values such that the missing values could affect classification tasks. In this process, the algorithm iteratively selects random values for the missing data, then optimizes it iteratively until convergence is achieved. At the point of convergence, the best (optimum) estimates of the missing values are used [19].

EM algorithm use-cases: The problem of non-adaptation of signal strength maps and indoor localization algorithms in SLAM is addressed in [64]. An online training technique using Sequential Monte Carlo (MC) and EM algorithm is used to produce maps of the environment. Therefore, any change in the environment will reflect in the estimation of the map and the location of the device. Results show that optimal localization error is achieved in an unknown environment with 50 runs of MC simulation. In [65], a distributed, online, cooperative localization of multiple

robots and their paths is presented. The system uses an EM algorithm to efficiently identify inlier multi-robot taking uncertainty into account thereby improving the robots' trajectories over time. The system achieved an error of 0.8 m (translation) and 0.04 rad (rotation). The low accuracy of WiFi fingerprinting-based systems caused by variations in RSS and device heterogeneity is addressed in [66] using EM with support set (EMSS). EMSS divides the indoor location into grids and forms a set of grids. When an online RSS is received, the most similar grids are optimally selected based on the Bayesian information criterion. Then EM algorithm is used to select the most likely location. EMSS achieves a Root Mean Square Error (RMSE) of 2.5 m.

H. Neural Networks (NN)

The NN is a perceptron-based approach to understand the relationship between inputs and outputs in data. To do that, weighted inputs are fed into neurons on which activation functions act to produce an output. The weights are selected and assigned to the inputs to obtain the output in a process called Forward propagation. To optimize the output to be more in-line with the inputs, the weights are adjusted from the output in a process called Backward propagation. NN may have one or a few hidden layers between outputs and inputs. NN performs well with unstructured data [19].
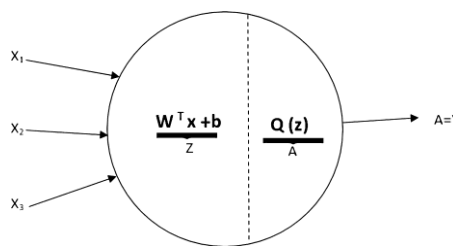

Fig.5. NN with input weighting (z), activation function (Q), and one hidden layer.

NN use-cases: In [67], a Bluetooth-based indoor localization system was designed for Wireless Sensor Networks (WSNs). To reduce the error arising from RSSI measurements, a trained Artificial Neural Network (ANN) was implemented to reduce the errors between the actual and obtained distance between a node and the receiver (mobile phone). A thirty percent (30%) decrease in error was observed in comparison with Centroid Localization (CL) algorithm. In [68], another indoor WSN localization system was presented but considering the best back propagation algorithm in modeling the inaccuracies in RSS. The system achieved an RMSE of 0.4991 m in a range of 100 m for the best algorithm (Elman Backpropagation). Back propagation-based ANN and KNN were combined in [33] to improve the accuracy of systems where room level classification was required before locating targets. Feed forward NNs were used in [69] to improve the performance of a CIR fingerprint-based system. The best structure with the least error was implemented and a location error of less than 0.05 m was obtained with a confidence interval (CI) of 90%. In [70], a Radial Basis Function (RBF) NN was used to design an RFID-based indoor positioning system. A reduced error was achieved by using an Artificial Immune System (AIS) on the NN. The difference in obtained RSS and actual RSS are used to train the network. An accuracy improvement of up to 58% was achieved (0.216942 m) over RBF NN. In [71], NN with Combined deterministic and Modified Monte Carlo (CDMMC) was used to reduce the impact of multipath reflection between an LED (transmitter) and a photodiode (receiver). The technique resulted in an average positioning errors of 1-2 cm.

I. Deep Learning (DL)

This involves mining large unstructured data sets with many hidden features and interdependencies. Hence adjusting weights of a NN will have to go through many layers and activation with each layer to establish an important feature used in computing the final output. Deep learning works very well with large unstructured data. The simple NN is not used for deep learning due to its complexity rather Convolutional NN (CNN) and Recursive NN (RNN) are used [19].

DL use-cases: Though significant progress was made in reducing the multipath effect in indoor environments, it persists. Recently, [72] developed an autoencoder-based technique 'AutLoc' to improve the performance of WiFi RSS fingerprint-based systems. AutLoc uses deep autoencoders to denoise the RSS measured during the offline fingerprinting phase. It then uses ensemble learning at the online phase to estimate the location of the target. AutLoc achieves a mean accuracy of 1.29 m. Similarly, an autoencoder-based Deep Extreme Learning machine (ADELM) was presented in [73]. ADELM uses an autoencoder to improve the accuracy of RSS signals measured in the offline phase. To reduce the training complexity, extreme learning is used rather than traditional technique of training. Results show an accuracy of 92.8% estimation rate for targets. In WiDeep [74], a stack of autoencoders was used in the offline phase to correctly map fingerprints to locations of APs thereby further reducing the effect of noise (from heterogeneous devices) in the environment. Furthermore, a regularization technique was used to boost robustness and avoid overfitting. The location of the target is then estimated in the online phase using a probabilistic estimator. WiDeep achieves a mean

