Wifi Based Indoor Localization Using Grossly Labeled Data For Smartphone Users

A thesis
submitted in partial fulfilment of the requirement for the Degree of Master of Technology in Computer Technology of
Jadavpur University

by
Sanchita Das
Exam Roll No - M6TCT22013
Registration No - 149846 of 2019-2020
Class Roll No - 001910504012
Session: 2019-2022

Under the Guidance and Supervision of
Dr. Chandreyee Chowdhury
Department of Computer Science and Engineering
Jadavpur University Kolkata-700032
India
2022

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Declaration Of Authorship

I, hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of Master in Computer Technology. All information in this document have 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 that are not original to this work.

Sign:________________________**Date:**______________________________

Name: Sanchita Das

Examination Roll No. M6TCT22013

Registration No. 149846 of 2019-2020

Thesis Title: WiFi Based Indoor Localization Using Grossly Labelled Data For Smartphone Users

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that the dissertation entitled “WiFi Based Indoor Localization Using Grossly Labelled Data For Smartphone Users" has been carried out by Sanchita Das (University Registration No: 149846 of 2019-2020, Examination Roll No: M6TCT22013) under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of Master in Computer Technology. The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other University or Institute.

Dr. Chandreyee Chowdhury
Dept. Of Computer Science Engineering
Jadavpur University, Kolkata-700032

Signature: ______________________________

Prof. Anupam Sinha
Head Of Department
Dept. Of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Signature: ______________________________

Prof. Chandan Mazumdar
Dean
Faculty of Engineering and Technology
Jadavpur University, Kolkata-700032

Signature: ______________________________

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Certificate Of Approval

This is to certify that the thesis "WiFi Based Indoor Localization Using Grossly Labelled Data For Smartphone Users " is a bonafide record of work carried out by Sanchita Das**(University Registration No: 149846 of 2019-2020, Examination Roll No: M6TCT22013)** in partial fulfilment of the requirements for the award of the degree of Master in Computer Technology in the Department of Computer Science and Engineering, Jadavpur University during the period of June 2019 to july 2022. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, the opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Examiner:

Signature: ____________________ Date: ________________________

Dr. Chandreyee Chowdhury
Dept. Of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Signature:______________________ Date: ______________________

Acknowledgement

I would like to express my sincere gratitude to my advisor **Dr. Chandreyee Chowdhury** for her continuous motivation, guidance and patience throughout my thesis work, I have been very lucky to have a advisor who cared so much about my work. More importantly, the scientific and personnel support along with her valuable suggestion softened the journey.

Special thanks to dear senior Manjarini Mallik for all her help in getting the things done.

Sanchita Das

Signature: ________________________

Table of Contents

Abstract

Localization in indoor areas deals with the serious problem of collecting data over a broad experimental region while maintaining the location points. Sensor values may be collected during normal movements and work schedules of the people, but those data will be either unlabelled or grossly labeled. This paper addresses the challenge of providing a localization solution for unlabelled or grossly labeled indoor data by a two-phased semi-supervised learning approach. In the first phase, a Rank-Based Iterative Clustering method (RICM) method is proposed that processes the entire dataset iteratively, generating a final cluster at the end of each iteration.

The experiments were conducted in a realistic indoor localization dataset. In the first phase, distinct temporary sets of clusters are produced by each clustering algorithm and their performances are evaluated by computing different clustering scores based on the respective temporary set of clusters obtained. Finally, the algorithms are sorted according to their rank. At each iteration, an inner join is performed among each possible pair of clusters obtained from those rank-wise sorted algorithms. Finally, an improved set of clusters is received and the cluster containing maximum data samples is kept in the final cluster set. These samples are removed from the primary dataset and in subsequent iterations, the remaining data samples are re-clustered. This way, a new final cluster is obtained at each iteration and the procedure is repeated until all the final clusters are obtained.

In the second phase, classification is performed using random test data and the obtained set of clusters as training data and 97% accuracy is obtained for different supervised classification algorithms. External and internal validation scores were utilized to evaluate the clustering techniques.

Chapter 1

Introduction

In several domains, technological advancements in recent decades have resulted in a massive rise in data-intensive applications. Localization and navigation in indoor environments are gaining importance because of various location-based services, e.g. user tracking [3], object detection [14], robot navigation [21], etc. However, collecting and maintaining data location point-wise over a broad experimental region is a difficult and time-consuming task. The biggest effort nowadays is no longer in the production of data, but in the analysis of that data by extracting usable information from it. Sensor values can be collected over the experimental region during a normal day schedule, with users doing their regular work, without bothering about the location points. This results in an unlabelled or grossly labelled (maybe, room-wise) dataset. The challenge is, to divide these kinds of datasets into sub-regions or simply, label the data samples into smaller location areas.

1.1. Indoor Localization system

Localization is the term which is used to find the exact or approximate location of an object or user. Though GPS is efficient for localization in outdoor environments, it cannot penetrate the complex building walls and other obstacles. Because there is no visual contact with the GPS satellites in inside settings, an ILS (indoor localization system) must rely on other techniques of locating. Hence, for indoor areas, a distinct way is followed based on other sensors that are available in indoor area, such as WiFi, Bluetooth Low Energy (BLE) etc. and that is where the term indoor localization came into existence. There are various ways for indoor localization. Traditionally, researchers used to find the location based on the time or angle of the arrival of the signal, or by proximity measurements. A newer approach is fingerprinting, which depends on the signal strengths received at different locations. The impact of these techniques in recent indoor localization works are explained in the subsection 2.1.

Location can be obtained based on different ML techniques. Supervised methods are applicable where the dataset is labelled location point wise. Otherwise, unsupervised approaches can be the solution. Here we have applied clustering approach based on similarity or dissimilarity of the received signal strength values of floor. In this work we are not diving the region geographically but based on characteristics of the radio signal, we are dividing region by using the clustering methods and matching with the building or property wise region. At the end, the user will be located in any one of the computed regions based on similarity or dissimilarity of the signal values received on the users' smartphone. To collect the data for locating user, first if we divide the complete region into grids and each grid is assigned to a unique label, which indicates that specific position. For example, in the dataset proposed in [25], the grid with label L4-14-36 indicates a location point physically positioned at 4 th floor, 14 th row and 36 th column number. In the active public place like metro station or shopping mall dividing the place into some equal size grids is way more difficult to collect data. So, collecting finger prints at every grid in a public place can be a laborious task and often invisible. So that's why in the public place to locate user we may collect data using static reference points, for example pillars, doors, start or end of staircases etc. The size of each grid indicates the granularity of the localization, either it is fine-grained or coarse-grained.

1.2. Applications of Indoor localization systems

There are mainly three types of applications for ILS. Below is a discussion of a few of them:

- Object navigation - The process of locating one or more items in an image and tracing a bounding box around their extent is known as object localization. These two tasks are combined in object detection, which locates and categorises one or more things in an image. A computer vision technology called object detection helps locate and identify things in an image or video. To be more precise, object detection creates bounding boxes around the items it has found, allowing us to determine their location inside (or how they move across) a scene. There is a method to reduce, while retaining the robustness of the steered response power-based phase transition algorithm for object detection[].This method employs three main techniques: receiver signal strength indicator (RSSI), time of arrival (TOA), and time difference of arrival (TDOA)[15].
- Robot navigation - The ability of a robot to determine its own position in its frame of reference and then plan a path to some goal location is referred to as navigation. Many

localization algorithms use the robot's previous position to update its current location during navigation. This is accomplished by continuously monitoring the robot's path. Odometry and sensor data are used in position tracking. People with motor or cognitive impairments are frequently afraid of moving in large, crowded spaces (e.g., because they could lose the sense of direction). To address this issue, the position tracking approach was used, and the position estimation and tracking technique developed within the project devices for assisted living (DALi) was implemented[9].

- User navigation - The map, directions, and route that a user takes to find information on a website or app are referred to as user navigation. User navigation detect where is the user in that particular place or building. User navigation technology is used to track user movement in an indoor environment. Using user localization technology, a model is introduced to identify a user's position, providing a smart solution that does not require a dedicated tracking device.A hybrid random forest model that takes signal strength from Wi-Fi devices has been proposed to achieve localization awareness in devices [16].

1.3. Overview of Machine Learning

Machine learning covers a wide range of topics, including regression analysis, feature selection methods, classification and clustering. The latter entails classifying the objects in a dataset. For classification, three approaches are available: supervised, semi-supervised, and unsupervised classification. In the first case, the classes, or labels, of some objects are known ahead of time, defining the training set, and an algorithm is used to generate the classification criteria. Semi-supervised classification involves training an algorithm with both labelled and unlabeled data. They are commonly used when manually labelling a dataset becomes prohibitively expensive. Finally, unsupervised classification, also known as clustering, is concerned with defining classes from data without prior knowledge of the class labels. Clustering algorithms are designed to identify groups of objects, or clusters, that are more similar to one another than to other clusters. This method of data analysis is closely related to the task of developing a data model, which is defined as defining a simplified set of properties that can provide intuitive explanations about relevant aspects of a dataset. Clustering methods are generally more difficult to implement than supervised approaches, but they provide more information about complex data. This type of classifier is the primary focus of the current work. Clustering algorithms' performance can vary significantly for different applications and data types because they involve several parameters, frequently operate in high-dimensional spaces, and must deal with noisy, incomplete, and sampled data. For these reasons, several approaches to clustering have been proposed in the literature. In practise, selecting a suitable clustering approach given a dataset or problem becomes a difficult task. Nonetheless, comparing different clustering methods can teach us a lot. Several previous comparison efforts for clustering algorithms have been reported in the literature.

Each clustering algorithm is based on a set of parameters that must be adjusted in order to achieve viable performance, which corresponds to an important point to consider when comparing clustering algorithms. A long-standing issue in machine learning is defining a proper procedure for setting parameter values. In principle, one can use an optimization procedure to find the parameter configuration that provides the best performance for a given algorithm. Nonetheless, there are two major issues with this approach. First, adjusting parameters to a given dataset may result in overfitting. That is, the specific values discovered to provide good performance may result in lower performance when new data is considered. Second, due to the time complexity of many algorithms and their typically large number of parameters, parameter optimization may be impossible in some cases. Many researchers eventually resort to using classifier or clustering algorithms with the software's default parameters. As a result, efforts must be made to evaluate and compare the performance of clustering algorithms in optimization and default scenarios. Following are some representative examples of algorithms used in the literature.

Clustering algorithms have been built into a variety of programming languages and software packages. It is common for changes or optimizations to be implemented during the development and implementation of such codes, resulting in new versions of the original methods. The current work compares several clustering algorithms found in popular packages available and this decision was popular in the data mining field, as well as the well-established clustering packages it contains. This study is intended to help researchers who have experience in data clustering.

1.4. Motivation

The existing ML-based indoor localization approaches provide a solution either based on supervised learning [22][23][24], or some specific unsupervised methodologies [9][10][12]. This sets a limitation that the dataset must be labeled (for supervised learning), or at least the experimenters have some idea about the regional divisions of the complete experimental region, in case of unsupervised learning. Also, the excellent performance of some specific unsupervised approaches does not guarantee that the proximity of the data points within a cluster will be as small as possible. Among numerous data points, there may be one or two data points, presenting certain location points within a cluster that are physically positioned far away. This fact may arise as a serious issue for real-time localization or tracking methodologies. This motivated us to design a combined clustering approach, in which, the clustering algorithms will be ranked dynamically based on their performances and each cluster will be generated by the combined decisions of all the algorithms. Accordingly, the contributions of the work are summarized in the next subsection. In existing work there is many way to detect the user using distinct clustering methods, but in our proposal the clusters are obtained by matching the location points with the floor.

