JADAVPUR UNIVERSITY

M.Tech. Computer Technology

Analysis Of Sparse Dataset for Multi-city Indoor Localization

A dissertation submitted in partial fulfilment of the
requirements for the award of
Master of Technology in Computer Technology

in

Department of Computer Science and Engineering
Faculty Of Engineering and Technology
Jadavpur University
Kolkata-700032

By

Rituparna Chakraborty
University Roll No.: 002110504025
University Registration No.: 160096 of 202124
Exam Roll No.: M6TCT24020

Under the Guidance of

Dr. Chandreyee Chowdhury
Supervisor
Department of Computer Science and Engineering
Faculty of Engineering and Technology,
Jadavpur University, Kolkata

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that the thesis entitled "**Analysis of Sparse Dataset for Multi-city Indoor Localization**" contains literature survey and original research work by the undersigned candidate, as a part of her degree of **Master of Technology** in **Computer Technology, Jadavpur University**. All the information has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Signed: _____

Date:

Name: Rituparna Chakraborty

Examination Roll No.: 002110504025

Registration No.: M6TCT24020 of 2021 - 2022

Thesis Title: Analysis of Sparse Dataset for Multi-city Indoor Localization

CERTIFICATE OF RECOMMENDATION

This is to certify that the work embodied in this thesis entitled **"Analysis of Sparse Dataset for Multi-city Indoor Localization"** has been satisfactorily completed by **Rituparna Chakraborty, Roll No: 002110504025, University Registration No.: 160096 of 2021-2022**. It is a bona-fide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata, for partial fulfilment of the requirements for the awarding of the **Master of Technology in Computer Technology** degree of the **Department of Computer Science and Engineering**, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021 - 2024.

Dr. Chandreyee Chowdhury
(Supervisor)
Department of Computer Science and Engineering
Jadavpur University
Kolkata - 700032

Forwarded By:

Dr. Nandini Mukhopadhyay
Head of The Department
Department of Computer Science and Engineering
Jadavpur University
Kolkata - 700032

Dr. Dipak Laha
Dean
Faculty Council of Engineering and Technology
Jadavpur University
Kolkata -700032

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Analysis of Sparse Dataset for Multi-city Indoor Localization"** is a bona-fide record of work carried out by **Rituparna Chakraborty** in partial fulfilment of the requirements for the award of the degree of **Master of Technology** in the **Department of Computer Science and Engineering**, **Jadavpur University** during the period of July 2023 to May 2024. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Examiners:

_____

Signature of Examiner

Date:

_____

Signature of Examiner

Date:

Abstract

Faculty of Engineering and Technology, Jadavpur University
Department of Computer Science and Engineering

M.Tech. Computer Technology

By Rituparna Chakraborty

The increasing use of smart technologies in the 21st century is changing the way we live, communicate and work. Outdoor solutions such as GPS (Global Positioning System) are almost 99% accurate for locating an object or person, but are not as effective underground, in a parking lot, inside a shopping center or similar indoor locations. In the era of automation, location-based services (LBS) have become more and more important for indoor navigation of a person or object. A wide range of context-aware applications, a growing number of location-based technologies, and the challenge of achieving greater indoor localization accuracy have greatly increased the volume and scope of research in this area using new fields such as ML, artificial intelligence, and so on and so forth. Indoor localization technologies such as RSSI, fingerprinting, Dead Reckoning, channel state information, scene analysis, and indoor localization technologies such as Wi-Fi, Bluetooth (BLE), ultra-wideband (UWB), and ultrasound are being studied by the researchers to discover the holy grail of indoor localization. High accuracy at lowest cost is a key requirement of LBS. This thesis would focus on indoor localization where complete “Sparse Indoor Loc” data was collected from a multi-city experimental region with 4852 samples, 49 classes and 187 APs. Proposed “Sparse Indoor Positioning” algorithm will be explored to predict the class label with 99% accuracy for selected threshold values of APs in this thesis.

Acknowledgements

The writing of the thesis as well as the related work has been a long journey with input from many individuals, right from the first day till the development of the final thesis. I would like to express my deepest gratitude to my supervisor, Dr. Chandreyee Chowdhury, department of Computer Science and Engineering, Jadavpur University for giving me the opportunity to do research and providing invaluable guidance throughout this work. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has taught me the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I would like to express my sincere, heartfelt gratitude to Manjarini Mallik, research scholar, department of Computer Science and Engineering, Jadavpur University for her patience, guidance, suggestions and moral support in times of need. I am also particularly thankful to Dr. Nandini Mukhopadhyay, Head of the department of Computer Science and Engineering, Jadavpur University for allowing us to carry out research in the department.

I would also like to thank all the faculty members of the department of Computer Science and Engineering of Jadavpur University for their continuous support. This thesis would not have been completed without the inspiration and support of my family and friends, and a number of wonderful individuals including my batchmates of Master of Technology in Jadavpur University — my thanks and appreciation to all of them for being part of this journey and making this thesis possible.

Rituparna Chakraborty

Examination Roll No.: M6TCT24020

Registration No.: 160096 of 2021-2024

Class Roll No.: 002110504025

Contents

1. Introduction 2
1.1 Application of Indoor Localization System 2
1.2 Research Challenges 4
1.3 Motivation 5
1.4 Contribution 6
1.5 Scope of the Work 7
2. Related Work 9
2.1 Table of existing dataset 9
2.2 Comparison of existing works 10
3. Dataset Design 14
3.1 Data Collection 14
3.2 Design of training & test datasets 21
4. Proposal Classification Framework/Algorithm for Sparse Indoor Localization Datasets 25
4.1 Problem statement 25
4.2 System Models 26
4.3 Algorithm Description 26
5. Experimental Results 31
5.1 Dataset Visualization 31
5.2 Result of the Algorithm 36
6. Conclusion 42
7. Future Work 44
References 45

Chapter - 1

1. Introduction

Increasing use of smart technologies in 21st century is changing the way we live, communicate and work. Outdoor solutions like GPS (Global Positioning system) which are accurate in detecting the position of an object or a person, however it is not that effective when underground, inside a parking, inside a shopping mall or any such indoor for that matter. In the age of automation, navigating a person or object indoor has become increasingly important for Location Based Services (LBS). Wide scope of context aware applications, higher number of location-based technologies and the challenge to achieve higher accuracy of indoor localization has increased the number and scope of research significantly in this area using emerging fields like ML, AI and many more.

1.1 Application of Indoor Localization System

Few important applications of Indoor Localization System:

- **User Localization**: User localization in indoor localization techniques is the process of estimating the position of a user or an object inside a building or an enclosed space. It is useful for providing services that rely on the user's location information [2], such as navigation, assistance or security. Some of the common indoor localization techniques are:
    - Wi-Fi fingerprinting: This technique uses the measured intensity of the Wi-Fi signals from different access points

  to create a radio map or a database of the indoor environment. User's location is measured by matching the

  current Wi-Fi measurements with the database. [5]
    - Angle of arrival (AoA): This technique utilizes the time of arrival of the Wi-Fi signals at a set of antennas

  separated by known distance to calculate the angel of arrival. Triangulation is used to estimate the user's

location [22]

- Time of arrival (ToA): This technique uses the time span of the Wi-Fi signal traveling between the emitter and the

receiver to calculate distance between them. Trilateration can be used to estimate the user's location [5].

- Other technologies: Besides Wi-Fi, other wireless technologies can also be used for indoor localization, such as Bluetooth, RFID, ultrasound, or infrared

- **Robot Navigation**: Robot navigation is the ability of a robot to move from one location to another in an indoor environment, where global localization sensors such as GPS cannot be used. Indoor localization techniques are methods that use different sensors and algorithms to estimate the robot's position and orientation within a frame of reference, such as a map or a coordinate system. Some of the common sensors and techniques used for indoor localization are Wi-Fi fingerprinting, BLE beacons, LIDAR, magnetic field, RFID, NFC, AprilTags, and HMM1234 [22]. A high-level sketch of some components in Robot navigation is included in the below.
  Some of the challenges and applications of robot navigation and indoor localization are:
  - The indoor environment may be cluttered, dynamic, or unknown, which requires the robot to adapt to changes and uncertainties.
  - The robot may need to perform tasks such as material handling, disaster relief, patrolling, rescue operation, or assistive technologies.
  - The robot may need to cooperate with other robots or humans in a multi-agent system.
  - The robot may need to optimize its path planning, energy consumption, accuracy, and efficiency. [22]

Moreover, in situations where GPS reception is spotty, indoor localization improves location-based services, asset monitoring, and navigation by allowing accurate positioning and navigation inside buildings of people and objects.

1.2 Research Challenges

Briefly main points of research challenges have been discussed in this section.

1.2.1. Data Annotation: Data annotation for indoor localization involves labeling data that represents various indoor environments. This annotated data is crucial for training machine learning models to accurately determine the position of objects or people within a building using signals from various sensors, such as Wi-Fi access points, Bluetooth beacons. Importance of Data Annotation in Indoor Localization is enhancing model accuracy. Accurate annotations of signal strength and corresponding locations improve the precision of localization models.

Facilitates Supervised learning - labeled data is essential for supervised learning algorithms to learn the relationship between sensor signals and spatial locations. Improves User Experience: High-quality localization models can provide reliable navigation and location-based services, enhancing the user experience in indoor environments.