accuracy of 2.6 m (629 m^2 -university building) and 1.21 m (65 m^2-apartment building). In [75], complementary features of deep learning techniques (CNN and DNN) were fused using Dempster-Shafer to design a regression model that can be trained offline and used to estimate the location of targets instantaneously. The initial weights of DNN are assigned using an autoencoder and are optimized by minimizing the mean square error of the model output and target location. Results of the system show an achieved Mean Square Error of 1.48 m. Rather than fusing deep learning techniques, DeepPositioning [76] fuses the pervasive WiFi and Magnetic field fingerprint in order to obtain a finer grain fingerprint. The intrinsic features of the multi-class fingerprint are combined with DNN to train the system in the offline phase, then target device position is estimated in the online phase. DeepPositioning achieves a mean estimation error of 1.45 m. ResLoc [77] uses a finer-grained CSI to form tensors which are used during the offline phase to train the deep learning model. Then new CSI of target devices is used to obtain the position of the target device. ResLoc achieves a median error of 0.98 m when a dual-channel is used. A novel kind of CSI that contains TOF, AoA, and amplitude was used to improve the accuracy of a CSI-based system in [78]. The novel CSI was used in the offline phase to train the CNN model used for location estimation. The system achieved a mean estimation error of 1.5762 m. In an attempt to reduce the effort involved in collecting RSS fingerprints of an environment and making sure it is updated according to changes in the environment, [79] introduced a technique that uses Deep belief networks to extract hidden features from the RSS fingerprints thereby reducing the collection workload. It then applies SVM and KNN to obtain the location of users. Implementation with 10% of fingerprint shows an improvement of 1.9 m in location detection over baseline approaches. CellinDeep [80] utilized a cellular network for indoor localization after identifying the limited ability of existing localization technologies in serving low-end (non-smart) phones. CellinDeep obtains RSS values from a neighboring base station across various geo-tagged positions and uses it to form a fingerprint which is combined with DNN at the offline phase. The trained model is then used to obtain the location of users in the online phase. Techniques to; avoid overfitting, avoid noisy RSS values and reduce training overhead were adopted. CellinDeep achieves a mean error of 0.78 m and it adds no extra energy consumption costs. In [81] and [82] deep learning techniques are used to enhance the performance of holographic-based RFID indoor localization by exploiting the intrinsic properties of a hologram and fixed positions of reference tags.

Table 2. Summary of techniques, technologies, and results of data mining use-cases

Use-case	Data mining technique(s)	Indoor Localization		Results
		Technology(ies)	*Technique(s)*	
INBS [20]	Naive Bayes	WiFi, Smartphone	RSS	Average accuracy of 5.048 m
[21]	Naive Bayes	WiFi	RSS	Mean error of 1.86 m-2.54 m
[22]	Naive Bayes	BLE	Fingerprinting	Accurate class (area) estimation
[23]	Naive Bayes, KNN	WiFi	RSS	Mean error of 1.80m
[24]	Hidden Naive Bayes	WLAN	RSS	Mean error of 1.32 m
[25]	Naive Bayes	WLAN	CSI	Mean accuracy of 80%
[26]	KNN	WiFi	Fingerprinting	Mean error of 1.7928 m
[27]	KNN	WiFi	Distributed CSI	Median accuracy of 1 m
[28]	KNN	RFID	RSS	Accuracy of 0.5 m
[29]	KNN	WiFi	Fingerprinting	Mean accuracy of 1.5 m for 15 APs
[30]	KNN, K-Means	WiFi	RSS	Average error of 1.6 m
IW-KNN [31]	KNN	BLE	Fingerprinting	Average error of 1.5 m
UILoc [32]	KNN	BLE, WiFi	Fingerprinting	Average error of 1.1 m
[33]	KNN, ANN	WiFi	RSS	Accuracy of 0.9 m
[34]	KNN	AM radio	Fingerprinting	Mean error of 2.76 m for k=1
[35]	KNN	Smart phone	Hybrid Fingerprinting	Mean distance error of 0.5 m
[36]	Regression	BLE	AoA	Mean error of 2 ^{0} at 30dB SNR, 0-49^0 elevation
ORRLC [37]	Logistic Regression	WiFi	RSS	Accuracy of 95.8%
[38]	Logistic Regression, DL	WiFi	RSS	Median distance error of 0.97 m
[39]	Regression, KNN	WiFi	Fingerprinting	Average error of 5.12 m
[40]	Regression	WLAN	RSS	Average error of 2.31 m
[41]	Regression	WiFi	RSS	Accuracy of 1.35 m
[42]	Regression	VLC	RSS	Average error of 0.22 m
[43]	SVM	ZigBee	Fingerprinting	Average error of 1 m
[44]	SVM	WiFi	Fingerprinting	Average error of 1.22 m
MalDIP [45]	SVM	WiFi	CSI	Accuracy of 98% cell identification
[46]	SVM	WiFi	Fingerprinting	Average rms error of 1.8 m
[47]	SVM	WiFi	Fingerprinting	Average error of 2.56 m
[48]	SVM	UWB	CIR	Accuracy of approx. 100% LOS identification
[49]	SVM	RFID	RSS	Average error of 0.12 m
[50]	K-Means	VLC	Fingerprinting	Average accuracy of 0.31 m in a 4.3x4x4m^3 area
[51]	K-Means, Regression	VLC	Fingerprinting	Average accuracy of 40 cm in a 5x5x4 m^3 area
[52]	K-Means	WiFi	Fingerprinting	Accuracy below 3 m in 80% of estimations
[54]	K-Means	WiFi	Fingerprinting	Mean error of 1.51 m
[53]	K-Means	WiFi	Fingerprinting	Fast floor estimation, low data size and high accuracy