1.5. Contribution

The contributions of the proposed works are as follows:

- It provides a two-phased semi-supervised localization solution for any unlabeled or grossly labeled indoor dataset.
- The proposed Rank-Based Iterative Clustering Method (RICM) algorithm provides a clustered dataset with negotiable chance of containing two physically apart location points within common cluster. Hence, it provides a real-time applicable solution.
- It provides a classification approach for dynamically generated test data, by using the clustered data as training dataset.
- It validates the proposed approach based on real life benchmark datasets.

1.6. Scope of the work

We have defined how the proposed method is working which is combining best two clustering algorithms which gives us the clustering labels and with that clustering label we are matching the location points of the floor and checking the partition in well done or not. The signal strengths received from APs fluctuates for different areas of the complete experimental region. As a result, we are able to partition the entire experimental zone into smaller areas and using our proposed method. We have matched the cluster's label to the location points of the specific floor of the building.

However, there are certain areas of concerns that can be improved. For example, determining the number of clusters to which the entire region is to be divided can be computed using some rules and experiments and we can go for an optimal decision. We'll try to improve these issues while working for the future works planned, as explained below.

Chapter 2

Related Work

2.1. Indoor localization techniques

Indoor localization can be done in two ways. First is statistical approach of machine learning and second is utilizing geometric property of the signal. The methods based on geometric property are as follows:

2.1.1 Angle Based Method

Arrival Angle (AoA)/Angulation The Angle of Arrival (AoA) technique determines the angle of arrival of a mobile signal arriving at multiple base stations from a known location. The AoA approach requires only two beacons to estimate position in a 2D dimension plane. Three or more beacons are used for location estimation to improve accuracy (triangulation). It requires highly directional antennas or antenna arrays to determine direction. The location of the intersection of two lines of bearing from known reference points can then be estimated using geometric relationships [17].

2.1.2. Time Based Methods

Lateration/Trilateration/Multilateration these are the three terms refer to a position determined by measuring distances. Lateration, also known as trilateration, is the process of determining an object's position by measuring its distance from multiple reference points. As a result, it is also known as range measurement technique. The "tri" in trilateration denotes that at least three fixed points are required to determine a position[8]. Lateration techniques are those that are based on the measurement of the propagation-time system (e.g., TOA, RTOF, and TDOA) as well as RSS-based and received signal phase methods [4][5].

- Time of Arrival (ToA) - Flight Time (ToF)/Time of Arrival (ToA) systems are based on the precise timing of a signal transmitted from a mobile device to several receiving beacons. ToA involves the mobile device sending a time-stamped signal to receiving beacons. When it is received, the distance between the mobile node and the receiving beacons is calculated based on the transmission time delay and signal speed.The ToA method necessitates exact knowledge of the transmission start time [18]. As a result, all receiving beacons and mobile devices are precisely synchronised with a precise time source[6][7]. ToA is the most accurate technique for filtering out multi-path effects in an indoor environment. The requirement for precise time synchronisation of all devices is one of the disadvantages of the TOA approach. An additional server will be required for time delay measurement, increasing the system's cost. Increased delay can also be propagated by a denser environment, which means more people.

- Time Difference of Arrival (TDoA) - Time Difference of Arrival (TDoA) techniques are used to measure the time difference between multiple pairs of known locations and use relative time measurements at each receiving node rather than absolute time measurements[19]. TDoA does not require a synchronised time source of transmission to resolve timestamps and determine location. TDoA receives a transmission with an unknown start time at multiple receiving nodes, with only the receivers requiring time synchronisation. Each difference in arrival time measurement results in a hyperbolic curve in the localization space, where the mobile node is located. The client's possible locations are defined by the intersection of multiple hyperbolic curves[7]. Multilateration is the term used to describe TDOA-based localization.

2.2. Clustering techniques

In the literature, many different types of clustering methods have been proposed but in Indoor localization mainly three clustering algorithm used precisely. This three main representative clustering algorithm is discussed as follows:

2.2.1. k-means algorithm

The k-means algorithm has been widely used by researchers in partitional approaches. This method requires the number of groups (k) as input parameters. Each data point is initially assigned to one of the k clusters based on its distance from the centroids (cluster centres) of each cluster. Regarding clustering approaches, the k-means algorithm is very popular for the researchers. This method select k number of clusters and group the dataset with k number times. Initially data point are assigned Randomly to each of the clusters according to the initial centroids. each data point is assigned to one of the *k* clusters based on the distance to the cluster centroids. An approach is shown [10] which greatly reduces the complexity and increases the accuracy of floor estimate compared to NN fingerprinting. The indoor wireless positioning approach based on WiFi k-means is suggested as a way to lessen the impact of indoor environmental factors on indoor wireless positioning, increase positioning accuracy, and increase the location area [11].The effect of attribute values is taken into account when using the enhanced distance formula, and the difference between various objects can be estimated more precisely. Therefore, the goal of this research is to develop a novel clustering technique that will shorten search times without sacrificing placement accuracy [2].So, these are some methods in which k-means is applied on Indoor Localization system or k-means is used to detect the users in indoor environment.

2.2.2. Hierarchical Clustering algorithm

An algorithm called hierarchical clustering, commonly referred to as hierarchical cluster analysis, divides objects into clusters based on how similar they are. The result is a collection of clusters, each of which differs from the others while having things that are generally similar to one another. The goal of hierarchical clustering is to create a hierarchy of nested clusters. This hierarchy is graphically represented by a diagram known as a dendrogram, which is an inverted tree that illustrates the order in which elements are combined (bottom-up view) or clusters are divided up (top-down view). The primary way that the suggested localization strategy differs from currently used hierarchical clustering-based approaches is through the automatic partitioning of the indoor environment of interest into zones that may or may not overlap . This results in fewer zone classification errors, which can be expensive, especially for traditional approaches, this scenario would undoubtedly return inaccurate localization output. As a result, the precision of indoor location is also improved. In a real hospital, the effectiveness of the hierarchical classification-based technique was confirmed (12).

2.2.3. Gaussian Mixture Clustering

A probabilistic model called a "Gaussian mixing model" posits that all of the data points were produced by combining a limited number of Gaussian distributions with unknowable parameters. Mixture models can be seen as a generalisation of k-means clustering to include details of the covariance structure of the data as well as the locations of the latent Gaussian centres. The expectation-maximization (EM) approach for fitting a mixture of Gaussian models is implemented by the Gaussian Mixture object. Additionally, it can compute the Bayesian Information Criterion to determine how many clusters there are in the data and create confidence ellipsoids for multivariate models. Using Bhattacharyya coefficient/ distance (BC/BD) and a novel GMM-based probabilistic framework, A paper suggests several fingerprinting zones for the real-time observed RSSI vectors. More specifically, values associated with each location are modelled using a multivariate GMM distribution following the measurement of the RSSI data (13).

2.3. Cluster Evaluation matrix

To detect how well the clustering algorithm performed, some commonly used internal validation indices or matrix has been used to validate the clustering models and the obtained results from clustering. There are two types of clustering validity techniques based on external and internal criteria. The external criteria evaluate clustering in relation to a pre-specified structure, whereas the internal criteria evaluate clustering in relation to a proximity matrix of the data objects. The most precisely used internal clustering validity indices are as follows:

- Silhouette Index – A technique for interpreting and validating consistency within data clusters is known as silhouette analysis. The silhouette value gauges an object's cohesion with its own cluster in comparison to other clusters (separation). It can be used to investigate how far apart the generated clusters are from one another. The silhouette plot offers a visual approach to evaluate factors like the number of clusters by displaying a measure of how close each point in one cluster is to points in the neighbouring clusters[1].

- Calanski Harbasz - When the labels for the ground truth are unknown, Calinski-Harabasz can be used to assess the model. In this case, the dataset's inherent quantities and characteristics are used to validate how successfully the clustering was performed. When compared to other clusters, the cohesion of an object is gauged by the CH Index (also known as the Variance ratio criteria) (separation). with higher Silhouette values and lower Davies-Bouldin index values indicating better clustering quality.

In this work some popular external indices are used to evaluate the quality of clusterings in the external criteria category which are as follows:

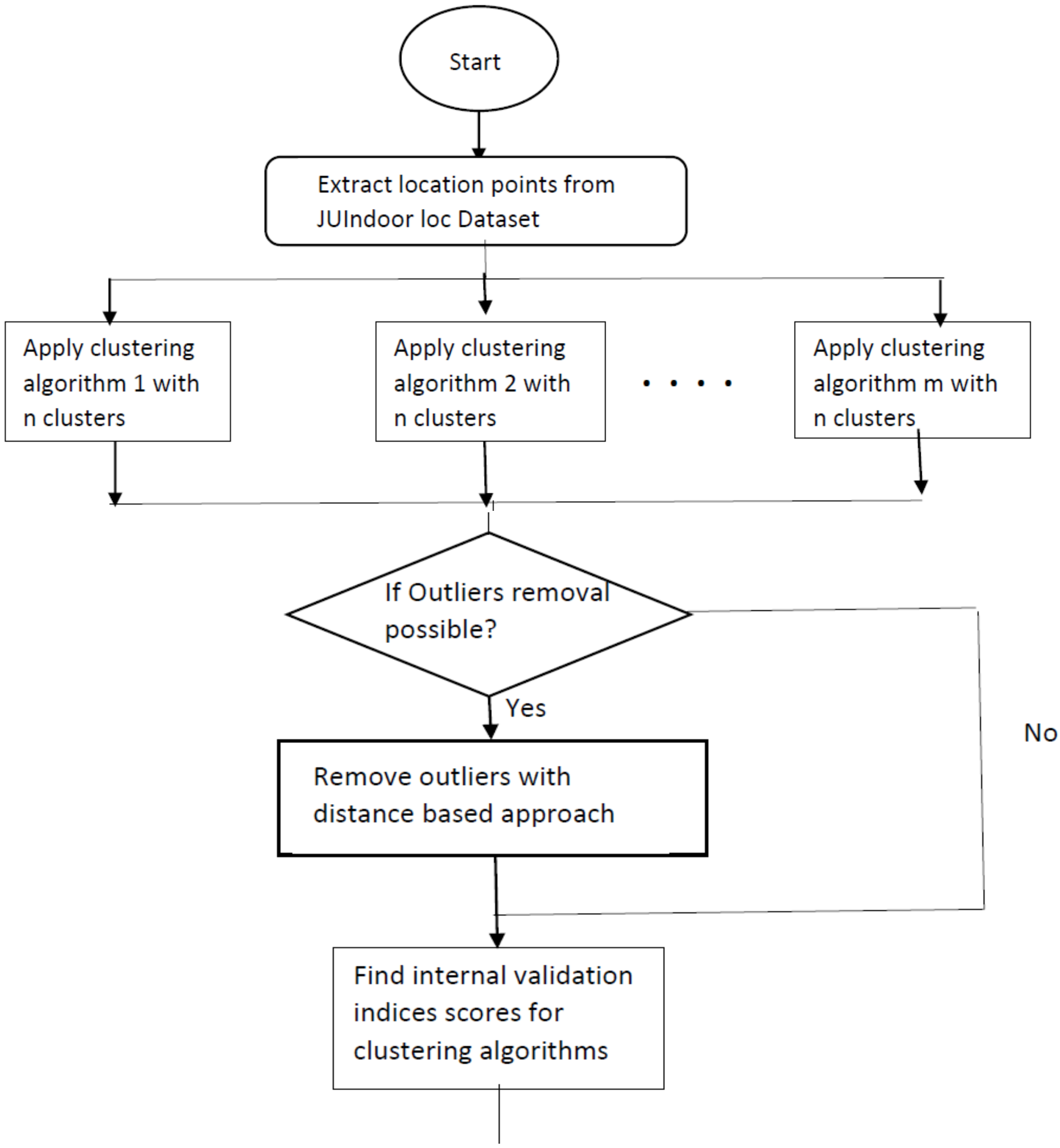
- normalised mutual information (NMI) - When we are given the cluster labels, normalised mutual information (NMI) provides us with the reduction in entropy of class labels. When we know the cluster labels, NMI sort of informs us how much the ambiguity about class labels diminishes. It is comparable to how decision trees gain knowledge. NMI has the benefit of allowing us to compare clustering models with various numbers of clusters because it is normalised.NMI can be calculated using the scikit-learn function normalized_mutual_info_score.
- adjusted Rand index (ARI) - To establish whether two cluster findings are similar to one another, the Adjusted Rand score is presented. The "RI" in the calculation stands for the rand index, which compares two cluster results by taking into account all points found in the same cluster. Where high values of NMI and ARI indicate better clustering results.