Data annotation for indoor localization is a critical step in developing accurate and reliable indoor positioning systems. By following systematic annotation processes and best practices, high-quality labeled data can be generated, leading to improved model performance and better localization accuracy. This, in turn, enhances the usability and reliability of indoor navigation and location-based services.

1.2.2. Data Sparsity: In multi-city indoor localization, sparse datasets frequently lack thorough coverage and diversity, which makes it difficult to create precise models that perform well across various buildings and urban settings.

1.2.3. Heterogeneous Environments: Using sparse datasets to establish a universal solution for accurate indoor positioning is challenging due to the enormous heterogeneity in signal propagation introduced by different building architectures and materials across cities.

1.3 Motivation

This thesis is motivated by the urgent need to improve indoor localization systems from a technology-independent aspect in highly-urbanized environments. Sparse data sets, typical in the multi-city context, suffer a variety of challenges (example- lack of data for practically all buildings and cities). Through this thesis, need to find ways to model these sparse data which can be used in various urban environment and not specific due uncertainties of the architecture or signal propagation. This will result in better indoor localization methods [1], benefiting urgent actions (example- emergency response applications), smart navigation and tracking of object or person in indoors, all leading to the progress on smart city technologies [2]. This thesis is primarily inspired by the gap often neglected in prior indoor localization studies/work on the matter of different architectural styles and building materials existing across cities. This thesis seeks to construct data having accounted for these variations (hence data collected from various experimental regions lab, office, canteen, restaurant etc.) and used it to increase the robustness and accuracy of localization models. Its purpose is to analyze sparse datasets, as one experimental region access point is not available in the other different city experimental region.

However, there are few methods [4] to enhance indoor localization using sparse Wi-Fi data, demonstrating improved accuracy in single-building scenarios. But this thesis will extend the work by focusing on multi-city scenarios [6], where the

variability between cities poses additional challenges. The thesis will develop scalable localization solutions considering all the above-mentioned facts that maintain high accuracy despite data sparsity [7]. This contribution will bridge the gap between localized indoor systems and a more universal approach applicable to diverse urban environments.

1.4 Contribution

Keeping "the accurate indoor positioning is challenging due to the enormous heterogeneity in signal propagation introduced by different building architectures and materials across cities" in mind this work proposes a novel indoor localization approach for multi-city environment.

The proposed approach addresses the research challenge to train a localization model on multiple indoor distinct regions.

- This work introduces a new RSSI fingerprint dataset collected from multi-city environment and from various experimental region. In multi-city 5 different experimental region considered. Different experimental regions are considered to ensure the localization model's/algorithm's robustness and generalizability across diverse urban environments with varying architectural styles, materials, and layouts.
- This dataset is sparse in nature as the features available at one experimental region (city) is not available at other experimental regions.
- The proposed algorithm "Sparse Indoor Positioning" works on two-way feature selection strategy. The location specific least active and most active features are divided into 2 groups for location estimation.
- More than 99% localization accuracy has been achieved using the proposed Sparse Indoor Positioning algorithm for selected threshold values for APs.

1.5 Scope of the Work

The remaining work is organized as follows – mentioned about the Related Work to the existing dataset in Chapter 2. Next go into great detail on the dataset design, collection and the division plan about training and test dataset in chapter 3. After that mentioned about proposed classification Framework/Algorithm for Sparse Indoor Localization Dataset in Chapter 4. Chapter 5 covers the topic of Experimental Results and Dataset Visualization along with result of the proposed "Sparse Indoor Positioning" algorithm. Chapter 6 covers the conclusion about the thesis work. Finally, wrap up discussion of potential future developments based on the Sparse Dataset for Multi-city Indoor Localization in Chapter 7.

Chapter - 2

2. Related Work

2.1 Table of existing dataset

Dataset Name	Main Features	Limitations
UJIIndoorLoc [14]	Consisted of 21,049 samples captured from 3 different buildings with 4 floors each by 21 devices. 33 reference points were covered and 520 Aps were identified	Comprehensive benchmark is missing
The Mannheim/compass dataset [15]	Wi-Fi based dataset was recorded from the 2nd floor of a building. Training data were captured from 130 reference points with a spacing of 1.5m. Test data were captured from 46 reference points and selected in a random fashion. The access points were covered by 12 Aps and additional 13 Aps were installed	Data were collected from only corridor. Number of devices used to capture data not mentioned clearly.
IPIN 2016 Off-Site Competition [16]	Dataset used in off-site competition of IPIN 2016 conference, sensor and RSSI samples were acquired from 4 buildings by multiple devices (approximately 6)	Data were collected from some trajectories and stored in the log files in a complex pattern. A parser, written in MATLAB, was provided to read the log files in an expected format.

ALCALA 2017 Tutorial dataset [17]	Data were collected in the same corridor as the IPIN 2016 Tutorial. The dataset consisted of 670 training samples and 405 testing samples. 154 attributes of the dataset contained the Wi-Fi fingerprint data from 152 Aps and the GPS coordinates where it was recorded.	Some training fingerprints were mislabelled for testing. A finer grid was used for data collection, but the grid size is not mentioned.
IPIN 2016 Tutorial dataset [17]	Considered 927 data samples for training and 702 for testing. 177 attributes of this dataset contained the Wi-Fi fingerprint data from 16 APs,	Small corridor was covered, to investigates small scenarios.
GEOTECH [18] laboratory dataset	Dataset was recorded by one device and consisted of 281 continuous samples, among them 270 were for training and 11 were for testing. Contained magnetometer, accelerometer and orientation sensor data.	Sensor data were found to mostly depend on user velocity the orientation of the device.

Table 1: Existing data set table

2.2 Comparison of existing works

UJIIndoorLoc: Stands apart as one of the most exhaustive datasets, offering a different scope of multi-building and multi-floor WLAN fingerprints. With broad

inclusion and fastidious comments, it fills in as a benchmark for indoor restriction research [14]. Be that as it may, its huge size might present difficulties for specific applications, requiring the requirement for more sensible datasets [1].

LocoPositioning: Gives a decent mix of BLE and Wi-Fi fingerprints custom-made for indoor following. While its size is moderate and explanations are accessible, the inclusion is restricted contrasted with bigger datasets [2]. This dataset provides food well to applications requiring moderate-scale indoor restriction.

OpenLocate: Offers an open-source Wi-Fi RSSI dataset, yet with restricted size and inclusion. While it needs point by point explanations, it fills in as a significant asset for specialists investigating Wi-Fi-based indoor restriction techniques, especially in situations with obliged assets [1].

Microsoft Indoor dataset: Organized by Microsoft Exploration, offers a moderate-sized assortment of Wi-Fi RSSI information. It finds some kind of harmony among size and inclusion, making it reasonable for different examination applications. Also, the dataset is all around clarified, working with calculation assessment and examination [3].

MIT Wi-Fi dataset: Got from Wi-Fi RSSI information gathered on the MIT grounds, flaunts broad inclusion and explanations. With its huge size and complete inclusion, it fills in as a strong benchmark for assessing indoor limitation calculations [4].

JUIndoorLoc: The dataset contains Wi-Fi signal strength data collected from various access points in the indoor environment. This information is used to determine the exact location of devices in the building using machine learning techniques [5].

The exact file size is not specified, but it contains extensive signal strength measurements from multiple access points, making it an important dataset for indoor localization tasks.

The material covers different floors and areas of Jadavpur University premises and provides a detailed and versatile data point for effective indoor localization.

Chapter - 3

3. Dataset Design

Data was collected at two cities Kolkata and Chandan Nagar, from various indoor regions under various ambient conditions. The experimental regions consist of Newtown Office, Restaurant, Hospital, Jadavpur University-Lab, Jadavpur University-Canteen, Jadavpur University-RoyBuilding, Jadavpur University-Departmental laboratory. Each experimental region was covered by distinct set of APs. The "Sparse Indoor Loc" dataset is consisting of 187 APs in total. Since data was collected from diverse areas, each AP cannot cover each region. Thus, there are multiple location points in the dataset that do not have any signal strength value from multiple APs, causing no records.

Causing no records in many locations as datasets with missing values cannot be fed to any ML/DL model. These missing records are filled with dummy values which is -110 dBm for this work.

3.1 Data Collection

The framework involves sensors capturing indoor environment information, which is then processed through data preprocessing and feature extraction. The data in this work has been captured using Wi-Fi access points [14] [Figure 3].

In this work the raw data has been collected from multiple locations like Newtown Office, Restaurant, Hospital, Jadavpur University-Lab, Jadavpur University-Canteen, Jadavpur University-Roy Building, Jadavpur University Departmental laboratory from Kolkata and Chandan Nagar. Experimental data is also captured from multiple positions from every location. Finally, the data captured was consolidated in excel sheets for processing. The app captures timestamps, MAC address and signal strength values in the raw data [1].

Timestamp	MAC Address	RSSI Value
1688562157.55473	00:a0:0a:27:ab:00	-90
1688562158.42608	00:a0:0a:27:ab:00	-70

Figure 1: Example of Data Entry in Raw collected dataset

Data was collected from diverse unbiased indoor spaces to compile a huge and trustworthy data set. These locations include office, a restaurant, a hospital, a laboratory (LAB), a canteen, a specific building (ROY_Building), and departmental laboratory dataset (Complete_ Departmental laboratory) [8].

Newtown Office Dataset: In the office environment data was gathered in five classes individually. Generally, an enclosed space consists of walls, office cubicles, laptops, phones and other electronic equipment. Such circumstances imply a more complicated indoor localization problem as signals will be distorted and disrupted as a result of the physical obstacles.