[56]	K-Means	WiFi	Fingerprinting	Mean accuracy of 1.65 m
[55]	K-Means	UWB	Fingerprinting	Mean error of 9.27 cm
[57]	RF	WiFi	RSS	Average accuracy of ± 36 cm
[59]	RF	WiFi	RSS	Average accuracy of 91%
[58]	RF	WiFi	CIR	Accuracy of 91%(NLOS), 95%(LOS)
[60]	RF	USRP	Fingerprinting	Prediction accuracy of 100% at SNR 15dB
[61]	RF	WiFi	RSS	Accuracy of 98.3% detection of devices
HybLoc [62]	RF	WiFi	RSS	Accuracy of 85% (room level) and 6.2 m (latitude-longitude)
[63]	RF	WiFi, Smart phone	RSS	Accuracy of 98% detection of device location
[64]	EM	WiFi	RSS (SLAM)	Optimal error with 50 runs of MC
[65]	EM	WiFi	RSS (SLAM)	0.8 m (translation) and 0.04 rad (rotation) error
EMSS [66]	EM	WiFi	Fingerprinting	RMSE of 2.5 m
[67]	NN	Bluetooth	RSSI	30% decrease in error over CL algorithm
[68]	NN	WLAN	RSS	RMSE of 0.4991 m in a 100 m range
[69]	NN	WLAN	CIR Fingerprinting	Error of less than 0.05 m with 90% CI
[70]	NN	RFID	RSS	Error of 0.216942 m at 1.5 m from reader
[71]	NN	VLC	RSS	Average error of 1-2 cm
AutLoc [72]	DL	WiFi	Fingerprinting	Mean accuracy of 1.29 m
ADLEM [73]	DL	WiFi	Fingerprinting	Successful estimation rate of 92.8%
WiDeep [74]	DL	WiFi, Smart phone	Fingerprinting	Mean accuracy of 2.6 m (university) and 1.21 m (apartment)
[75]	DL	WiFi	Fingerprinting	Mean Square Error (MSE) of 1.48 m
[79]	DL, SVM, KNN	WiFi	Fingerprinting	Improved existing results by 1.9 m using just 10% of fingerprint data
[76]	DL	WiFi, Smart phone	Fingerprinting	Mean estimation error of 1.45 m
CellinDeep [80]	DL	Cellular	Fingerprinting	Mean error of 0.78 m
ResLoc [77]	DL	WiFi	CSI	Mean error of 0.98 m
[78]	DL	WiFi	CSI	Mean error of 1.5762 m
FaHo [81]	DL	RFID	RSS (Hologram)	Mean error of 1.31 cm (lateral and 5.28 cm (radial)
AdaRF [82]	DL	RFID	RSS (Hologram)	Mean error of 1.79 cm (lateral) and 3.39 cm (radial)

The Table 2 provides a tabular summary of each use-case scenario discussed in the paper. This gives readers an easy tool to compare the results of existing use-cases under each data mining technique. And also allow for easy verification of results of future works that deploy data mining to enhance the performance of indoor localization.

6. Challenges and Future Work

This section explains some of the challenges related to applying data mining techniques to enhance the performance of indoor positioning systems. It also highlights future research work that could be done to address those challenges.

1) IoT applications with varying protocols have been deployed in environments with high mobility and multipath effects [83][84]. The goal of completely eliminating the effects of multipath on position estimation is a daunting task that is still far from being achieved despite significant improvements in techniques that slow it down. With this realization, the next generation of indoor localization systems should define acceptable thresholds in the various deployment environments. Rather than focus on eliminating the multipath effect, the focus should be on working within the set acceptable threshold.

2) Fingerprinting techniques have shown significant improvements in the accuracy of indoor localization. But a pending issue with the fingerprinting technique is the inability of the fingerprint-based models to autonomously adapt to changes in the fingerprints caused by moving objects or obstacles. Therefore, solving this problem using adaptive models would greatly improve the accuracy of fingerprinting-based indoor localization. Another novel technique proposed by authors in [85] that involves dividing an indoor environment into sectors could help in improving the performance of fingerprinting technique.

3) The advent of ultra-low power sensing through (backscatter communication) promises an energy-efficient, low cost and sustainable technique for indoor localization. Recent efforts such as; [83] [84] have shown the possibility of achieving cm-level accuracy with the aid of backscatter communication. Hence, BackCom aided localization could allow for the dense deployment of Reference Nodes (RNs) instead of active transmitters thereby improving the accuracy of indoor localization systems without adding cost. A similar application would be employing BackCom aided motion tracing to replace cameras in certain indoor environments. The data of reflected signals off the body of a person could be mined taking into account the moving reflections and static reflection to track the person's motion. Authors in WiDeo [85] have laid the foundation for further work in that direction.