Chapter 3

The proposed methodology

3.1. System model

For the system model, The below flowchart has been drawn where it can be clearly seen that n number of clustering algorithm has applied on dataset and compute the best two clustering algorithm and combine them by intersecting their clustering label's corresponding dataset and find the best region and mapped them as a new cluster. Figure 1 shows the system model of the following proposed method.



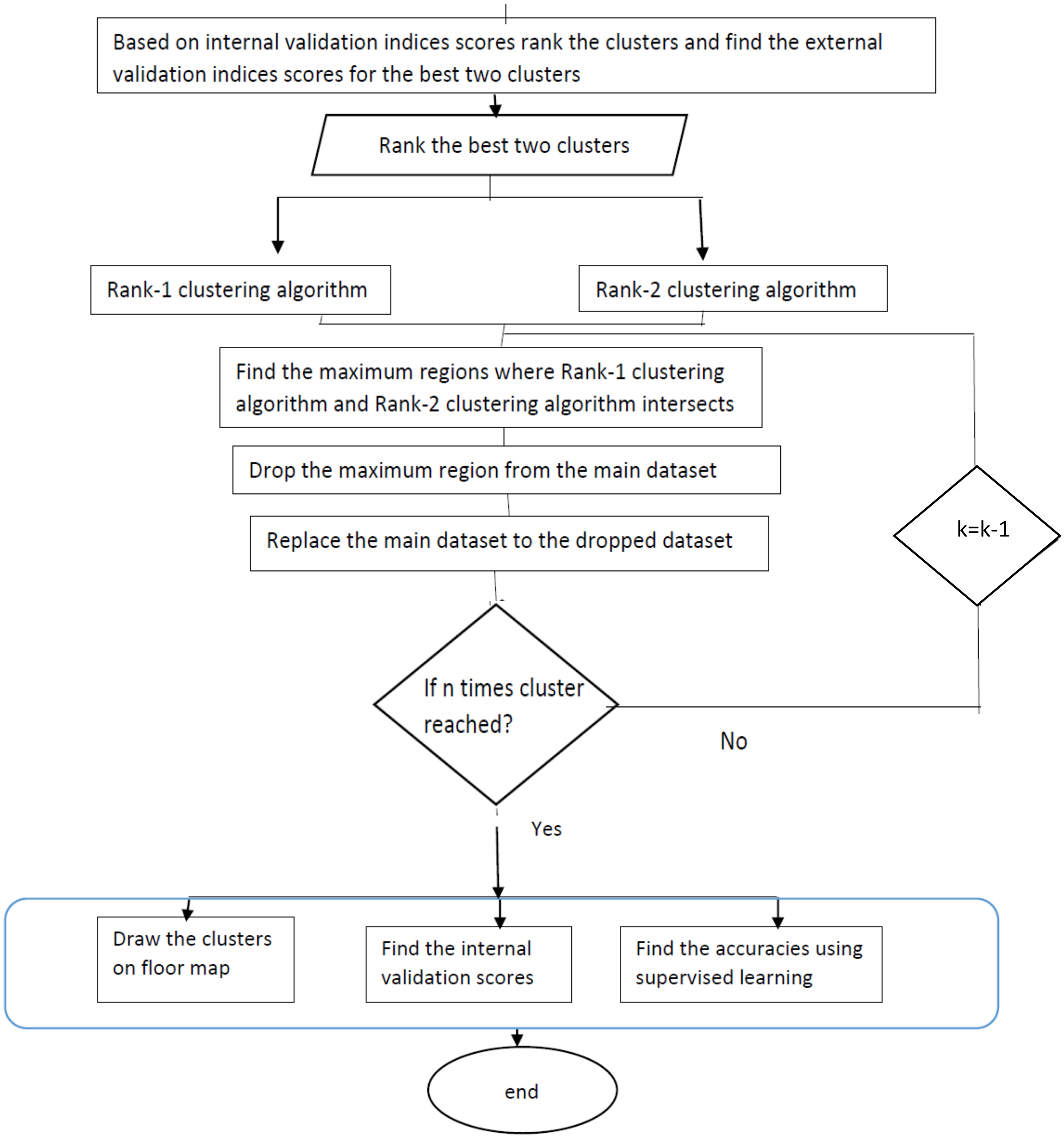


Figure 1 : System model diagram of the proposed method RICM

3.2. Proposed algorithm

A Rank-based iterative clustering algorithm has been designed in this work that comprises different types of clustering algorithm. For our problem, we have used k-means clustering algorithm, Hierarchical and Gaussian Mixture clustering algorithm as algorithm 1 and algorithm 2 and algorithm 3. The proposed methodology is discussed in the following steps:

I. Inputs – For the first step we have taken the JUIndoorLoc dataset and drop the labels (The location points) and we have applied different types of clustering algorithm on that unlabelled dataset with K number of clusters where n number of clustering algorithms are taken in a set C_i and all n number of algorithms are executed parallelly.

II. Outliers Removal – For n number of algorithms in the set C_i ,we have checked if outliers removal approached can be proceed with that particular clustering algorithm or not. Here in our work, outcome of outliers removal on k-means has been successfully executed. So, in the outlier removal we have used a distance based approach where we are finding the data points who are having the maximum distance from the cluster centroid to itself and marking them as an outliers. we are finding the outliers and After detecting outliers we have dropped them from the main dataset for a maximum iteration time which has been fixed after tuning. Here three terms has been used which are Euclidean distance, a threshold value, outlier ratio formula. First, The Euclidean distance formula has been applied to detect the distance from each of the data points to its cluster centroid and calculate the maximum distances for all the data points. The threshold value which is taken as T=0.921 after tuning. Then the outlier ratio formula has been calculated as

If square root of (location point^2 - centroid^2) / max(location points) > Threshold point ==True **then** it can be called as an outlier.

III. Internal validation scores – To detect which clustering algorithm is performing well, some representative internal validation score have been computed. As internal validation indices, Silhouette indices and Calinski harabasz indices score have been applied on the clustering models. and rank them based on these scores. After ranking two best performed clustering model is taken as rank 1 clustering model and rank 2 clustering model.

IV. Intersection of rank 1 and rank 2 clustering algorithm – For K number of times rank 1 algorithm and rank 2 algorithm have been executed parallelly and form k number of clusters each. So, first thing is we have intersected this k numbers of clusters of rank 1 algorithm to k numbers of clusters of rank 2 algorithm and form K*K clusters total. We have chosen that cluster which is covering the maximum region or maximum data points and mapped that cluster as one of the final cluster and delete that cluster from the main dataset and replace the main dataset to the dropped dataset for K number of times. That's how K number of times the dataset will be labelled. How the intersection of each cluster is occurring is shown in the Figure 2.

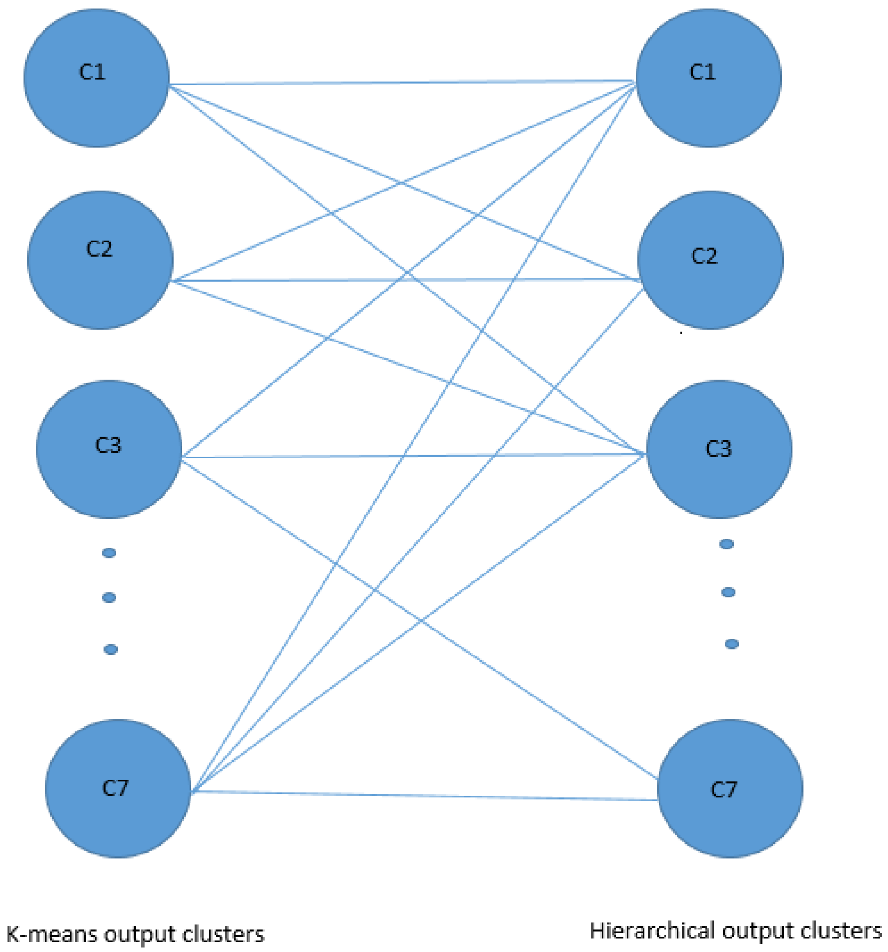


Figure 2: Output clusters of rank 1 algorithm and rank 2 algorithm is intersecting

V. Result checking - After this clustering method we have checked the results in three way. First is matching the clustering label with respect to the location points or the labels of the dataset. Second we have computed the accuracy of the clustering method using four supervised algorithms which are Random forest , SVM, Logistic Regression, KNN , Random forest and decision tree. Third, we applied three internal validation indices to check or validate the clustering model.

3.2.1. The Rank-based Iterative Clustering Method (RICM)

The Rank-based Iterative Clustering Method (RCCM) is an iterative process that uses an agreement strategy between two clustering algorithms to find the major region for each clustering after intersection of the clustering region between two clustering algorithm. The basic working of clustering has been shown in brief with this simple algorithm bellow.