Such data sets include:

- RSSI measurements: Received signal strength indicator (RSSI) values from Wi-Fi access points, Bluetooth signals, or other radio frequency sources.
- Floor plans: Detailed floor plans of the office environment, including dimensions, spaces and obstructions.
- Ground true stations: The exact coordinates of the known reference points of the office are the basis of localization accuracy.
- Timestamps: Timestamps associated with each data point to synchronize and correlate measurements. Environmental Factors: Information about environmental conditions such as temperature, humidity and presence of obstacles or people.

Stands apart as one of the most exhaustive datasets, offering a different scope of multi-building and multi-floor WLAN fingerprints.

Restaurant Dataset: The environment mirrored that of a restaurant in a class two setting. It could be considered as open-planned spaces with few walls, many people, and different types of density and thus this creates a few distinctive challenges to the office. RSSI here might be computerized due to other people and movable furniture that are around.

Hospital: The ambiance at Hospital contained seven classes. The reality of a hospital is filled with many rooms, devices and potential sources of interference, which provides an excellent opportunity to test the efficiency and workability of an algorithm in an unsteady area [8].

Jadavpur University-Lab Dataset: The Lab-Jadavpur University Dataset is a medium-sized dataset collected specifically for indoor localization research. Here are the key points about this dataset: Source and Collection, Purpose, Data Characteristics, Size, and Insights gained from the dataset help in designing more accurate and reliable indoor navigation systems. A laboratory data was gathered from eight classes. Labs mostly are equipped with a special type of gear which may sometimes result in electromagnetic interference making the problem of RSSI-based localization a special kind.

Jadavpur University-Canteen Dataset: The canteen had six groups of classes. The canteen, like the restaurant, is both a semi-open and a more flexible place to inhabit and get variations of space and furniture due to the movement of people. This factor results to a more dynamic and variable localization trend [9]. Here are some key points from this data set:

- Location: Data are collected from the Jadavpur University canteen, which provides a controlled environment for indoor localization experiments.
- Purpose: The purpose of the dataset is to facilitate the development and testing of indoor positioning systems, particularly those using Wi-Fi

fingerprinting, Received Signal Strength Indicator (RSSI) measurement, and other signal-based positioning methods.

- Collected data: The dataset usually includes various signal measurements such as RSSI values from Wi-Fi hotspots, timestamps and possibly ground truth locations. These measurements are taken from several reference points within the canteen.

Jadavpur University-Roy Building Dataset: The ROY Building environment had the largest number of lessons, numbering up to 18. Also, heterogeneous interior architecture of the building is to be expected with multi-level spaces and different-sized rooms which, in turn, results in more complex or tougher indoor localization task. Key highlights are:

- Collection Environment: Data was collected from the Roy Building of Jadavpur University, which offers a diverse and complex indoor environment typical of university buildings, including classrooms, corridors, stairs and open spaces.
- Data: The dataset contains various sensor readings such as Wi-Fi signal strength (RSSI values), Bluetooth signal strength, typically used for indoor localization. Annotations: Each data point in the dataset is accompanied by precise ground truth location information, usually represented as coordinates or spatial identifiers, enabling localization algorithms to be evaluated accurately.
- Size and diversity: As a large dataset, it provides comprehensive coverage of various scenarios and conditions, including human presence, structural obstructions and signal strength variations at different times of the day.
- Research Utility: This dataset is invaluable for developing and testing indoor localization algorithms, as it allows researchers to train and validate their models in real environments.

- Benchmarking: The dataset can be used as a benchmark to compare the performance of different indoor localization techniques, helping to advance the industry by providing a standard dataset for evaluation.

Jadavpur University Departmental Laboratory Dataset: The Jadavpur University Departmental laboratory data is a large dataset collected specifically for indoor localization research. This includes extensive measurements of Wi-Fi signal strengths (RSSI) from various access points in the university building. The purpose of the dataset is to provide real-world information for the development and evaluation of indoor positioning algorithms, using signal strength data to determine a precise location in an indoor environment. This dataset is valuable for advancing research in areas such as machine learning-based localization, fingerprinting techniques and signal processing methods in indoor positioning systems. The Complete_ Departmental laboratory dataset divided into three classes. This environment was well suited for us and we could check out its power to generalize across different environments.

Data Stratification and Analysis: Then the data was divided into training and test sets each of 80% and 20%. Each set is used for training and testing respectively. Hence, this stratification is embedded form which the model can be trained on different data points and the accuracy is evaluated on a separate dataset. [9]

Each AP was considered individually and -110 in all lines with the same classes of objects. The calculated specific AP values were useful in comprehending the AP's performance characteristics and dependability in varying conditions. A . csv format table is drawn in which the values of the exact classes and AP are entered, which is then revised to model improvement [11].

Data Preprocessing

Data processing is an essential element in ensuring the integrity of the data to be used.

- **Normalization:** Levelling out RSSI values from different scales for assuring the same contribution of each in the model training process.
- **Outlier Detection:** The place of the device and the replacement or rectification of any wrong RSSI readings that can possibly impact the model results [2].

Data preprocessing involves interpretation of missing values and transforming raw sensor data for meaningful features using signal processing and statistical analysis.

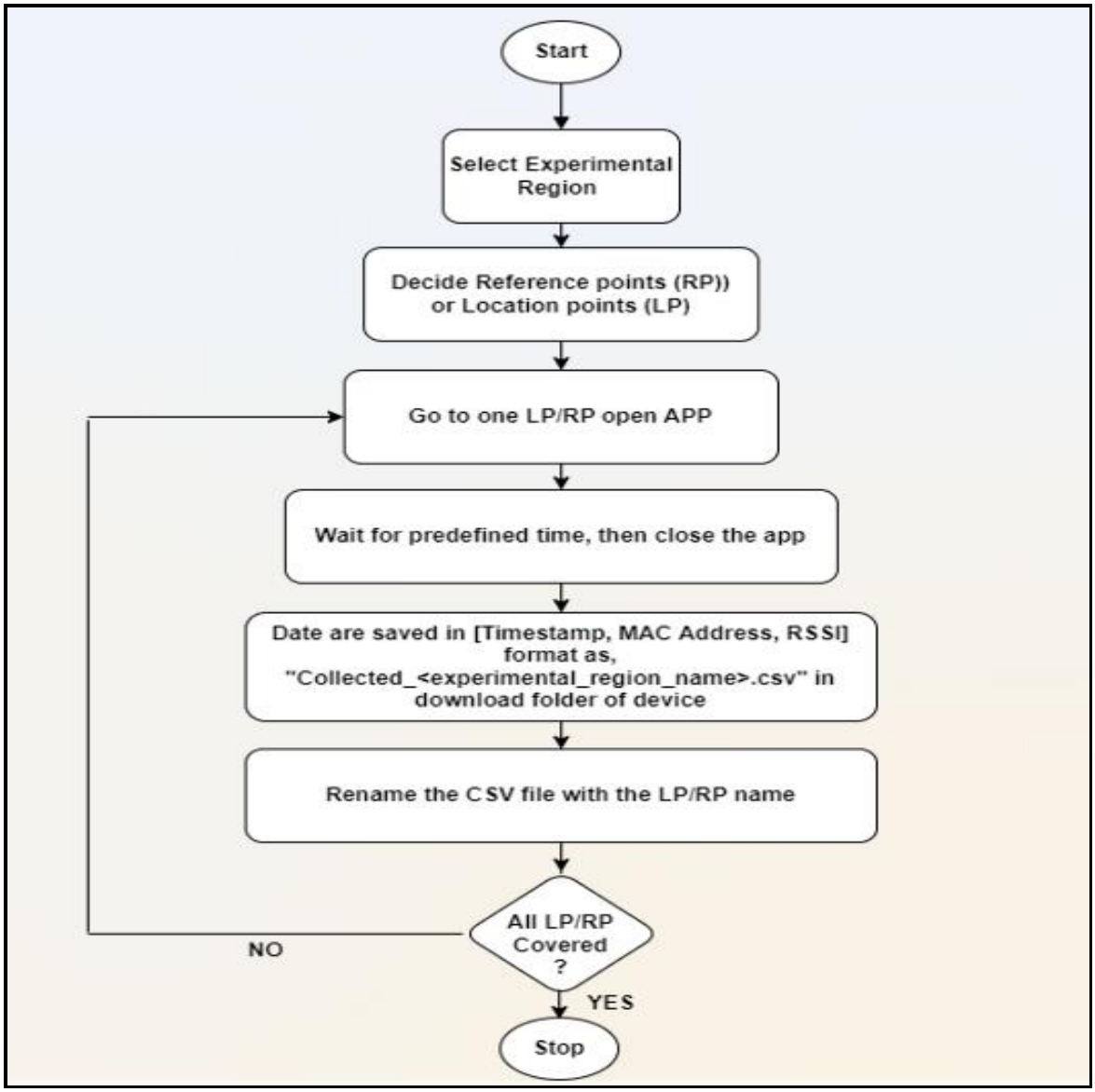
The process followed in this work are:

- Raw data was collected and processed using computing techniques and then converted to structured data with Timestamp, Cellid (for location tagging), MAC Addresses (segregated with multiple different addresses) along with signal strength values.
- As a next step the processed data is further tagged with Timestamp, Cellid, CLASS and AP values from AP1 to AP187 was captured.
- Based on Cellid (location tagging) input CLASS label is assigned in each dataset that will help to identify the actual position/space for the assigned set of access points.
- For a specific location, entire data captured are joined together using Python program developed, to convert into a complete “Sparse Indoor Loc” dataset for the specific location. Example - Newtown Office raw data collected from 5 different locations and each location created 5 different

excel/csv for the raw data. Similar process followed for each individual dataset.