4) The use of Ensemble/Hybrid data mining techniques has shown an improved performance of data mining tasks. In a similar vein, deep learning techniques have also provided improvements in the accuracy of localization algorithms. Therefore, the next generation of indoor localization systems should incorporate Ensemble (Hybrid) data mining and deep learning techniques. But a challenge would be the non-availability of large amounts of data to train the model. To address this, pre-training of the ensemble/hybrid model with data from similar environments could be used. In addition, deep learning techniques require large amounts of data to train the indoor positioning systems. Therefore, techniques to reduce the latency of mining as a result of increased data size would present a promising research problem.

5) The integration of technologies that could complement one another is a promising technique for improving indoor localization. For instance, combining cellular technology with WiFi. The Base Stations in small cellular cells can provide GPS level accuracy while WiFi can provide cm-level accuracy in indoor environments. Designing an indoor-based system on both technologies would have a significantly higher accuracy especially when the base stations could serve as reference nodes for indoor localization.

6) Another challenge of indoor positioning is the inherent noise generated by the electronic devices used for positioning. These noise signals are random thereby making their effect on the positioning data complex. Research effort towards estimating the effect of noise contributed by the positioning systems and eliminating it will go a long way in improving the quality of positioning data and hence accuracy of position estimation.

7. Conclusion

The notion of enhancing the performance of Indoor location systems is very vital in the current industrial revolution. Though existing works have discussed the subject, we provided a more comprehensive survey in our work by considering various techniques, technologies and data mining techniques. In this paper, an introduction to data mining, its various techniques and how they are used to improve the performance of indoor localization was presented. Despite seeing the improvements shown by adopting data mining techniques in the reduction of location estimation errors in the use-cases discussed, there still exists a certain degree of inaccuracy largely due to multipath effect. These needs addressing through adoption of possible solution such as; hybrid data mining techniques, low energy devices-assisted localization and integrating multiple technologies in indoor localization systems. Though not exhaustive, the future research issues highlighted in the paper provide a list of promising research directions that can address the challenges of data mining-assisted indoor localization and serve as open areas for future work.

Appendix a Summary of abbreviations used

Abbreviation	Meaning	Abbreviation	Meaning
AoA	Angle of Arrival	AP	Access Point
AM	Amplitude Modulation	BackCom	Backscatter Communication
BLE	Bluetooth Low Energy	CDMMC	Combined Deterministic and Modified Monte Carlo
CH	Cluster Head	CIR	Channel Impulse Response
CL	Centroid Localization	CNN	Convolutional Neural Network
CSI	Channel State Information	AIS	Artificial Immune System
DL	Deep Learning	DNN	Deep Neural Network
EM	Expectation Maximization	GMM	Gaussian Mixture Model
GOOF	Group of Fingerprint	HO	Handover
IoT	Internet of Things	KNN	K Nearest Neighbor
LED	Light Emitting Diode	LoS	Line of Sight
LoRA	Long Range Radio	LS-SVM	Least Squares Support Vector Machines
MC	Monte Carlo	MFS	Magnetic Field Strength
MIMO	Multiple Input Multiple Output	MSE	Mean Square Error
MUSIC	Multiple Signal Classification	NLoS	Non-Line of Sight
NN	Neural Network	ORRLC	One-to-all Regularized Logistic Regression Classifier
PCA	Principal Component Analysis	PoA	Phase of Arrival
PSO	Particle Swarm Optimization	RBF	Radial Basis Function
RF	Random Forest	RFID	Radio Frequency Identification
RMSE	Root Mean Square Error	RNN	Recursive Neural Network
RSSI	Received Signal Strength Indicator	SDN	Software Defined Networking
SLAM	Simultaneous Localization and Mapping	ToF	Time of Flight
UHF	Ultra-High Frequency	UWB	Ultra-Wide Band
SVM	Support Vector Machines	VLC	Visible Light Communication
WLAN	Wireless Local Area Network	GPS	Global Positioning System

References

[1] Alagha, A., Singh, S., Mizouni, R., Ouali, A., & Otrok, H. (2019). Data-driven dynamic active node selection for event localization in iot applications-a case study of radiation localization. *IEEE Access, 7, 16168–16183.*

[2] Nguyen, C., & Raza, U. (2019). LEMOn: Wireless Localization for IoT Employing a Location-Unaware Mobile Unit. *IEEE Access, 7, 40488–40502.*

[3] Farnham, T. (2019). Indoor Localisation of IoT Devices by Dynamic Radio Environment Mapping. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)* (pp. 340–345).

[4] Kshirsagar, J., Shue, S., & Conrad, J. (2018). A Survey of Implementation of Multi-Robot Simultaneous Localization and Mapping. In *SoutheastCon 2018* (pp. 1–7).

[5] Van Huynh, N., Hoang, D., Lu, X., Niyato, D., Wang, P., & Kim, D. (2018). Ambient backscatter communications: A contemporary survey. *IEEE Communications Surveys & Tutorials, 20(4), 2889–2922.*

[6] Anoopkumar, M., & Rahman, A. (2016). A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 122–133).

[7] Khan, D. (2008). CAKE–Classifying, Associating and Knowledge DiscovEry-An Approach for Distributed Data Mining (DDM) Using PArallel Data Mining Agents (PADMAs). In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 596–601).

[8] Kaur, B., Ahuja, L., & Kumar, V. (2019). Crime Against Women: Analysis and Prediction Using Data Mining Techniques. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 194–196).