Algorithm : Rank-based clustering algorithm

Data: a set of n data points D= $\{d_1 , d_2 .... d_n\}$

A set of clustering algorithms S_i where I = {1,2......c}

Output: the entire dataset has been clustered by the algorithm.

Initialization:

Run the c algorithms of the set S_i in parallel on the dataset D.

Compute all of the clustering algorithm and Compute the consensus solution from

After applying clustering algorithms, the internal validation scores are computed to see which clustering algorithm is performing well in this JUIndoor dataset and select those clustering models. The structure can be seen in the following Figure 3.

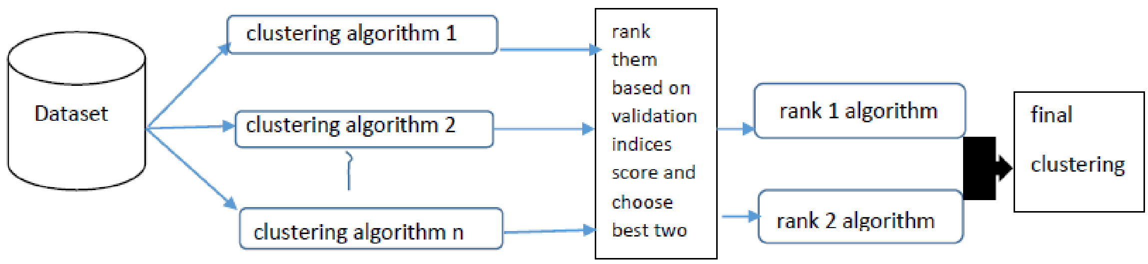


Figure 3. The diagram shows the procedure of the RICM algorithm

The algorithm of RICM is as follows where **Function** outlier_removal is used to remove dataset and **Function** clustering is used to intersect rank 1 algorithm and rank 2 algorithm.

Algorithm 1: Rank based iterative clustering algorithm

---

input : K: no. of cluster
D: dataset containing data points
C_i: A set of clustering algorithms where i = {1,2……n}
output: A set of K cluster

initialization:

1. For each n clustering algorithms of the set C_i:

    Run the algorithm on the dataset D and check
    if outliers removal is possible in the algorithms **then**
        Apply a distance based approach to remove outlier by using outlier removal function outlier_removal().

    else add the label with the dataset without removing outliers.

    Function outlier_removal(D,K):
        for each maximum iteration:
            Apply clustering algorithm on dataset D with K no of cluster
            d_vector = np.array([Euclidean distance of (data point, centroid of cluster) for each data points])
            d_max = d_vector.max()
            Take T as 0.921
            outliers = (distance(centroid, row) / d_max > T, for each cluster points)
            new_cluster = (distance(centroid, row) / d_max <= T, for each cluster points)

    return new_cluster with labels

    Rank all the n algorithms of the set C_i by computing the internal validation scores, Silhouette indices and Calinski harabasz indices score.

2. Perform **Function** clustering(D,K) on rank 1 algorithm and rank 2 algorithm.

3. **foreach** cluster k:

```
    max_df= clustering(D,k)

    Function clustering(D,k):
     foreach i in range of k:
       foreach j in range of k:
         dataset df = D with algorithm 1 cluster label==i intersects with D with  algorithm
         2 cluster label==j
         If max length < length of df :
                 max length  =  length of df
                 max df=df.copy()
      add column cluster in max df with the value of k
    return max df

    dataset M= delete max_df from main dataset D with the help of left outer join
    dataset D= dataset M
```

end

Chapter 4

Experimental setup

We have experimented on the benchmark indoor localization dataset JUIndoorLoc [25]. This dataset is publicly available[1]. The data was collected using an android application which is capable of sensing all available access points (APs) in the corresponding environment and collect RSSI values for specific location points. RSSI was represented by negative integer values as the signal strength and its unit is dBm. Table 1, shows the different signal strength integer values with the respective priorities. The data set is divided into 1750 rows and 119 columns where columns represent the features, i.e., the APs.

Table 1: RSSI signal strength [25]

Unit	dBm
-30 and upper	Excellent signal strength
-30 to -90	Good/Average
-90 and lower	Bad signal strength

The dataset represent the floor of Jadavpur University's Computer science building that included some active learning classroom (ALC) and a hallway, as depicted in Figure 4. The classroom's adjustable desks, tables, and chairs give students a variety of seating options. The classroom has a large capacity and the space is designed to provide consumers complete control. All of the spaces in this floor are anticipated to be used while carrying out various tasks thanks to the design elements. Although the active learning floor is big, It is intended to serve as a test site for data collection. To perform localization, we need to divide the experimental region into sub regions, which will act as location points. Here is total 42x21 grids, so there are total 882 location points where the data was collected from L4-4-9 to L4-42-12. The details of the floor can be shown in the table 2.

[1] https://drive.google.com/open?id=1_z1qhoRIcpineP9AHkfVGCfB2Fd_e-fD

Table 2: Details of dataset[2]

Number of Entries	7550
Number of device used	4
Number of features	119
Data points collection starts from	L4-4-9 to L4-4-21
Data points collection ends	L4-42-10 to L4-42-12

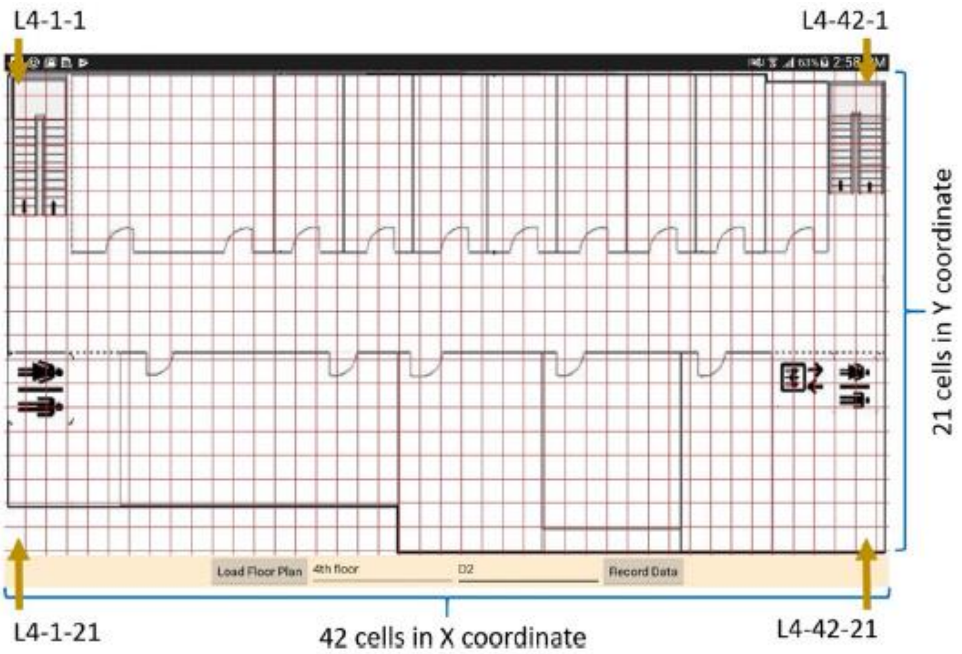


Figure 4: Floormap of JUIndoorLoc Dataset [25]

[2] https://drive.google.com/open?id=1_z1qhoRIcpineP9AHkfVGCfB2Fd_e-fD

Chapter 5

Experimental results

5.1. Experimental results for individual clustering algorithms

In this chapter, the performance of the proposed clustering algorithm is analyzed. The experimental results are detailed as follows. For the RICM three clustering model k-means, Hierarchical, Gaussian Mixture clustering model have been applied on indoor localization dataset.

5.1.1 Results of k-means Clustering

The main problem in k-means is that the outliers the anomalies. They create the issue on the performances of clustering. To detect and delete outliers in k-means clustering algorithm, the approach which have been applied on the dataset is a distance based approach. Here three terms has been used which are Euclidean distance, a threshold value, outlier ratio formula. In the last phase after detecting the outliers for this dataset and run k-means clustering algorithm again until it reach the maximum iterations. The result of removing outliers are shown below where Figure 5 represent the data points in each cluster before removing outliers and Figure 6 , Figure 7.1, Figure 7.2 represents the data points after removing the outliers.

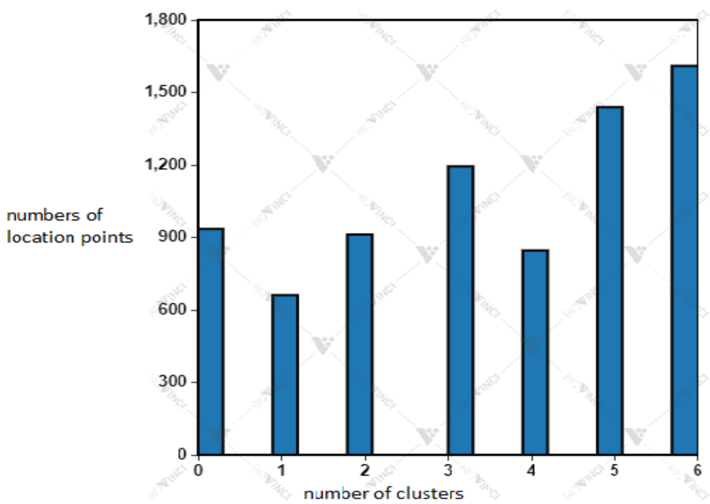


Figure 5 :Before removing outliers the data points are clustered .

Table 1. Result of k-means Clustering before and after removing outliers

Score	Before outlier removal	After outlier removal
Davies bouldin score	0.8325	0.8731
Silhouette score	0.3721	0.3856
Calinski harabasz score	7889.9656	8023.9052

To investigate the effectiveness of this outliers removal approach Davies bouldin, Sil houette, Calinski harabasz has been implemented. How the scores are increased that is show n in the table 1.

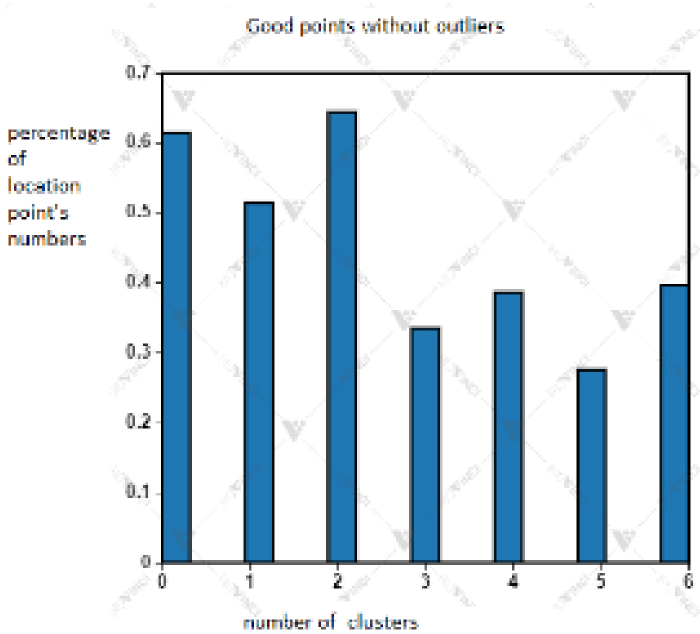


Figure 6 : Data points are clustered after removing outliers.