- The processed data is fed as input to machine learning supervised classification models, using algorithms like K-nearest neighbor (kNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RT), or Artificial Neural Network (ANN) classifier, trained on labeled data.
- Prediction of the location of new unlabeled data is performed by using trained machine learning models.
- The localization results can be represented as coordinates or labels indicating the estimated location of the person or object being [18] localized. Flowchart [1] overviews indoor localization data collection framework however implementation details may vary based on system requirements and sensor technologies.

WIFI hotspot RSSIs are eliminated from the gathered dataset due to movable and temporary nature of such signals. Noise is thus introduced into these recordings throughout the localization prediction process. Missing entries are result of not hearing from all of the APs at all locations. Prior to the dataset's analysis, these missing items are set to a dummy value that doesn't show up in the collected dataset. The gathered dataset RSSIs range from -40dBm to -100dBm. Considering these missing entries are filled with −110dBm.



Flowchart 1: Data Collection Procedure

3.2 Design of training & test datasets

The indoor localization problem is also resolved by machine learning techniques. This method starts by creating a training model from the data that was gathered during the offline stage. Subsequently, during the online phase, a machine learning classifier uses the previously created training model to analyze test data collected at the current instance in order to estimate an unknown location. To improve localization accuracy and stability, researchers used supervised

learning [8], semi-supervised [19], unsupervised [20], and classifier fusion [9] based techniques.

A train set is built during the learning and classification phase taking into account the cell size, cs, and the RSSI data of APs. To build the classification model, this train set is subjected to cutting-edge classifiers. Based on the test dataset(s) that were obtained from the unknown user or device, this classification model is utilized to estimate their position.

In this work the train and test datasets were divided into 80:20 format to get the quality accuracy. Also, **stratify** parameter from machine learning is used when a dataset is split into a training set and a test set to ensure that the distribution of a given variable (usually the target variable) is preserved in both subsets. This is especially important for unbalanced data sets, where some classes are much more common than others. Through use of stratify, it is ensured that the proportion of each class in the target variable is approximately the same in both the training and testing. This consistency helps train models that generalize better to unseen data. Without stratification, there is a risk that the training or testing may contain a disproportionate number of samples from certain classes, leading to model bias and poor generalization. If the distribution of the training set is similar to the original data set, the model can learn more effectively from each class and improve its performance, especially in minority classes. A test set reflecting the original class distribution provides a more reliable assessment of model performance based on real data.

"**Random state**" is another variable that has been used to ensure repeatability of results. Specifying a random state (often called a seed) allows to get the same random distribution of data with every code run, which is critical for debugging, comparing results, and ensuring consistency of experiments. By setting random state, ensure that random processes in code (such as splitting data into training and test sets) restoring neural network weights, etc.) produces the same results every time the code is run. This is necessary to ensure the reproducibility of the experiments so that others (or later) can accurately validate and compare the results. When tuning hyperparameters or comparing different models, it is important that the distribution of the data remain the same to ensure a fair comparison. Setting the random mode ensures that the training and test set are identical at different times. Fixed random state makes it easier to debug problems because results are consistent between different runs. If something goes wrong or notice unexpected behavior, can repeat the same scenario and find out the cause more efficiently. In this code Random State 9 was used.

Computed accuracy of each individual dataset for different classifiers like K-nearest neighbor (kNN), support vector machine (SVM), Gaussian Naive Bayes (GNB), Decision Trees (DTs), Random Forest (RF) and Artificial Neural Network (ANN) finetuning with specific configuration. For SVM considered kernel as 'rbf' and 'gamma' parameter as 'auto'. In DT considered minimal sample split from 2 to 21 and calculated accuracy for each dataset.

Chapter - 4

4. Proposal Classification Framework/Algorithm for Sparse Indoor Localization Datasets

4.1 Problem statement

Collected datasets from various location for indoor localization are often Sparse due to weak signal and variable signal strength from the Wi-Fi access points. If only few APs are present or operating at any given location and then incomplete and noisy signal fingerprints are generated typically the value of these signal are -110 dBm. These APs for the specific location cannot be considered to identify the actual location of the object or person in an indoor location. Hence multiple distinct indoor area there exist sufficient indoor localization systems that are developed for individual specific indoor areas for example some specific office or university, building, hospital, shopping mall etc.

As a result, one needs to install different localization application on his/her device to get location information for multiple desired indoor areas. But is it possible to get location information for multiple distinct indoor areas from one single localization system. How localization accuracy gets effected if a localization system covers different indoor areas present at a significant distance or even at multiple cities?

This query remains unaddressed in existing research in RSSI based indoor localization.

The proposed work investigates the feasibility and efficiency of a multi-city indoor localization.

4.2 System Models

System Configurations For Proposed Algorithm Sparse Indoor Loc

Processor	Intel(R) Core(TM) i3-8145U CPU @ 2.10GHz 2.30 GHz
Installed RAM	12.0 GB (11.9 GB usable)
Device ID	D77D247A-0ED0-4641-ACA2-5AED31C0CBF9
Product ID	00331-10000-00001-AA323
System Type	64-bit operating system, x64-based processor
Edition	Windows 11 Pro
Version	23H2
OS Build	22631.3593
Experience	Windows Feature Experience Pack 1000.22700.1003.0
Python Version	3.10.12

Figure 2: System configurations

After multiple tests run the above configuration of the system has been found to be perfect for optimum performance of the "Sparse Indoor Positioning" algorithm.

4.3 Algorithm Description

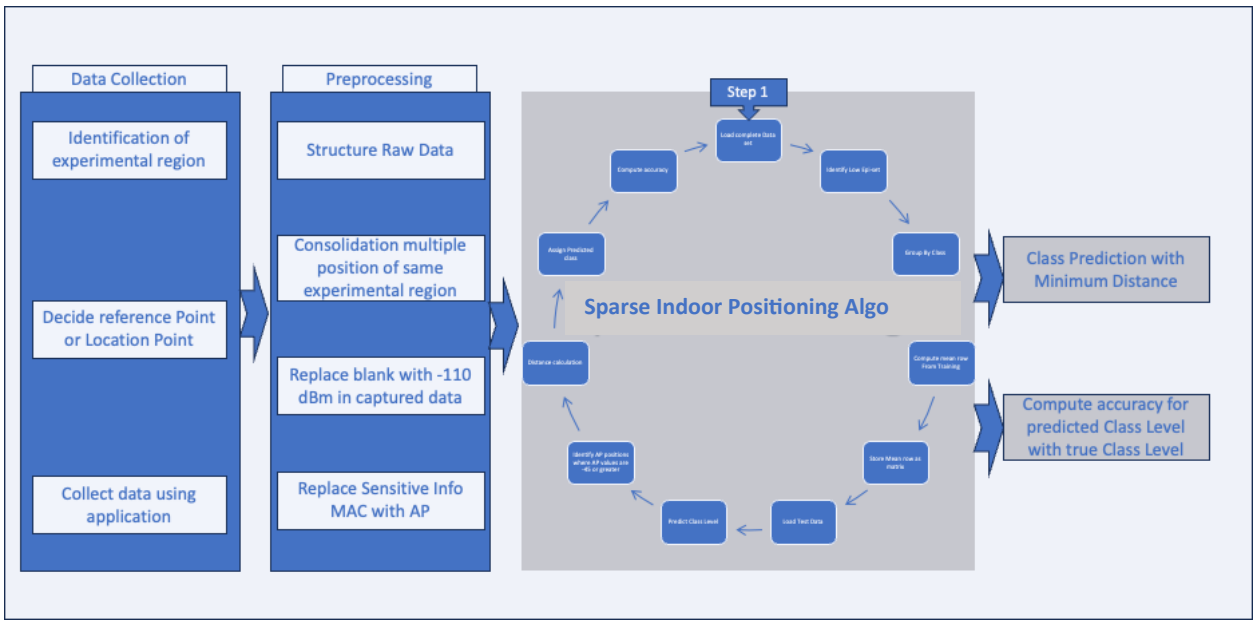


Figure 3: Algorithm pictorial representation

“Sparse Indoor Positioning” algorithm is created to solve above stated problem statement [4.1] existing in previous works, using Machine Learning Methods. To achieve this rigorous step by step process was followed including data capturing, data preprocessing, data sanitization, data labeling, prediction, Python Code development using machine learning for repetitive high accuracy outcome, followed by accuracy determination testing for complete “Sparse Indoor Loc” data sets and finally documenting the entire process to ensure future enhancements and reference. Subsequent sections have covered these stages in detail.

The input to the Algorithm is finally a complete dataset set which is elaborated in earlier sections.

The “Sparse Indoor Loc” dataset for the proposed algorithm was generated using office, a restaurant, a hospital, a laboratory (LAB), a canteen, a specific building (ROY_Building), and departmental laboratory [8] dataset all together.

The **“Sparse Indoor positioning”** Algorithm accepts the **“Sparse Indoor Loc”** dataset as input and takes it though different steps as following to finally produce the desired outcome which is maximum accuracy and high prediction for class labels dataset of multi-city and multi-location.