[9] Zamani, Z., Pourmand, M., & Saraee, M. (2010). Application of data mining in traffic management: case of city of Isfahan. In *2010 2nd International Conference on Electronic Computer Technology* (pp. 102–106).

[10] Gandge, Y., & others (2017). A study on various data mining techniques for crop yield prediction. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)* (pp. 420–423).

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

[12] Liu, W., Cheng, Q., Deng, Z., Chen, H., Fu, X., Zheng, X., Zheng, S., Chen, C., & Wang, S. (2019) Survey on CSI-based Indoor Positioning Systems and Recent Advances. In *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–8).

[13] Laoudias, C., Moreira, A., Kim, S., Lee, S., Wirola, L., & Fischione, C. (2018). A survey of enabling technologies for network localization, tracking, and navigation. *IEEE Communications Surveys & Tutorials, 20(4), 3607–3644.*

[14] Jang, B., & Kim, H. (2018). Indoor positioning technologies without offline fingerprinting map: A survey. *IEEE Communications Surveys & Tutorials, 21(1), 508–525.*

[15] Ferreira, A., Fernandes, D., Catarino, A., & Monteiro, J. (2017). Localization and positioning systems for emergency responders: A survey. *IEEE Communications Surveys & Tutorials, 19(4), 2836–2870.*

[16] Jung, S.H., Lee, G., & Han, D. (2017). Methods and tools to construct a global indoor positioning system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(6), 906–919.*

[17] Jung, J.I., Cho, H.W., & Lee, S.S. (2012). A study of data mining method for indoor positioning on smartphones. In *International Conference on Hybrid Information Technology* (pp. 683–697).

[18] Sattarian, M., Rezazadeh, J., Farahbakhsh, R., & Bagheri, A. (2019). Indoor navigation systems based on data mining techniques in internet of things: A survey. *Wireless Networks, 25(3), 1385–1402.*

[19] Aggarwal, C. (2015). *Data mining: the textbook*. Springer.

[20] Zhang, W., Wang, L., Qin, Z., Zheng, X., Sun, L., Jin, N., & Shu, L. (2014). INBS: An Improved Naive Bayes Simple learning approach for accurate indoor localization. In *2014 IEEE International Conference on Communications (ICC)* (pp. 148–153).

[21] Chitsobhuk, O., Warunsin, K., & Udomthanapong, S. (2018). Multidestination Indoor Navigation Using Path Planning and WiFi Fingerprint Localization. In *2018 3rd International Conference on Computer and Communication Systems (ICCCS)* (pp. 406–410).

[22] Farahiyah, D., Romadhoni, R., & Pratomo, S. (2018). Na\"\ive Bayes Classifier for Indoor Positioning using Bluetooth Low Energy. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference* (pp. 181–185).

[23] Nascimento, H., Cavalcanti, F., Rodrigues, E., & Paiva, A. (2017). An algorithm for three-dimensional indoor positioning based on Bayesian inference, Fingerprinting method and Wi-Fi technology. *International Journal of Advanced Engineering Research and Science, 4(10).*

[24] Song, C., Wang, J., & Yuan, G. (2016). Hidden naive bayes indoor fingerprinting localization based on best-discriminating ap selection. *ISPRS International Journal of Geo-Information, 5(10), 189.*

[25] Wu, Z., Xu, Q., Li, J., Fu, C., Xuan, Q., & Xiang, Y. (2018). Passive indoor localization based on csi and naive bayes classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(9), 1566–1577.*

[26] Zhao, L., Wang, H., Li, P., & Liu, J. (2017). An improved WiFi indoor localization method combining channel state information and received signal strength. In *2017 36th Chinese Control Conference (CCC)* (pp. 8964–8969).

[27] Xiao, Y., Zhang, S., Cao, J., Wang, H., & Wang, J. (2017). Exploiting distribution of channel state information for accurate wireless indoor localization. *Computer Communications, 114, 73–83.*

[28] Zhou-guo, H., Fang, L., & Yi, Y. (2017). An improved indoor UHF RFID localization method based on deviation correction. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 1401–1404).

[29] Cheng, L., Li, Y., Zhang, M., & Wang, C. (2018). A fingerprint localization method based on weighted KNN algorithm. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)* (pp. 1271–1275).

[30] Qin, H., Shi, S., & Tong, X. (2019). A New Weighted Indoor Positioning Algorithm Based on the Physical Distance and Clustering. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)* (pp. 237–242).

[31] Peng, Y., Fan, W., Dong, X., & Zhang, X. (2016). An iterative weighted KNN (IW-KNN) based indoor localization method in bluetooth low energy (BLE) environment. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 794–800).

[32] Chen, J., Zhang, Y., & Xue, W. (2018). Unsupervised indoor localization based on Smartphone Sensors, iBeacon and Wi-Fi. *Sensors, 18(5), 1378.*

[33] Adege, A., Yayeh, Y., Berie, G., Lin, H.p., Yen, L., & Li, Y. (2018). Indoor localization using K-nearest neighbor and artificial neural network back propagation algorithms. In *2018 27th Wireless and Optical Communication Conference (WOCC)* (pp. 1–2).

[34] Rahman, M., Moghtadaiee, V., & Dempster, A. (2017). Design of fingerprinting technique for indoor localization using AM radio signals. In *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–7).