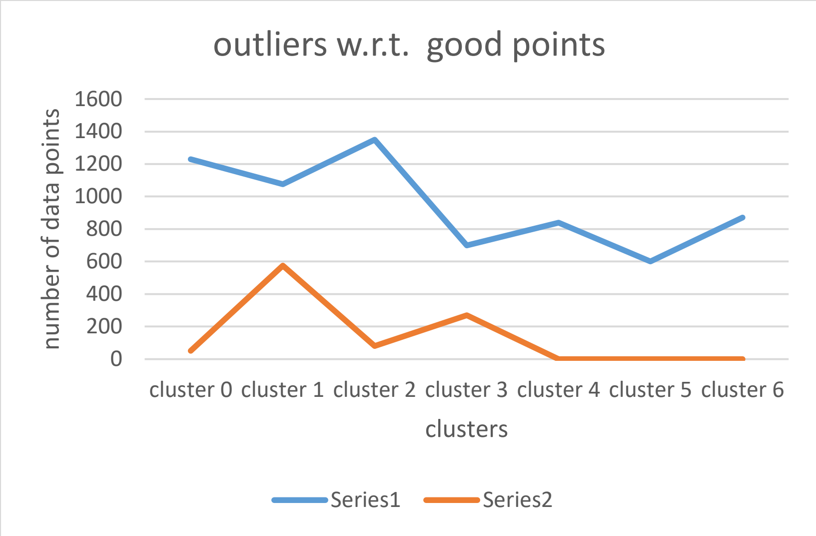


Figure 7.1. k-means Clustering after removing outliers for each cluster

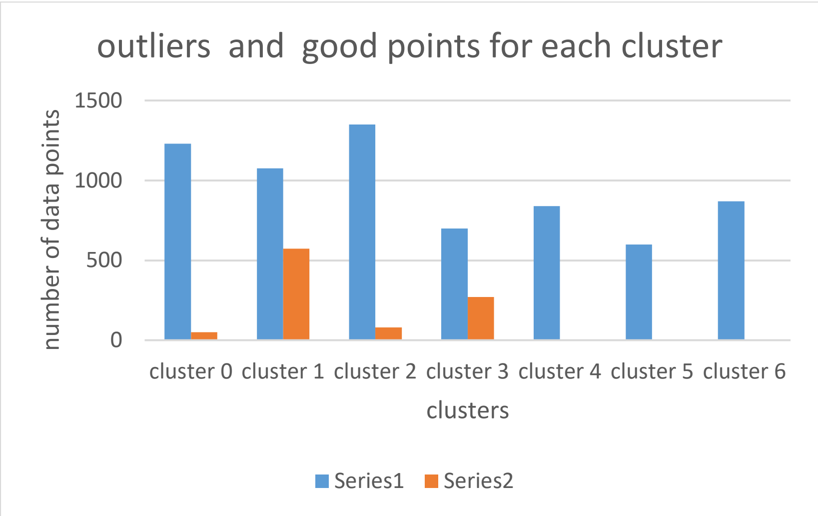


Figure 7.2. k-means Clustering after removing outliers for each cluster

K-means clustering with all the features(150 Aps) , Distortion Score Elbow for k-means Clustering is shown below in Figure 8 where The average of the squared distances from the cluster centres of the individual clusters is used to calculate it. Here we can see the clustering of k-means in seven number of cluster with the distortion score is good.

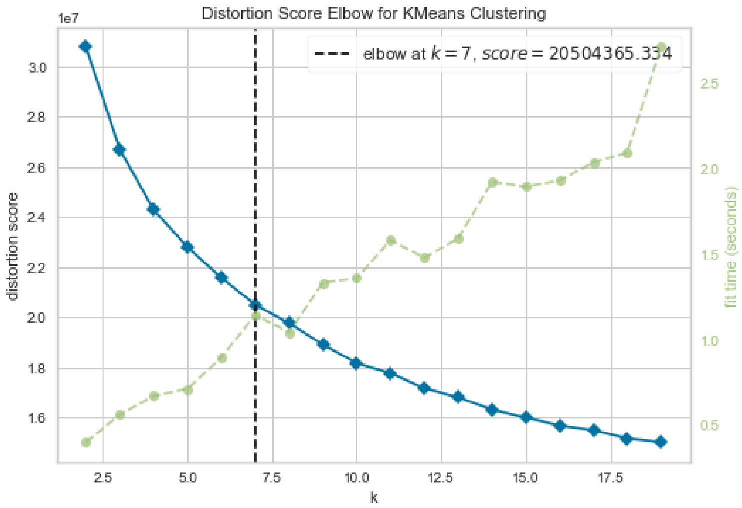


Figure 8 **:** Distortion Score Elbow for k-means clustering algorithm

The graph has been plotted for Davies bouldin, Silhouette after removing the outliers. that is shown in the below where davies bouldin score minimize the ratio of intra cluster and inter cluster variation and measure the average similarity of each cluster with its most similar cluster. we can see the results from figure 9 to 14. These scores evaluates that the clustering model is performing well with cluster number three , five and seven for Davies bouldin and and Silhouette score also quite good when cluster numbers are three and seven and same goes with Calinski Harbasz. That's why we computed others graph to see the cluster number where the model is performing good. For each clustering we have computed The Elbow method is shown for each cluster and k-distance graph for k-means . From these graph we can say that for cluster number seven the result of clustering is quite good.

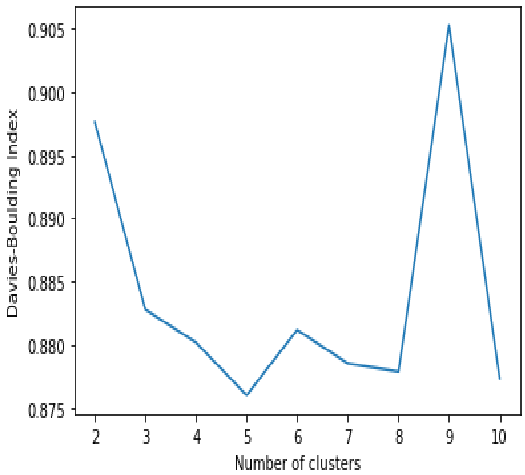


Figure 9: Davies Bouldin index for each cluster on k-means

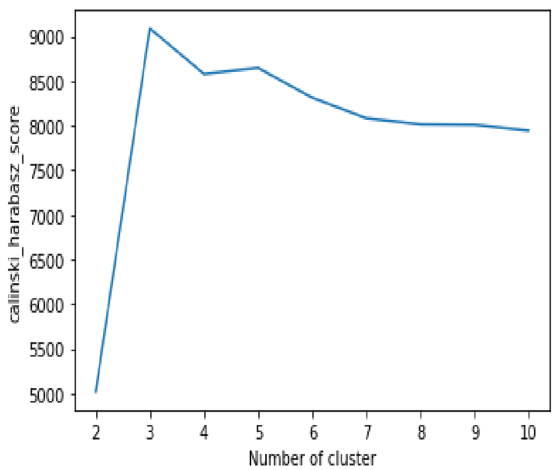


Figure 10: The Calinski Harbasz is shown for each cluster on k-means

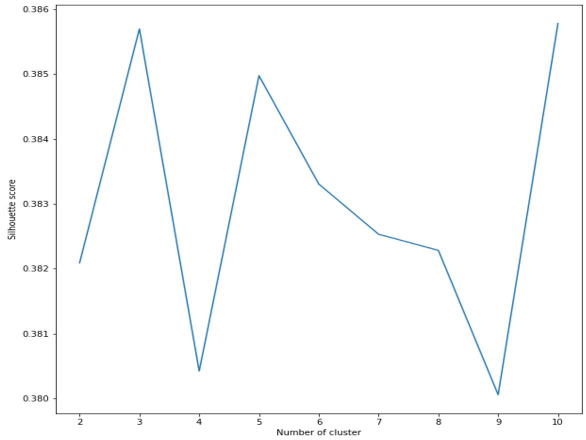


Figure 11: Silhoutte score for each of the clusters on k-means clustering

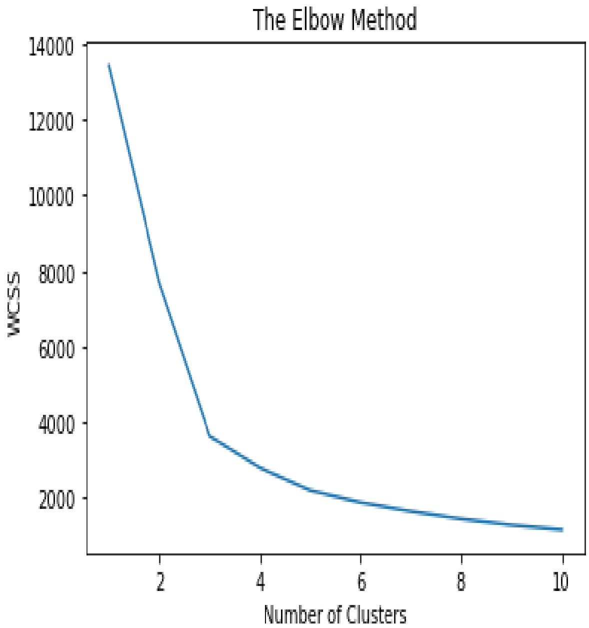


Figure 12: The Elbow method is shown for each cluster of k-means

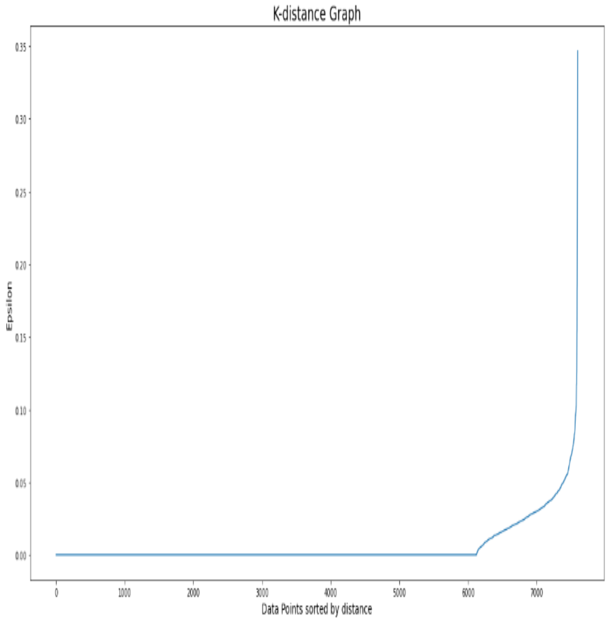


Figure 13: K-distance graph for k-means clustering

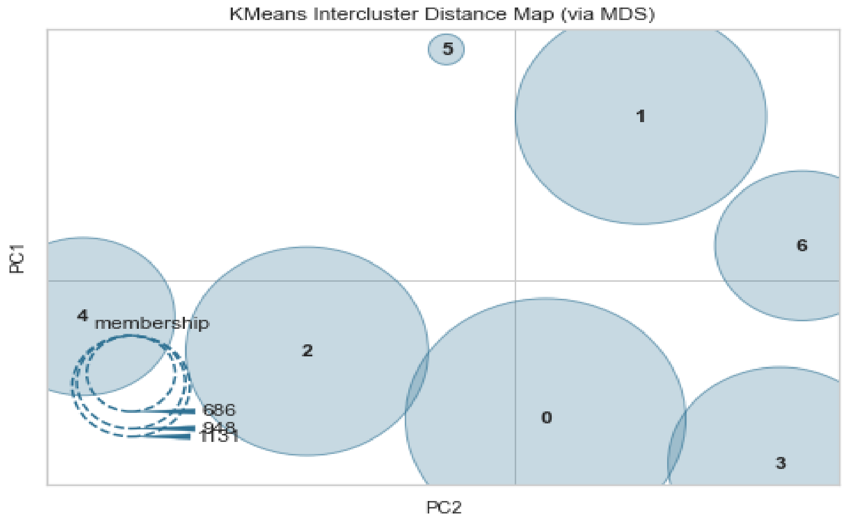


Figure 14: k-means intercluster distance graph has shown in this figure, after using PCA on this dataset

Applying PCA (principal component analysis) before a clustering technique is standard procedure (such as k-means). It is thought to enhance the clustering outcomes in actual use (noise reduction). Additionally, the results of the two techniques differ significantly in that clustering decreases the number of data-points by averaging multiple points according to their expectations or means, whereas PCA helps to reduce the number of features while maintaining variance (in the case of k-means). Therefore, PCA seeks to compress the T features whereas clustering aims to compress the N data-points if the dataset consists of N points with T characteristics each. After using PCA on this dataset. Silhoutte graph has been plotted below in the diagram 15 and from the Figure 16 how the two features clustered can be shown. Here, we can see most of the data points are in their clusters but in cluster six and cluster one we can see some values are negative so few data points are in wrong cluster.