The flow of the algorithm goes as follows:

Step 1- Data Preparation:

Data preprocessing contains normalization, outlier detection and dimensionality reduction as its steps for cleaning and standardize the feature vector [8]. This uses machine learning techniques to detect locations of devices based on RSSI signals from wireless APs captured from Multi city, Multi location experimental region. To generate a balanced consistent result, attributes like mean, variance, and

spatial distribution from RSSI values are used as the main steps for feature extraction.

The steps are:

- Load the complete “Sparse Indoor Loc” dataset.
- Identify Low AP Set
- For each sample, identify AP values that are -70 in all training samples.
- Group by Class
- Group the training data by CLASS label

Step 2- Training Phase

- Compute Mean Row
- For each CLASS compute the mean values across all samples in the group.
- Store Mean Rows
- Create a matrix to store the mean rows for each CLASS.
- The mean row matrix is used as the reference during the prediction phase.

Step 3- Prediction Phase

- Load Test Data:
- Load the test data containing values without CLASS labels.
- Identify Significant AP Values:
- For each test sample, identify AP positions where AP values are -45 or greater.
- Distance Calculation:
- For each test sample, calculate the distance between the sample's RSSI vector and the mean vectors in the matrix.
- Only consider the AP positions where AP values are -45 or greater.
- Assign CLASS:

- Identify the CLASS with the minimum distance for each test sample.
- Assign the corresponding CLASS to the test sample's prediction (Y_pred).

Step 4- Evaluation

- Compute Accuracy:
- Compare predicted CLASS (Y_pred) with true CLASS labels in the test set.
- Compute the accuracy of the predictions.
- Print Accuracy to output file

Chapter - 5

5. Experimental Results

5.1 Dataset Visualization

Data collected from diverse unbiased indoor spaces to compile a huge and trustworthy data set. Where Newtown Office consisting with 5 classes, 187 APs with 870 samples, Restaurant – consisting with 2 classes, 15APs with 75 samples, Hospital consisting with 7 classes, 68APs with 517 samples, Jadavpur University-Lab consisting with 8 classes, 62 APs with 424 samples, Jadavpur University-Canteen consisting with 6 classes, 27 APs and 270 samples, Jadavpur University-RoyBuilding consisting with 18 classes, 1 APs and 982 samples, Jadavpur University- Departmental laboratory consisting with 3 classes, 35 APs and 1716 samples from Kolkata and Chandan Nagar experimental regions.

Data collected from different experimental regions are visualized with different visualization methods of python, as follows:

Visualization

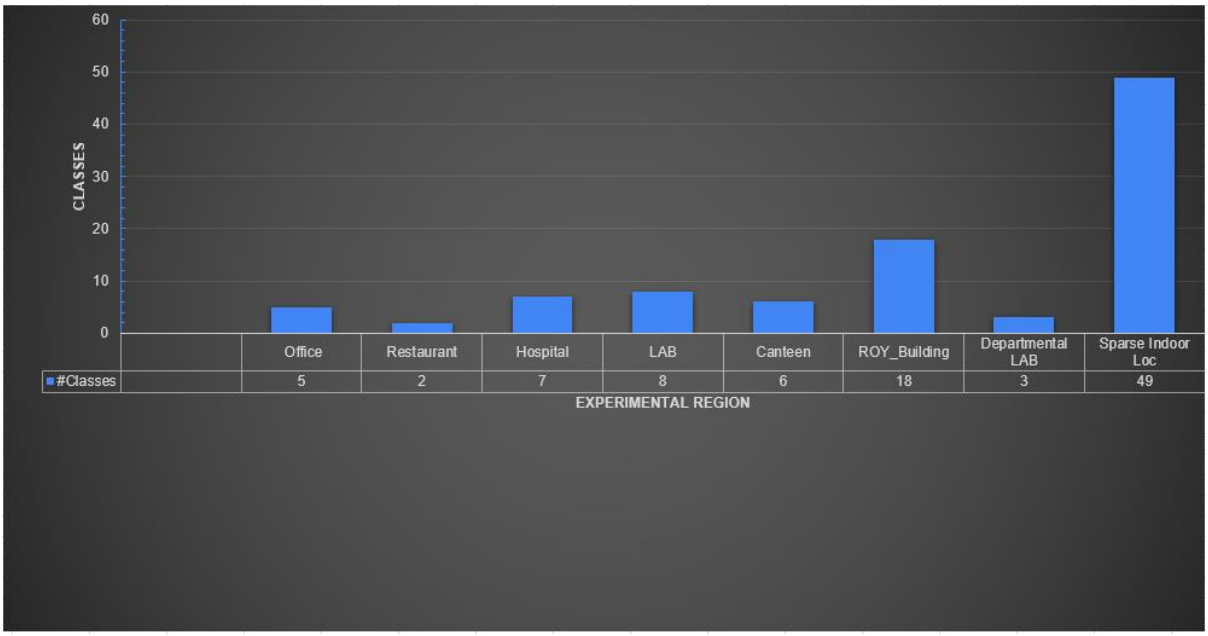


Figure 4: The distribution of classes across different experimental regions

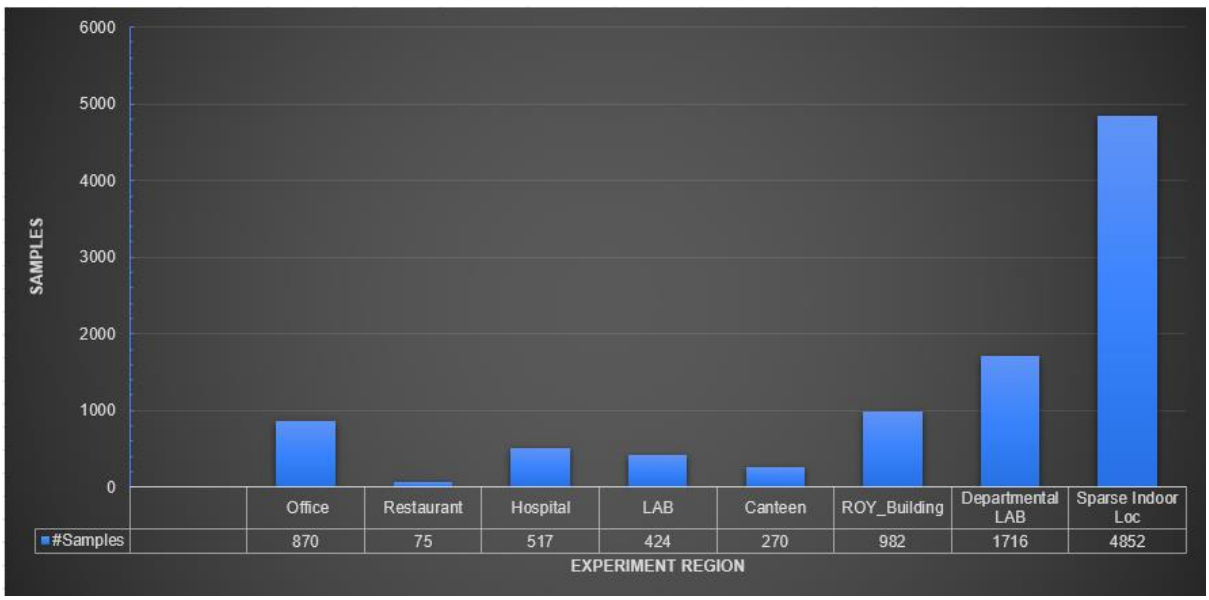


Figure 5: The distribution of samples across different experimental regions

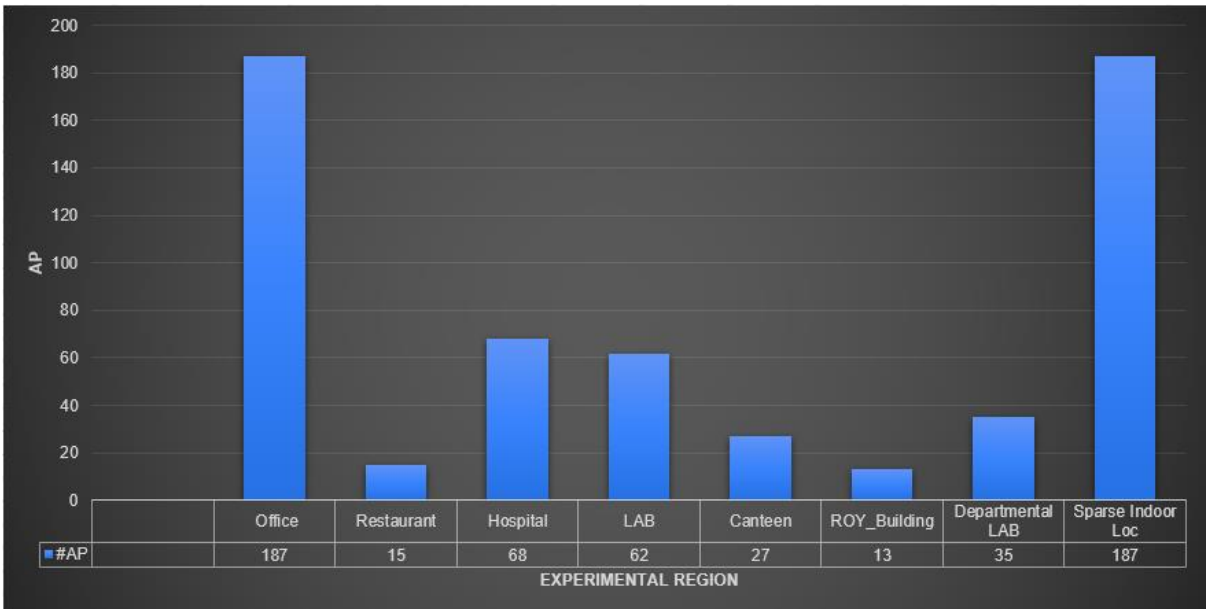


Figure 6: The distribution of APs across different experimental regions

Distribution of Classes: Above Figure [4] is created to show the range of the classes on experimental zones. The variation in the range of classes and examples present in each of the environments, show the format as well as the complexity of the dataset. This bar graph depicts the ratio of samples extracted from different labs, an office, restaurant, a hospital, a lab, a canteen, and the ROY Building. Each the bar shows the number of classes and the number of samples for the selected area indicating the structure and complexity of the data in an easily understandable way [12]. For example, the ROY Building has the largest number of classes (18) and samples (982), which indicates the detailed and inclusive nature of the data set. The Restaurant is a dataset of only 2 classes and 75 samples, so it stands apart from other datasets that are larger and more complex. The heterogeneity of the number of classes as well as samples results with inhomogeneous ecological complexity and data collecting for each region differently [10].

Accuracy Comparison: As represented in Figure [7] line plot was utilized to evaluate each of the classifiers against each other in the experimental area. This graph clearly shows how models have worked in different scenarios, and thus, the ones that are more robust as well as generalizable become apparent. The line graph shows the classification models (kNN, SVM, GNB, DT, RF, ANN) performances in several sectors (experimental region A, B, C,…). This model shows of the findings from the offices, restaurants, hospitals, labs, canteen, ROY Building, and Complete Departmental laboratory on the diagram [13].

It enables an optimizer to identify which models perform best in different scenarios. High accuracy level in all aspects, often displayed by Random Forests and Decision Trees, shows their robustness and generalization. There is a type of model, which is Artificial Neural Networks that could be unstable since they could give high performances in some designs and pull down in others. Hence in doing these comparisons, it is revealed the unique attributes of each model to recommend the right model for a particular indoor localization scenario [13].

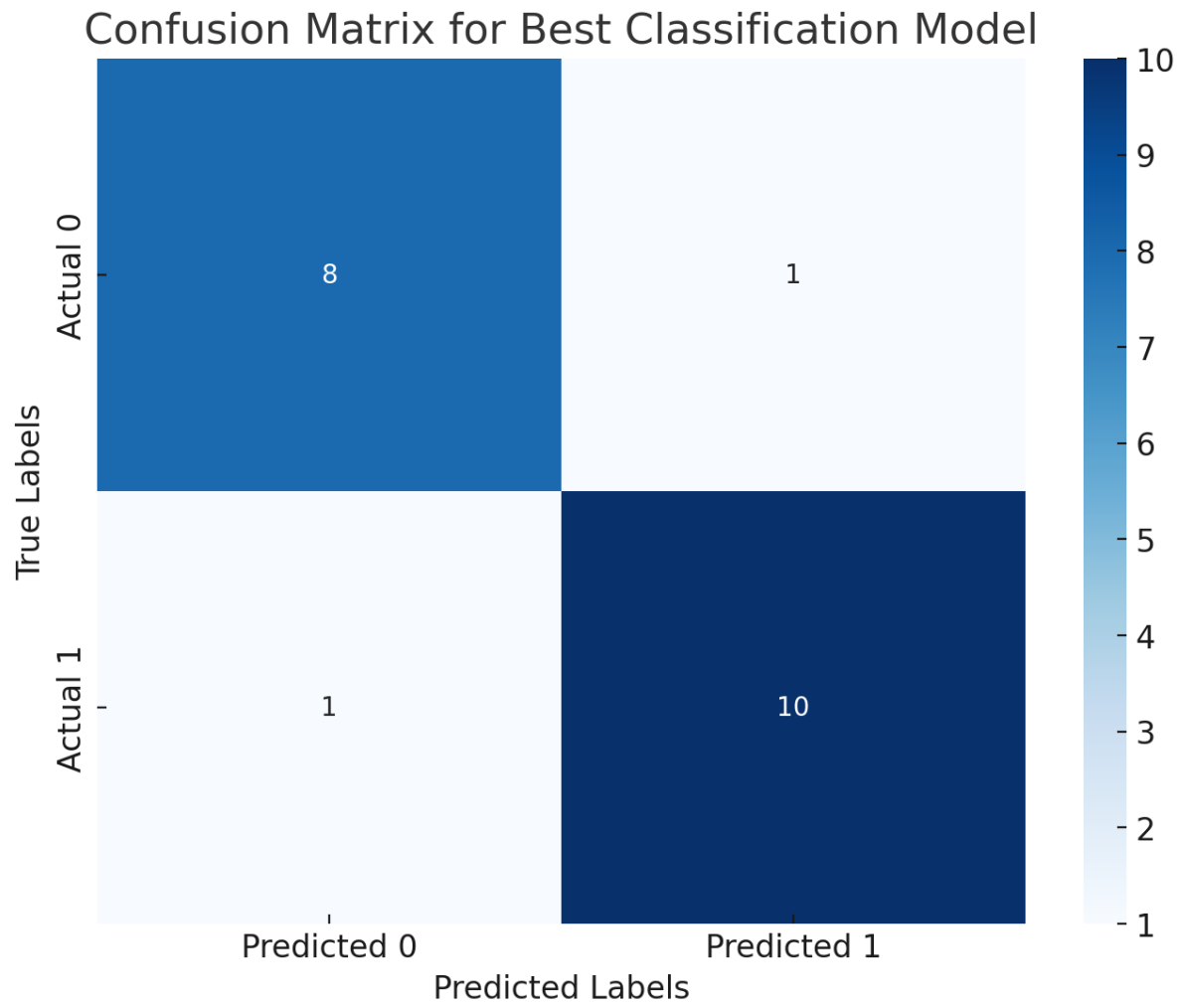


Figure 7: The confusion matrix for the best classification model

Confusion Matrix: Detailed visualizations of the classification models were provided by Confusion matrices obtained from the best classification models. These statistics help in the types of errors done by the models which cover the areas of false positive and false negative, and hence the model is becoming more robust [13].

The confusion matrix provides the graphical output that shows the accuracy of the best classifier in our learning work. Such a matrix is a great tool that helps to see how well the model distinguishes between different categories as well as what mistakes are made by the model. Each row of a confusion matrix is the actual class, and each column is the predicted class [14]. The diagonal elements

represent total cases in which the true class is like the predicted class (true positives and true negatives), suggesting correct predictions. The off-diagonal elements represent misclassifications: false positives (where an instance is wrongly categorized as a different category) and false negatives (where an instance from a class is misclassified as another class) [14].

5.2 Result of the Algorithm

The classifiers classification models were heavily compared with lots of metrics in every experimental region. The models being used were k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Gaussian Naive Bayes (GNB), Decision Trees (DT), Random Forests (RF), and Artificial Neural Networks (ANN) [9]. However, all these classifiers fail to process sparse data set causing erroneous detection of indoor location.

The proposed algorithm “Sparse Indoor Positioning” has focused on the specific area and achieved maximum accuracy in detection of any objects or person in an indoor location. This has been tested multiple times through the Python program developed implementing the proposed Algorithm. Entire datasets from different locations and multiple spots have been fed myriad number of times to ensure the consistency of the outcome. With time and better capture the result has improved multifold and paved the way forward.