[35] Li, P., Yang, X., Yin, Y., Gao, S., & Niu, Q. (2020). Smartphone-Based Indoor Localization With Integrated Fingerprint Signal. *IEEE Access, 8, 33178–33187.*

[36] Khan, A., Wang, S., & Zhu, Z. (2018). Angle-of-arrival estimation using an adaptive machine learning framework. *IEEE Communications Letters, 23(2), 294–297.*

[37] Peng, Z., Xie, Y., Wang, D., & Dong, Z. (2016). One-to-all regularized logistic regression-based classification for WiFi indoor localization. In *2016 IEEE 37th Sarnoff Symposium* (pp. 154–159).

[38] Xiang, C., Zhang, S., Xu, S., Chen, X., Cao, S., Alexandropoulos, G., & Lau, V. (2019). Robust Sub-Meter Level Indoor Localization With a Single WiFi Access Point—Regression Versus Classification. *IEEE Access, 7, 146309–146321.*

[39] Sun, W., Xue, M., Yu, H., Tang, H., & Lin, A. (2018). Augmentation of fingerprints for indoor WiFi localization based on Gaussian process regression. *IEEE Transactions on Vehicular Technology, 67(11), 10896–10905.*

[40] Zhang, L., Ma, L., Xu, Y., & Li, C. (2017). Linear regression algorithm against device diversity for indoor WLAN localization system. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1–6).

[41] Pandey, A., Vamsi, R., & Kumar, S. (2019). Handling Device Heterogeneity and Orientation Using Multistage Regression for GMM Based Localization in IoT Networks. *IEEE Access, 7, 144354–144365.*

[42] Shawky, S., El-Shimy, M., El-Sahn, Z., Rizk, M., & Aly, M. (2017). Improved VLC-based indoor positioning system using a regression approach with conventional RSS techniques. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)* (pp. 904–909).

[43] Farjow, W., Chehri, A., Hussein, M., & Fernando, X. (2011). Support vector machines for indoor sensor localization. In *2011 IEEE Wireless Communications and Networking Conference* (pp. 779–783).

[44] Zhou, R., Chen, J., Lu, X., & Wu, J. (2017). CSI fingerprinting with SVM regression to achieve device-free passive localization. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (pp. 1–9).

[45] Sanam, T., & Godrich, H. (2018). An Improved CSI Based Device Free Indoor Localization Using Machine Learning Based Classification Approach. In *2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 2390–2394).

[46] Abdou, A., Aziem, M., & Aboshosha, A. (2016). An efficient indoor localization system based on Affinity Propagation and Support Vector Regression. In *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)* (pp. 1–7).

[47] Khatab, Z., Moghtadaiee, V., & Ghorashi, S. (2017). A fingerprint-based technique for indoor localization using fuzzy least squares support vector machine. In *2017 Iranian Conference on Electrical Engineering (ICEE)* (pp. 1944–1949).

[48] Kristensen, J., Ginard, M., Jensen, O., & Shen, M. (2019). Non-Line-of-Sight Identification for UWB Indoor Positioning Systems using Support Vector Machines. In *2019 IEEE MTT-S International Wireless Symposium (IWS)* (pp. 1–3).

[49] Yang, L., Liu, Q., Xu, J., Hu, J., & Song, T. (2018). An Indoor RFID Location Algorithm Based on Support Vector Regression and Particle Swarm Optimization. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)* (pp. 1–6).

[50] Elbasiony, R., & Gomaa, W. (2014). WiFi localization for mobile robots based on random forests and GPLVM. In *2014 13th International Conference on Machine Learning and Applications* (pp. 225–230).

[51] Jedari, E., Wu, Z., Rashidzadeh, R., & Saif, M. (2015). Wi-Fi based indoor location positioning employing random forest classifier. In *2015 international conference on indoor positioning and indoor navigation (IPIN)* (pp. 1–5).

[52] Ramadan, M., Sark, V., Gutierrez, J., & Grass, E. (2018). NLOS identification for indoor localization using random forest algorithm. In *WSA 2018; 22nd International ITG Workshop on Smart Antennas* (pp. 1–5).

[53] Guo, X., Ansari, N., Li, L., & Li, H. (2018). Indoor localization by fusing a group of fingerprints based on random forests. *IEEE Internet of Things Journal, 5(6), 4686–4698.*

[54] Gomes, R., Ahsan, M., & Denton, A. (2018). Random forest classifier in SDN framework for user-based indoor localization. In *2018 IEEE International Conference on Electro/Information Technology (EIT)* (pp. 0537–0542).

[55] Akram, B., Akbar, A., & Shafiq, O. (2018). HybLoc: Hybrid indoor Wi-Fi localization using soft clustering-based random decision forest ensembles. *IEEE Access, 6, 38251–38272.*

[56] Lee, S., Kim, J., & Moon, N. (2019). Random forest and WiFi fingerprint-based indoor location recognition system using smart watch. *Human-centric Computing and Information Sciences, 9(1), 6.*

[57] Liu, J., Liu, N., Pan, Z., & You, X. (2018). AutLoc: Deep Autoencoder for Indoor Localization with RSS Fingerprinting. In *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)* (pp. 1–6).