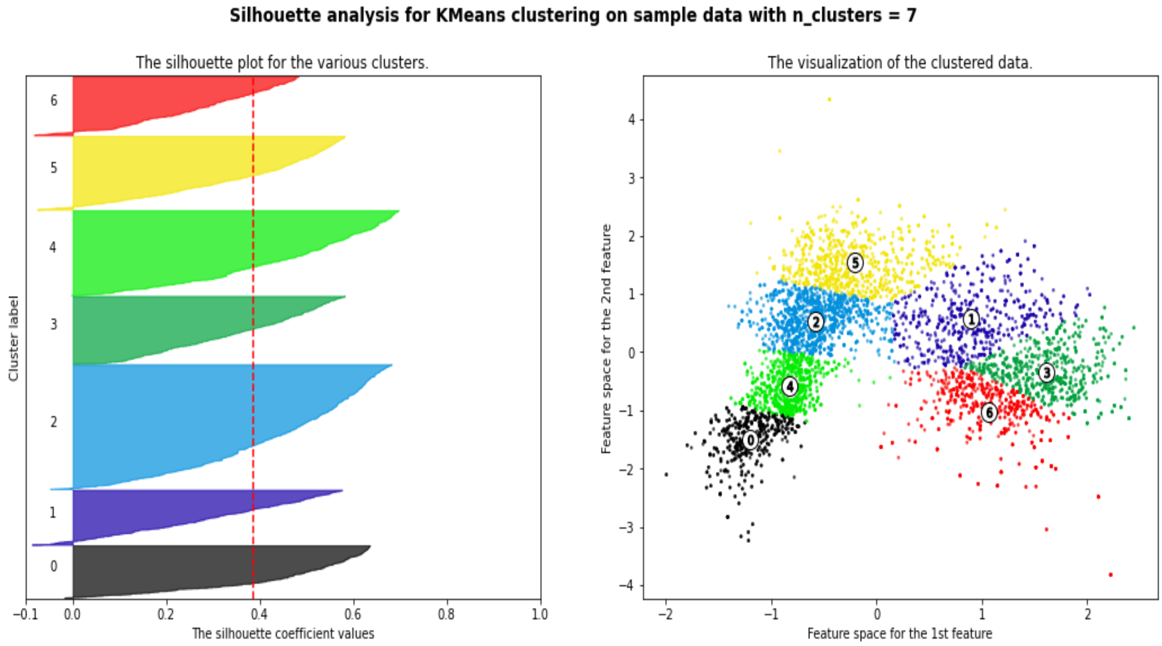


Figure 15 : Silhouette score for the final clustering of k-means

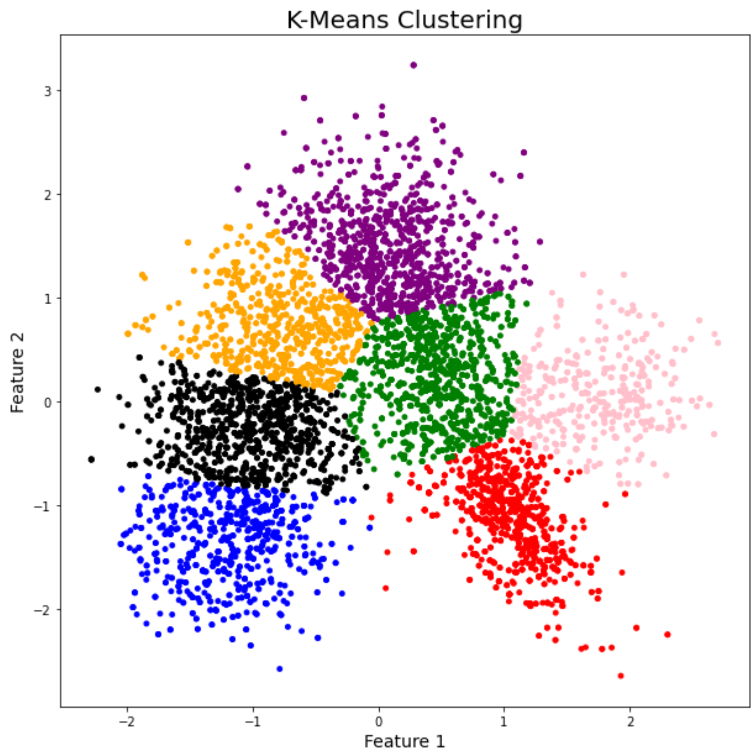


Figure 16 : k-means clustering after using PCA for two features.

Clustering has been applied on the JuIndoorLoc dataset which is consist with location points of the fourth floor of Jadavpur University. The dataset is consist of location points which is from L4-1-1 to L4-42-1 and L4-1-1 to L4-1-21. We have taken the clustering labels after removing the outliers and matched with the existing label of the dataset which is consist of location points from L4-4-9 to L4-42-12 with respect to data points.The clustering is shown in the floor map after manually coloring where each color is representing a cluster , here we can see some of the location points are overlapped but the main problem in this clustering we can define that some data points of the sky colored cluster are in L4-4 to L4-11 and some are in L4-37,L4-39. So we can say some of the datapoints are wrongly clustered from the below diagram.

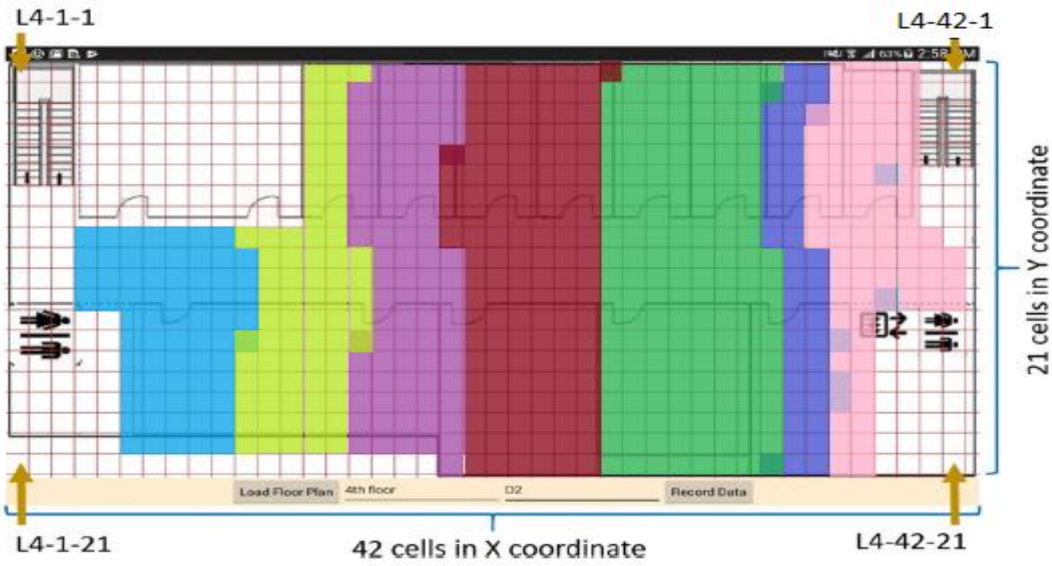
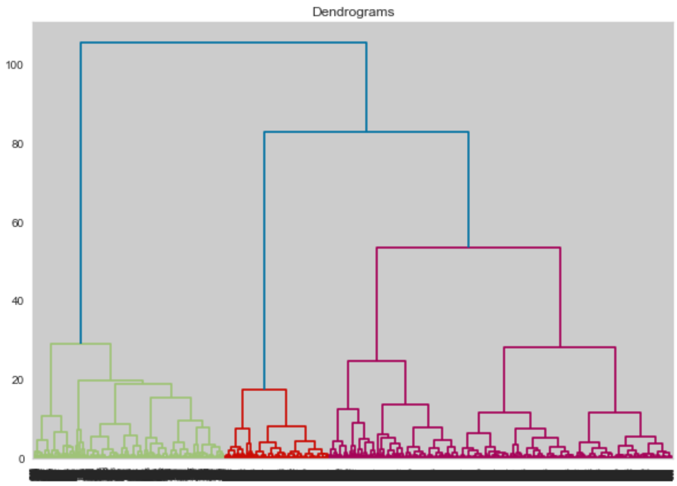


Figure 17 : Performance of k-means clustering on floor map after removing outliers

5.1.2 Hierarchical Clustering

The graph plot on Hierarchical Clustering are shown below where the Hierarchical Clustering has been applied on JUIndoor localization dataset.The main point of drawing the Figures is to detect how the clustering model is working in the dataset.The main feature of Hierarchical Clustering is dendogram which is shown in the Figure 18.



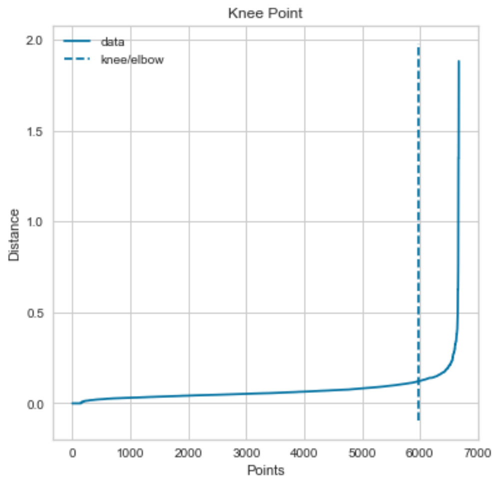


Figure 18 : Dendogram of Hierarchical Clustering on dataset

Figure 19: Knee point of clustering

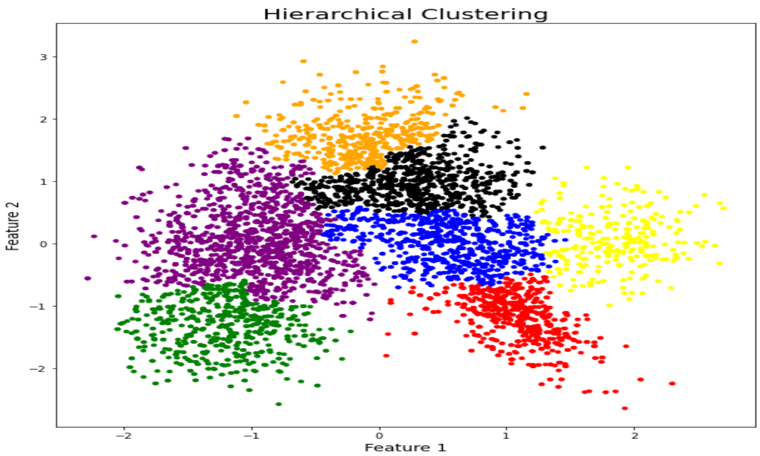


Figure 20 : After using PCA on the dataset, the graph plot of hierarchical clustering

After using PCA on this dataset. Silhoutte graph has been plotted below in the diagram 21 for hierarchical clustering. Here in the silhouette Figure, the clustering have not partitioned perfectly well because some data points are in wrong cluster because some of the values of silhouette are negative.