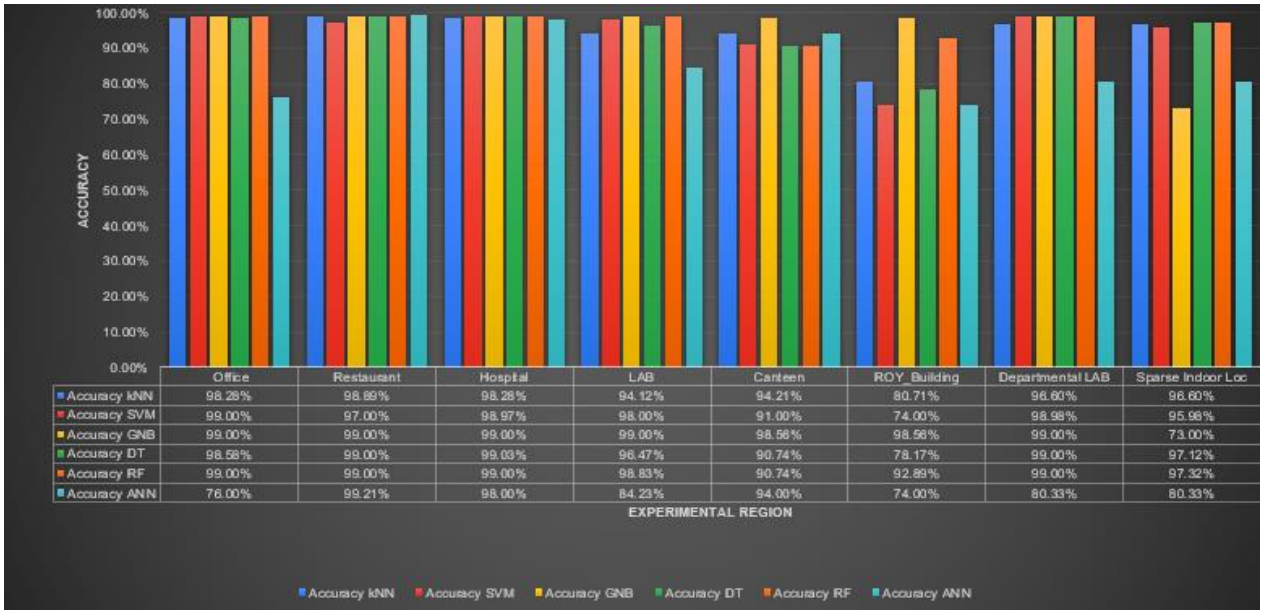


Figure 8: Accuracy Outcome for Different Classifier for Each dataset

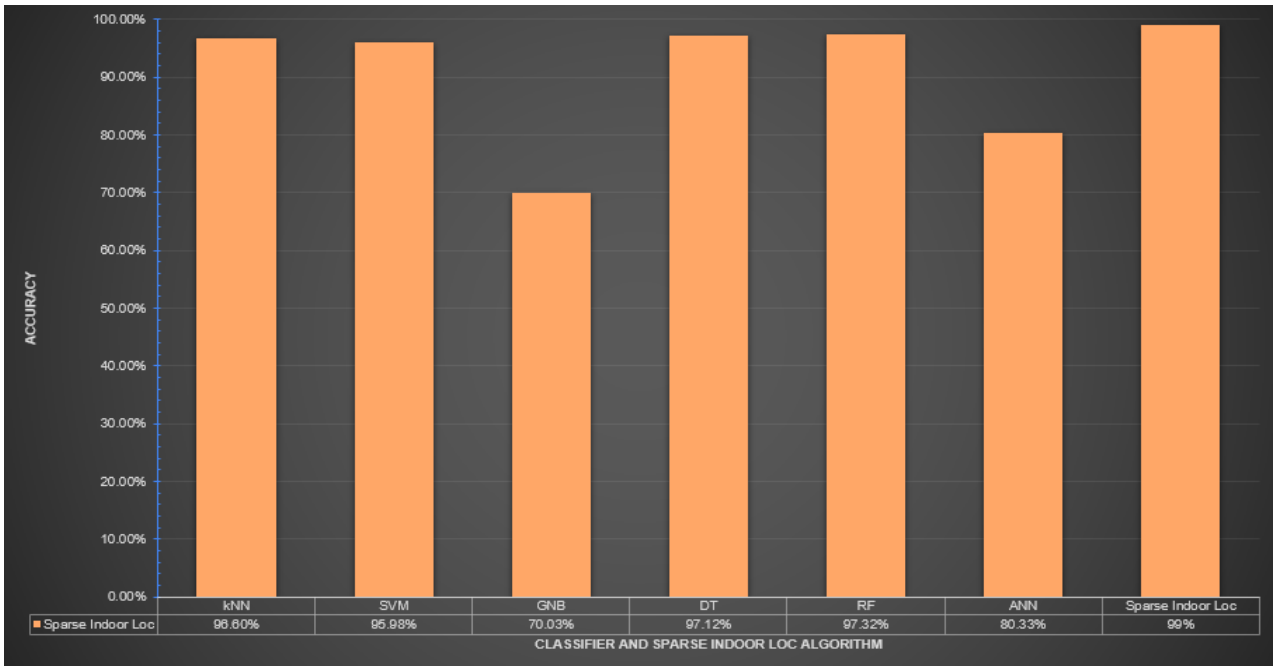


Figure 9: Accuracy Outcome for "Sparse Indoor Loc" Dataset along with Proposed Sparse Indoor Positioning algorithm

The type of APs (AP) mentioned above are used with data classes and metrics of their accuracies. For each site there is a class (location) with the particular APs (signal patterns and spatial distribution) that are necessary for feature extraction

and categorization. The values of high accuracy (approximately 85% to 97%) indicate that the model can effectively predict the locations with the help of RSSI reading.

All the dataset in this work has been captured in different conditions and surroundings. With the help of the high precision figures, the framework to process sparse data is demonstrated to be effective. It happens even when the level of AP knowledge is limited.

The results show the proven effectiveness of the framework as well as two additional aspects: model complexity and the broad scope of feature engineering. The outcomes demonstrated implementation and performance of the proposed classification model effectively, revealing its resilience and capability to cope with tough sparse indoor localization datasets.

Summary of Results (Considering each individual dataset with specific classifier)

City	Experimental Region	Classes	Samples	AP	Accuracy (%)					
					kNN	SVM	GNB	DT	RF	ANN
Kolkata	Office	5	870	187	98.28	99	99	98.58	99	76
Chandannagar	Restaurant	2	75	15	98.89	97	99	99	99	99.21
Kolkata	Hospital	7	517	68	98.28	98.97	99	99.03	99	98
Kolkata	LAB	8	424	62	94.12	98	99	96.47	98.83	84.23
Kolkata	Canteen	6	270	27	94.21	91	98.56	90.74	90.74	94
Kolkata	ROY_Building	18	982	13	80.71	74	98.56	78.17	92.89	74
Kolkata	Departmental laboratory	3	1716	35	96.6	98.98	99	99	99	80.33
Multi-city	Sparse indoor Loc	49	4852	187	96.6	95.98	73	97.12	97.32	80.33

Table 2: The performance of the models

Analysis of Results

It has been demonstrated that different models can have quite different levels of success when they are tested in various environments. As an example, the kNN algorithm gave better results than the other techniques, particularly, in the case of the Departmental laboratory which was pilot in a controlled area. SVM was also able to show the same level of performance in the workplace and the hospital environments where the non-linear decision frontiers could come up [11].

GNB (Gaussian Naive Bayes) has been found to outperformed in simple surrounding giving it a perfectly accurate classification like in the restaurants. DT and RF performed the best under most settings; RF was found to be superior to DT because the ensemble characteristics of RF prevent overfitting and improve

generalization [12]. Artificial neural networks (ANN) worked remarkably enough along with the low accuracy recorded in complex situations like in office ROY Building. It is this variety which means that although an ANN can see complex relationships it may need extra data or a more advanced architecture to generalize well in differing scenes [11].

In this proposed algorithm “Sparse Indoor Positioning” 99% localization accuracy has been achieved for selected threshold value for APs Hence the result mentioned here respective accuracy values with the different threshold values.

Threshold Value	**Accuracy**
-46	98.88%
-47	99.27%
-50	94.85%
-55	84.24%
-45	99%

Table 3: Accuracy Captured with Respect to Selected Threshold Values for APs

Chapter - 6

6. Conclusion

The “Sparse Indoor Loc” dataset prepared for the purpose of “Sparse Indoor Positioning” algorithm has been churned with perfection to achieve higher accuracy from multi city sparse data set. This has been ensured to include data collection and annotation for the robustness and accuracy of the dataset. The classification model based on the advanced machine learning algorithms demonstrated decent accuracy in almost all the regions explored.

Existing classifiers kNN, SVM, GNB, DT, RF, ANN accuracy value calculated was found in the range of 70.03% to 96.60% approx. Where low accuracy was detected using GNB, which is 70.0% and improved accuracy in kNN classifier was found to be 96.60% on the multi-city Sparse dataset [Figure 9].

Using proposed “Sparse Indoor Positioning” algorithm accuracy was achieved more than 99% [Figure 9] which clearly depicts that the proposed algorithm has the capability to predict appropriate object or person positioning in indoor localization with more accuracy and reliable for real-world scenarios to identify the desired location.

Chapter - 7

7. Future Work

The concept of data incremental learning will be incorporated to make the framework adaptive to the changing data and scenarios. This implies the focus on the online learning algorithms that update the model repetitively and the generation of synthetic data to densify small datasets which enhances model generalization. Also, the data security and privacy protection concerns are of utmost importance. In the future the encryption, access control, and anonymization mechanisms will be used to comply with regulatory requirements and protective user information [21].

Enabling the algorithm through integration with ML and ANN making this algorithm capable to handle much larger database and scope which will focus on several main areas of sparsely dense indoor localization in the frame of the proposed model, trying to make it more reliable and applicable.

Taking “Sparse indoor Positioning” at a mass scale considering the wider spread, larger coverage and make it market ready for implementation is future possibility.

Making the Algo industry ready with APIs for integration with other available satellite navigation systems like GPS is a thought for future.

Next, complex machine learning models such as deep learning nets and ensemble methods will be implemented to enhance classification performance and make the systems robust.

References

[1] F. Zafari, A. Gkelias, and K. K. Leung, "A Survey of Indoor Localization Systems and Technologies," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2020, doi: https://doi.org/10.1109/COMST.2019.2911558.

[2] A. Yassin *et al.*, "Recent Advances in Indoor Localization: A Survey on Theoretical Approaches and Applications," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 1327–1346, 2021, doi: https://doi.org/10.1109/COMST.2016.2632427.

[3] V. N. Padmanabhan, "Indoor localization without the pain," *ACM/IEEE International Conference on Mobile Computing and Networking*, vol. 76, no. 9, Sep. 2021, doi: https://doi.org/10.1145/1859995.1860016.