[58] Khatab, Z., Hajihoseini, A., & Ghorashi, S. (2017). A fingerprint method for indoor localization using autoencoder based deep extreme learning machine. *IEEE sensors letters, 2(1), 1–4.*

[59] Abbas, M., Elhamshary, M., Rizk, H., Torki, M., & Youssef, M. (2019). WiDeep: WiFi-based accurate and robust indoor localization system using deep learning. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom* (pp. 1–10).

[60] Zou, J., Guo, X., Li, L., Zhu, S., & Feng, X. (2018). Deep Regression Model for Received Signal Strength based WiFi Localization. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)* (pp. 1–4).

[61] Le, D., Meratnia, N., & Havinga, P. (2018). Unsupervised deep feature learning to reduce the collection of fingerprints for indoor localization using deep belief networks. In *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–7).

[62] Zhang, W., Sengupta, R., Fodero, J., & Li, X. (2017). DeepPositioning: Intelligent fusion of pervasive magnetic field and WiFi fingerprinting for smartphone indoor localization via deep learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 7–13).

[63] Rizk, H., Torki, M., & Youssef, M. (2018). CellinDeep: Robust and accurate cellular-based indoor localization via deep learning. *IEEE Sensors Journal, 19(6), 2305–2312.*

[64] Wang, X., Gao, L., Mao, S., & Pandey, S. (2016). CSI-based fingerprinting for indoor localization: A deep learning approach. *IEEE Transactions on Vehicular Technology, 66(1), 763–776.*

[65] Li, H., Zeng, X., Li, Y., Zhou, S., & Wang, J. (2019). Convolutional neural networks based indoor Wi-Fi localization with a novel kind of CSI images. *China Communications, 16(9), 250–260.*

[66] Xu, H., Wang, D., Zhao, R., & Zhang, Q. (2019). FaHo: deep learning enhanced holographic localization for RFID tags. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems* (pp. 351–363).

[67] Xu, H., Wang, D., Zhao, R., & Zhang, Q. (2019). AdaRF: Adaptive RFID-based Indoor Localization Using Deep Learning Enhanced Holography. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(3), 1–22.*

[68] Dumont, T., & Le Corff, S. (2013). Online EM for indoor simultaneous localization and mapping. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6431–6435).

[69] Dong, J., Nelson, E., Indelman, V., Michael, N., & Dellaert, F. (2015). Distributed real-time cooperative localization and mapping using an uncertainty-aware expectation maximization approach. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 5807–5814).

[70] Guo, X., Li, L., Xu, F., & Ansari, N. (2018). Expectation maximization indoor localization utilizing supporting set for Internet of Things. *IEEE Internet of Things Journal, 6(2), 2573–2582.*

[71] Saadi, M., Ahmad, T., Zhao, Y., & Wuttisttikulkij, L. (2016). An LED based indoor localization system using k-means clustering. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 246–252).

[72] Saadi, M., Saeed, Z., Ahmad, T., Saleem, M., & Wuttisittikulkij, L. (2019). Visible light-based indoor localization using k-means clustering and linear regression. *Transactions on Emerging Telecommunications Technologies, 30(2), e3480.*

[73] Zhong, Y., Wu, F., Zhang, J., & Dong, B. (2016). WiFi indoor localization based on K-means. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)* (pp. 663–667).

[74] Jiang, D. (2019). Bisecting K-Means Based Fingerprint Indoor Localization. In *Simulation Tools and Techniques: 11th International Conference, SIMUtools 2019, Chengdu, China, July 8-10, 2019, Proceedings* (pp. 1).

[75] Razavi, A., Valkama, M., & Lohan, E.S. (2015). K-means fingerprint clustering for low-complexity floor estimation in indoor mobile localization. In *2015 IEEE Globecom Workshops (GC Wkshps)* (pp. 1–7).

[76] Zhang, G., Ke, B., Jiang, T., & Lu, W. (2019). Indoor localization algorithm based on combination of Kalman filter and clustering. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)* (pp. 520–524).

[77] Arsan, T., & Hameez, M. (2019). A Clustering-Based Approach for Improving the Accuracy of UWB Sensor-Based Indoor Positioning System. *Mobile Information Systems, 2019.*

[78] Tuncer, S., & Tuncer, T. (2015). Indoor localization with bluetooth technology using artificial neural networks. In *2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES)* (pp. 213–217).

[79] Ibrahim, A., Rahim, S., & Mohamad, H. (2015). Performance evaluation of RSS-based WSN indoor localization scheme using artificial neural network schemes. In *2015 IEEE 12th Malaysia International Conference on Communications (MICC)* (pp. 300–305).

[80] Zouari, R., Zayani, R., & Bouallegue, R. (2014). Indoor localization based on feed-forward Neural Networks and CIR fingerprinting techniques. In *2014 IEEE Radio and Wireless Symposium (RWS)* (pp. 271–273).

[81] Li, Z., He, G., Li, M., Ma, L., Chen, Q., Huang, J., Cao, J., Feng, S., Gao, H., & Wang, S. (2018). RBF neural network based RFID indoor localization method using artificial immune system. In *2018 Chinese Control And Decision Conference (CCDC)* (pp. 2837–2842).

[82] CAVDAR, A., & Kadir, T. (2018). Artificial Neural Network Based Indoor Positioning in Visible Light Communication Systems. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1–7).