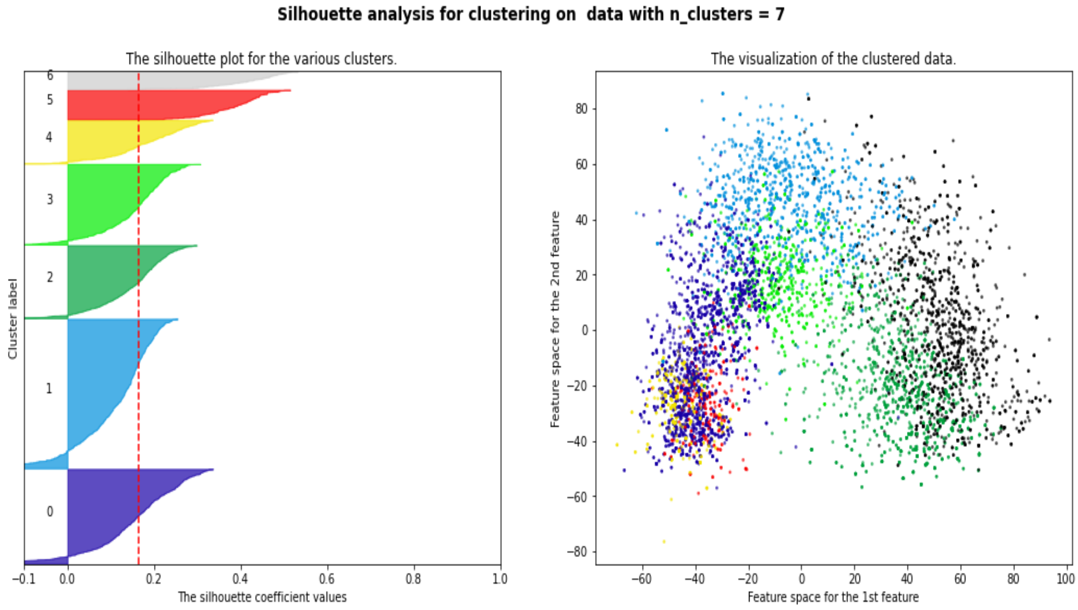
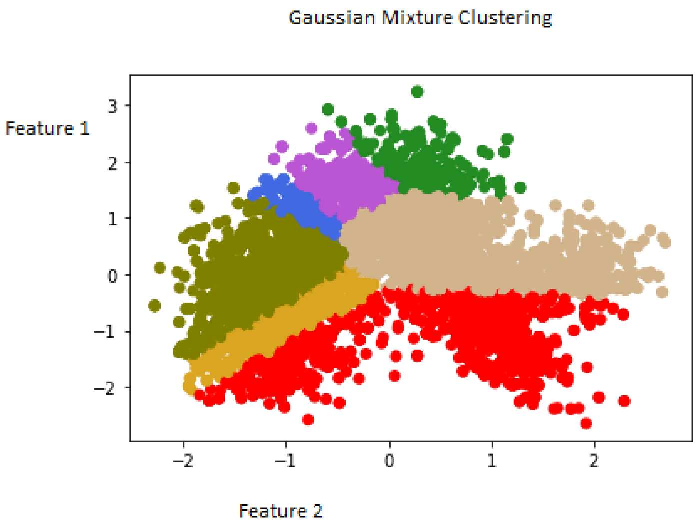


Figure 21 : Silhouette score for the final clustering for Hierarchical Clustering

5.1.3. Gaussian Mixture Model

How the Gaussian mixture clustering is partitioning that is shown in the Figure 22 after using PCA on this localization dataset.



.

Figure 22 : after using PCA on the dataset, the graph plot of GMM

Silhoutte graph has been plotted below in the diagram 23 for Gaussian mixture clustering. where we can see the datapoints are in wrong cluster because the values of clusters are negative.

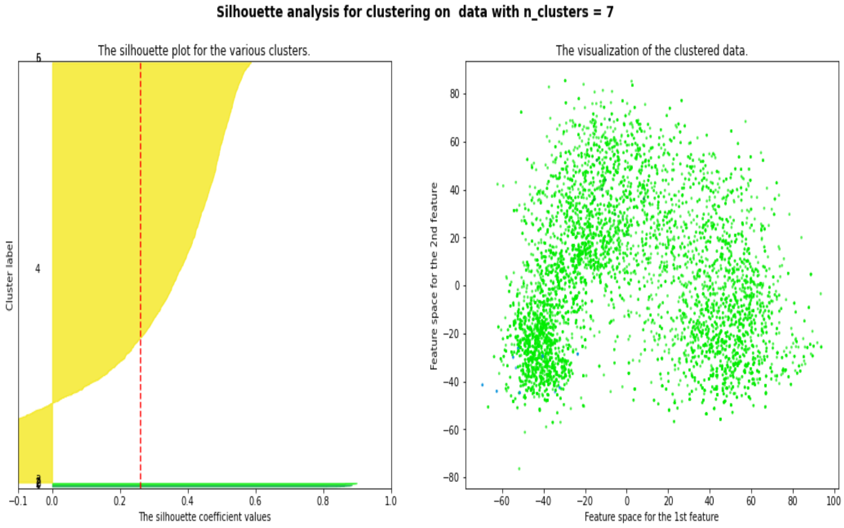


Figure 23 : Silhouette score for the final clustering for Gaussian Mixture

5.2: Experimental results for combined clustering algorithm

The most interesting part of these clustering methods is comparison of these clustering algorithms.The comparison of clustering algorithms gets interesting when the comparison is done with some valid internal validation indices who recognize Clustering validation properly. Internal validation indices measure how well clustering structure has been made without any help of external data. For that Silhouette Coefficient and Calinski-Harabasz has been applied on k-means , Hierarchical Clustering and Gaussian Mixture Clustering.

The performance of partitions for each clustering algorithms can make an issues in cluster analysis. The correct partition can be called correct when the quality of clustering will be valid. The performances measure how well a cluster is separated from other clusters and estimate how each location points are close to their cluster centroids. The validation scores are mainly used to detect clustering algorithm by estimating the better performance for the specific dataset.For each cluster *k* in the clustering algorithms, the score value of Silhouette Coefficient and Calinski-Harabasz has been plotted in the below diagram. The Calinski-Harabasz measures performance based on compactness of cluster that defines how a cluster differ from others and the similarity of location points inside a cluster. The higher score defines the better performances.

From the diagram 23 and 24, it can be observed, The Calinski-Harabasz scores are giving better performances based on the number of clusters for k-means, Hierarchical Clustering because the scores are higher so that defines that the clusters are dense and well separated cluster but Gaussian Mixture score are really low for number of clusters.The Silhouette Coefficient gives a performance score by measuring distances for each location points with respect to other location points. The performance score depends upon how close a location points with other points in a same cluster by measuring the mean distance between a location point and all other location points in a same cluster and also how a location point is far away from the location points who belongs to different clusters. A higher Silhouette Coefficient score defines that the model is making better clustering for the specified dataset.The Silhouette Coefficient scores of k-means, Hierarchical Clustering score are in between 0 to 1 for all the clusters so it defines that k-means, Hierarchical Clustering makes highly dense clustering, but it can be observed that for Gaussian Mixture Silhouette Coefficient scores are between 0 to -1 for some of the clusters, so these clusters are specified as incorrect clusters and overlapping clusters. So, from the observation of these two validation indices k-means and Hierarchical Clustering method has been taken for this Indoor localization dataset.

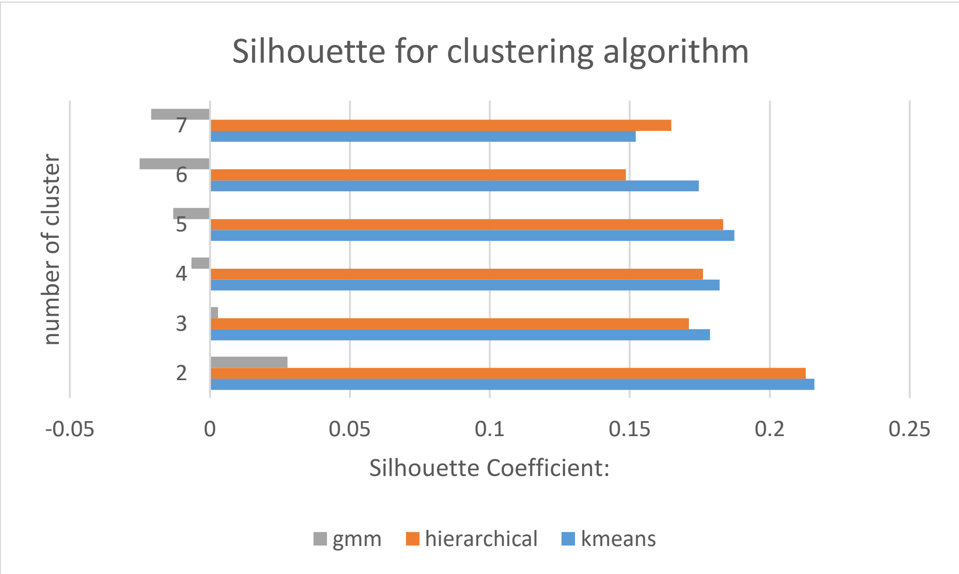


Figure 23 : Silhouette score for GMM, Hierarchical, k-means clustering algorithm

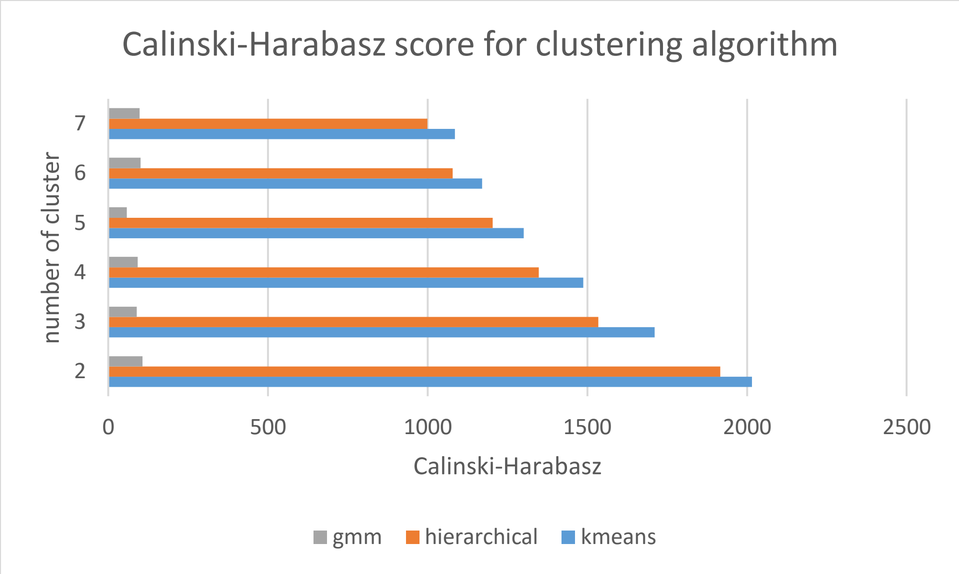


Figure 24 : Calinski-Harabasz score for GMM, Hierarchical, k-means clustering algorithm

Clustering validation technique has been elevated as an important issues in the path of good performance of clustering. Clustering validation are two types. One is internal clustering validation and another is external clustering validation. The internal clustering validation is already discussed on k-means Clustering, Hierarchical Clustering and Gaussian Mixture Clustering.