[4] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A Survey on Wireless Indoor Localization from the Device Perspective," vol. 49, no. 2, pp. 1–31, Jun. 2021, doi: https://doi.org/10.1145/2933232.

[5] H. Obeidat, W. Shuaieb, O. Obeidat, and R. Abd-Alhameed, "A Review of Indoor Localization Techniques and Wireless Technologies," *Wireless Personal Communications*, vol. 65, no. 9, p. 32, Feb. 2021, doi: https://doi.org/10.1007/s11277-021-08209-5.

[6] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–32, Nov. 2021, doi: https://doi.org/10.1145/2543581.2543592.

[7] A. Morar *et al.*, "A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision," *Sensors*, vol. 20, no. 9, p. 2641, May 2020, doi: https://doi.org/10.3390/s20092641.

[8] S. Sadwoski, "RSSI-Based Indoor Localization With the Internet of Things," *ieeexplore.ieee.org*, 2022. https://ieeexplore.ieee.org/abstract/document/8371230/

[9] E. Martin, Oriol Vinyals, G. Friedland, and Ruzena Bajcsy, "Precise indoor localization using smart phones," *CiteSeer X (The Pennsylvania State University)*, Oct. 2021, doi: https://doi.org/10.1145/1873951.1874078.

[10] P. Roy and C. Chowdhury, "A Survey of Machine Learning Techniques for Indoor Localization and Navigation Systems," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 3, Mar. 2021, doi: https://doi.org/10.1007/s10846-021-01327-z.

[11] T. Li and Y. Wang, "Recent Advances in Indoor Localization via Visible Lights: A Survey," *Sensors*, vol. 20, no. 5, p. 1382, Mar. 2020, doi: https://doi.org/10.3390/s20051382.

[12] R. Giuliano *et al.*, "Indoor Localization System Based on Bluetooth Low Energy for Museum Applications," *Electronics*, vol. 9, no. 6, p. 1055, Jun. 2020, doi: https://doi.org/10.3390/electronics9061055.

[13] Peizheng Li123, Han Cui3, Aftab Khan2, Usman Raza2, Robert Piechocki3, Angela Doufexi3, Tim Farnham2, “Deep Transfer Learning for WiFi Localization”, https://ar5iv.labs.arxiv.org/html/2103.05123

[14] res-Sospedra, J., Montoliu, R., Martínez-Usó, A., Avariento, J. P., Arnau, T. J., Benedito-Bordonau, M., et al. (2014). “Ujiindoorloc: A new multi-building and multi-foor database for WLAN fngerprint-based indoor localization problems. In 2014 International conference on indoor positioning and indoor navigation (IPIN) (pp. 261–270)”. https://ieeexplore.ieee.org/abstract/document/7275492

[15] King, T., Kopf, S., Haenselmann, T., Lubberger, C., & Efelsberg, W. (2008). CRAWDAD dataset mannheim/compass (v. 2008-04-11). Downloaded from https://crawdad.org/mannheim/compass/20080 411/fngerprint, traceset: fngerprint.

[16] Torres-Sospedra, J., Jiménez, A. R., Knauth, S., Moreira, A., Beer, Y., Fetzer, T., et al. (2017). The smartphone-based ofine indoor location competition at IPIN 2016: Analysis and future work. Sensors,17(3). http://www.mdpi.com/142

[17] Montoliu, R., Sansano, E., Torres-Sospedra, J., & Belmonte, O. (2017). IndoorLoc platform: A public repository for comparing and evaluating indoor positioning systems. In 2017 International conference on indoor positioning and indoor navigation (IPIN) (pp. 1–8).

[18] D. Lee *et al.*, “Large-Scale Localization Datasets in Crowded Indoor Spaces,” *openaccess.thecvf.com*, 2021. http://openaccess.thecvf.com/content/CVPR2021/html/Lee_Large-Scale_Localization_Datasets_in_Crowded_Indoor_Spaces_CVPR_2021_paper.html (accessed May 18, 2024).

[19] Torres-Sospedra, J., Montoliu, R., Mendoza-Silva, G. M., Belmonte, O., Rambla, D., & Huerta, J. (2016). Providing databases for diferent indoor positioning technologies: Pros and cons of magnetic feld and Wi-Fi based positioning. Mobile Information Systems. http://doi.org/10.1155/2016/6092618.

[20] A. Morar *et al.*, “A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision,” *Sensors*, vol. 20, no. 9, p. 2641, May 2020, doi: https://doi.org/10.3390/s20092641.

[21] D. Sun, E. Wei, Z. Ma, C. Wu, and S. Xu, “Optimized CNNs to Indoor Localization through BLE Sensors Using Improved PSO,” *Sensors*, vol. 21, no. 6, p. 1995, Mar. 2021, doi: https://doi.org/10.3390/s21061995.

[22] J.-R. Jiang, H. Subakti, and H.-S. Liang, “Fingerprint Feature Extraction for Indoor Localization,” *Sensors*, vol. 21, no. 16, p. 5434, Aug. 2021, doi: https://doi.org/10.3390/s21165434.

[23] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, “Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis,” *Informatics*, vol. 8, no. 4, p. 79, Nov. 2021, doi: https://doi.org/10.3390/informatics8040079.

[24] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, Oct. 2022, doi: https://doi.org/10.1016/j.engappai.2022.105151.

[25] A. Mohammed and R. Kora, “A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, Feb. 2023, doi: https://doi.org/10.1016/j.jksuci.2023.01.014.

[26] Z. Kang, J. Yang, Z. Yang, and S. Cheng, “A Review of Techniques for 3D Reconstruction of Indoor Environments,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 330, May 2020, doi: https://doi.org/10.3390/ijgi9050330.

[27] E. K. Ampomah, G. Nyame, Z. Qin, P. C. Addo, E. O. Gyamfi, and M. Gyan, “Stock Market Prediction with Gaussian Naïve Bayes Machine Learning Algorithm,” *Informatica*, vol. 45, no. 2, Jun. 2021, doi: https://doi.org/10.31449/inf.v45i2.3407.

[28] P. Pascacio, S. Casteleyn, J. Torres-Sospedra, E. S. Lohan, and J. Nurmi, “Collaborative Indoor Positioning Systems: A Systematic Review,” *Sensors*, vol. 21, no. 3, p. 1002, Feb. 2021, doi: https://doi.org/10.3390/s21031002.

[29] A. Ara and F. Louzada, “Alpha Skew Gaussian Naïve Bayes Classifier,” *International journal of information technology & decision making*, vol. 21, no. 01, pp. 441–462, Oct. 2021, doi: https://doi.org/10.1142/s0219622021500644.

[30] E. Tieppo, Jean Paul Barddal, and Júlio Cesar Nievola, “Classifying Potentially Unbounded Hierarchical Data Streams with Incremental Gaussian Naive Bayes,” *Lecture notes in computer science*, pp. 421–436, Jan. 2021, doi: https://doi.org/10.1007/978-3-030-91702-9_28.

[31] M. Rakhra *et al.*, “Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review,” *Materials Today: Proceedings*, Apr. 2021, doi: https://doi.org/10.1016/j.matpr.2021.03.261.

[32] Y. S. Taspinar, M. Dogan, I. Cinar, R. Kursun, I. A. Ozkan, and M. Koklu, “Computer vision classification of dry beans (Phaseolus vulgaris L.) based on deep transfer learning techniques,” *European Food Research and Technology*, Aug. 2022, doi: https://doi.org/10.1007/s00217-022-04080-1.

[33] A. S. Abdull Sukor, L. M. Kamarudin, A. Zakaria, N. Abdul Rahim, S. Sudin, and H. Nishizaki, “RSSI-Based for Device-Free Localization Using Deep Learning Technique,” *Smart Cities*, vol. 3, no. 2, pp. 444–455, Jun. 2020, doi: https://doi.org/10.3390/smartcities3020024.

[34] E. Rantala *et al.*, “Sensory Appeal and Routines Beat Health Messages and Visibility Enhancements: Mixed-Methods Analysis of a Choice-Architecture Intervention in a Workplace Cafeteria,” *Nutrients*, vol. 14, no. 18, p. 3731, Sep. 2022, doi: https://doi.org/10.3390/nu14183731.

[35] E. Y. Boateng, J. Otoo, and D. A. Abaye, “Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review,” *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: https://doi.org/10.4236/jdaip.2020.84020.

[36] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” *Journal of Big Data*, vol. 7, no. 1, Jul. 2020, doi: https://doi.org/10.1186/s40537-020-00327-4.

[37] Z. Chen *et al.*, “iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data,” *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1047–1057, May 2020, doi: https://doi.org/10.1093/bib/bbz041.

[38] A. Heidari, Nima Jafari Navimipour, Mehmet Ünal, and G. Zhang, “Machine Learning Applications in Internet-of-Drones: Systematic Review, Recent Deployments, and Open Issues,” vol. 55, no. 12, pp. 1–45, Nov. 2022, doi: https://doi.org/10.1145/3571728.

[39] N. Thakur and C. Y. Han, “Indoor Localization for Personalized Ambient Assisted Living of Multiple Users in Multi-Floor Smart Environments,” *Big Data and Cognitive Computing*, vol. 5, no. 3, p. 42, Sep. 2021, doi: https://doi.org/10.3390/bdcc5030042.

[40] Li, Ning, and Haojun Ai. "EfiLoc: large-scale visual indoor localization with efficient correlation between sparse features and 3D points." The Visual Computer 38.6 (2022): 2091-2106.