[83] Derek Johnson, Mohammed Ketel,"IoT: Application Protocols and Security", *International Journal of Computer Network and Information Security(IJCNIS)*, Vol.11, No.4, pp.1-8, 2019.DOI: 10.5815/ijcnis.2019.04.01.

[84] Syed Kashan Ali Shah, Waqas Mahmood, " Smart Home Automation Using IOT and its Low Cost Implementation ", *International Journal of Engineering and Manufacturing (IJEM)*, Vol.10, No.5, pp.28-36, 2020. DOI: 10.5815/ijem.2020.05.03.

[85] Ahmed k. Daraj, Alhamzah T. Mohammad, Mahmood F. Mosleh, "Indoor Localization Enhancement Based on Time of Arrival Using Sectoring Method", *International Journal of Intelligent Systems and Applications(IJISA)*, Vol.12, No.3, pp.1-7, 2020. DOI: 10.5815/ijisa.2020.03.01.

[86] Luo, Z., Zhang, Q., Ma, Y., Singh, M., & Adib, F. (2019). 3D backscatter localization for fine-grained robotics. In *16th $\$USENIX$\$ Symposium on Networked Systems Design and Implementation ($\$NSDI$\$ 19)* (pp. 765–782).

[87] Ma, Y., Hui, X., & Kan, E. (2016). 3D real-time indoor localization via broadband nonlinear backscatter in passive devices with centimeter precision. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (pp. 216–229).

[88] Joshi, K., Bharadia, D., Kotaru, M., & Katti, S. (2015). WiDeo: Fine-grained Device-free Motion Tracing using $\$RF$\$ Backscatter. In *12th $\$USENIX$\$ Symposium on Networked Systems Design and Implementation ($\$NSDI$\$ 15)* (pp. 189–204).

Authors' Profiles

Usman S. Toro is currently working as Lecturer in Department of Computer and Communications Engineering, Faculty of Engineering, Abubakar Tafawa Balewa University, Nigeria. He received M.Sc. Electronic Communication and Computer Engineering from Nottingham University in 2014. He is currently working towards a PhD at Shenzhen University, China. His research interests include; Wireless networks, Indoor localization and Low-power IoT Applications

Nasir A. Yakub is currently a Ph.D. Student, at the Laboratory of Computer Science, Robotics and Microelectronics (LIRMM), Universite de Montpellier, France. He received a B.Eng. Computer Engineering from Bayero University Kano (BUK), Kano city, Nigeria, and degree of Master of Engineering (Electrical-Computer and Microelectronic system) from Universiti Teknologi Malaysia (UTM), Skudai, Johor Bahru, Malaysia. He is currently a lecturer in the Department of Computer and Communications Engineering, Abubakar Tafawa Balewa University Bauchi (ATBU), Bauchi, Nigeria. His research interest includes Computer and Robotic Vision, Image processing and Embedded systems.

Aliyu Dala received B.Sc. degree in electrical and electronic engineering from Abubakar Tafawa Balewa University, Bauchi, Nigeria in 2010 and M.Sc. degree in communication engineering from the University of Manchester, Manchester, United Kingdom, in 2013. He is currently working towards the Ph.D. degree from the University of Edinburgh, Edinburgh, United Kingdom. His research focus is on the investigation of radio frequencies in sensing and communication underwater. He is a lecturer at Abubakar Tafawa Balewa University, Nigeria since 2014.

Murtala A. Baba obtained his first degree in 2010 from Abubakar Tafawa Balewa University Bauchi, Nigeria (ATBU) in Electrical/Electronic Engineering. He further received his MSc. Degree in Information Technology from the University of Wolverhampton, U.K, in 2013. In 2020 he got his Ph.D. from School of Electrical Engineering, Faculty of Engineering, University Teknologi Malaysia UTM, Malaysia. His areas of research interest include meta-material antenna for MIMO applications and have been with ATBU Bauchi since 2014 as a lecturer.

Kabiru I. Jahun received a B.Eng. Electrical and Electronics Engineering from the Federal University of Technology, Yola Adamawa State, Nigeria. M.Eng (Communication Engineering) from Universiti Tun Hussein Onn Malaysia, Parit-Raja Johor Malaysia. Currently a Lecturer in the Department of Computer and Communications Engineering, Abubakar Tafawa Balewa University, Bauchi, Nigeria. His research interest includes communication systems, Microwave, Antenna, and Propagation.

Usman I. Bature is currently a Ph.D. Candidate in the Department of Electrical Engineering, Universiti of Teknologi Petronas (UTP) Malaysia. He received a B.Eng. Computer Engineering from Bayero University Kano (BUK), Kano city, Nigeria, and degree of Master of Engineering (Electrical - Computer and Microelectronic system) from Universiti Teknologi Malaysia (UTM), Skudai, Johor Bahru, Malaysia. He is currently a lecturer in the Department of Computer and Communications Engineering, Abubakar Tafawa Balewa University Bauchi (ATBU), Nigeria. His research interest includes Image processing, Nano-material design, embedded systems, and Biomedical Systems.

Abbas M. Hassan is currently working as Lecturer in Department of Computer and Communications Engineering, Faculty of Engineering, Abubakar Tafawa Balewa University, Nigeria. He received M.Eng. Computer and Microelectronic Systems from Universiti Teknologi Malaysia in 2017. His research interests include Computer vision, video and image processing.