As external clustering validation, seven validation score has been computed on k-means Clustering and Hierarchical Clustering. The k-means Clustering label and Hierarchical Clustering label are taken as an input for these validation models. The seven validation indices which have been applied on this indoor localization dataset clustering are Adjusted Rand Index (ARI), Normalized mutual info, Adjusted mutual info, Fowlkes mallows, Homogeneity, Completeness and V measure. These external validation indices computes the similarity between the predicted k-means Clustering label and Hierarchical Clustering label. they computes how correctly the partitioning of the dataset occurs based on k-means and Hierarchical Clustering.These external clustering validation indices shows how k-means and Hierarchical Clustering labels are closely related. The differences are depends on partitioning of cluster sizes.

Adjusted Rand Index is computing similarity score between two clustering labels. Clustering labels on JUIndoor Location dataset have positive ARI score so that defines similar clustering. From Figure 35, 26, 27 we can see there are no -1 value in ARI score, so no independent labelling occurred.Adjusted mutual information is using for comparing clustering labels. Here Mutual Information function has been used to measures the agreement of two assignments by ignoring permutations. This measurement makes two different type of normalized versions, Normalized mutual information and Adjusted mutual information.Mutual information values on this dataset are not close to zero that defines that the labels are not largely independent. While the values are close to one so the labels of k-means and Hierarchical clustering have significant agreement. Homogeneity applied on these clustering labels to observe uniform in composition. Completeness is putting the property of statistics on these models for observing the location points of dataset. V-measure is computing the mean of Homogeneity and Completeness. The clustering labels performed good based on Homogeneity, Completeness, V-measure score who are close to one.The higher value of the Fowlkes–Mallows index on these models is defining a greater similarity between the clusters and the benchmark classifications.

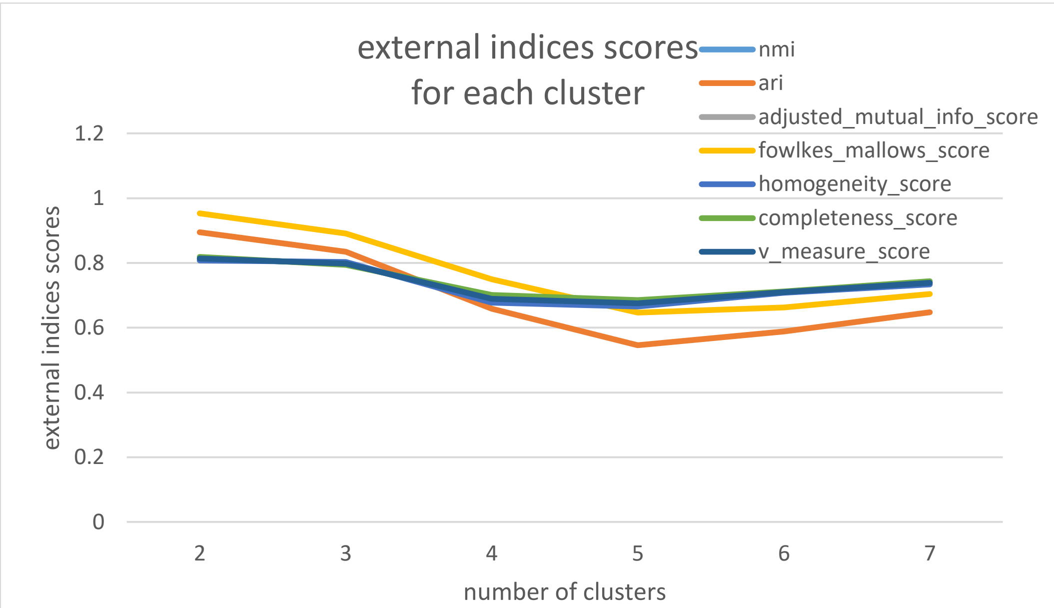


Figure 25: external indices scores for each cluster based on k-means and Hierarchical Clustering

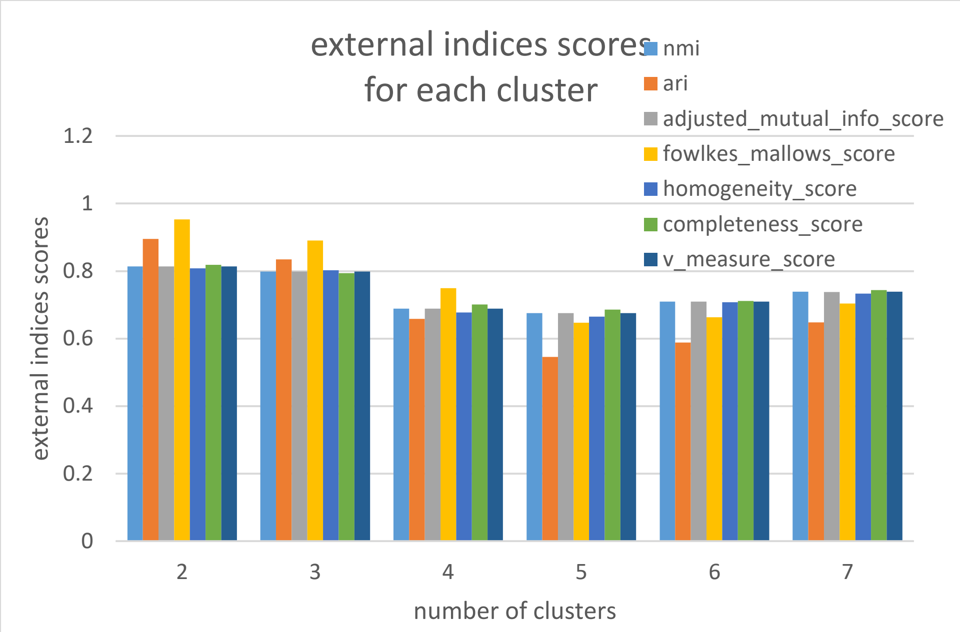


Figure 26 : Bar graph of external indices scores of k-means and Hierarchical Clustering for each cluster

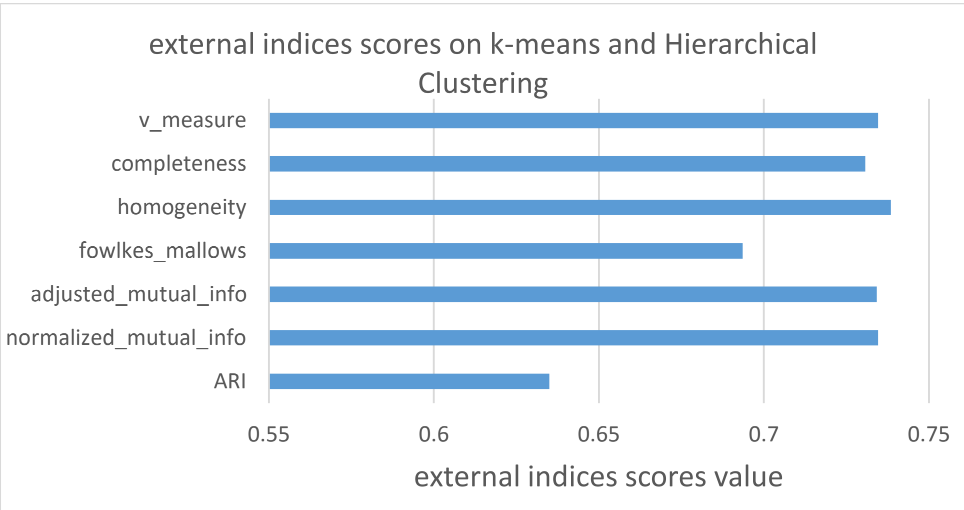


Figure 27 : Difference types of external indices scores based on k-means and Hierarchical Clustering

5.3. Results of Rank base iterative clustering algorithm

performance evaluation of the output of the RICM algorithm, after mapping the cluster in the floor map where each cluster represent a cluster and each cluster is partitioning the regions. The clustering have occurred very well according to the floor map and supervised algorithms and internal validation indices scores. The result is in Figure 28

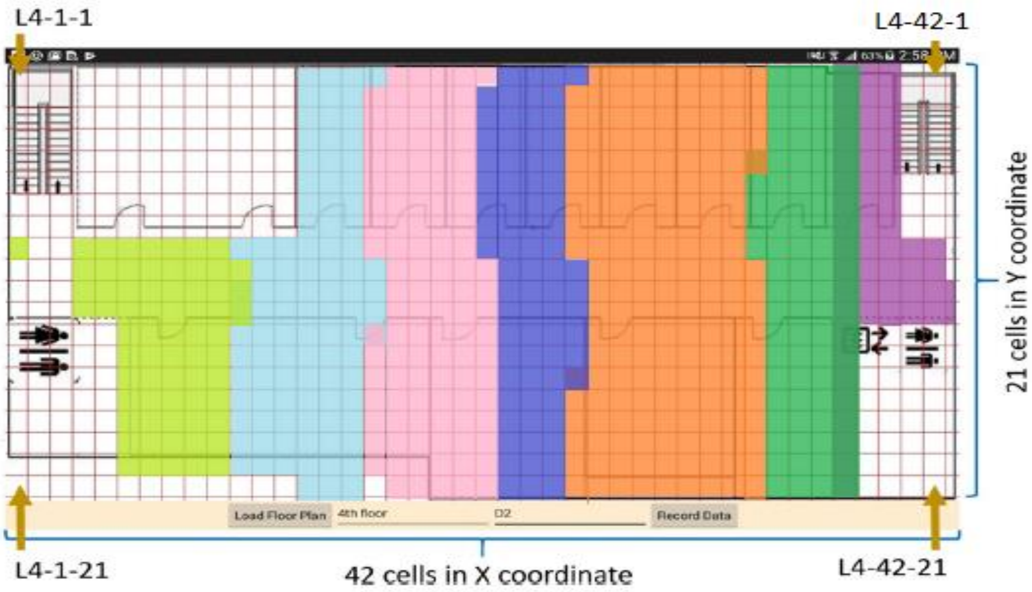


Figure 28 : RICM's Cluster performance on Floor map

For the supervised algorithm, Random forest , SVM, Logistic Regression,KNN , Random forest, decision tree have been applied on the JUIndoor dataset where Random Forest and Decission Tree are working very well and for all these supervised algorithms the label which was taken was the clustering label which is computed by RICM algorithm and result is shown in the Figure 29. For internal validation score Silhuette ,davis bouldin andcalinski-harbash have been computed in the Figure 30.

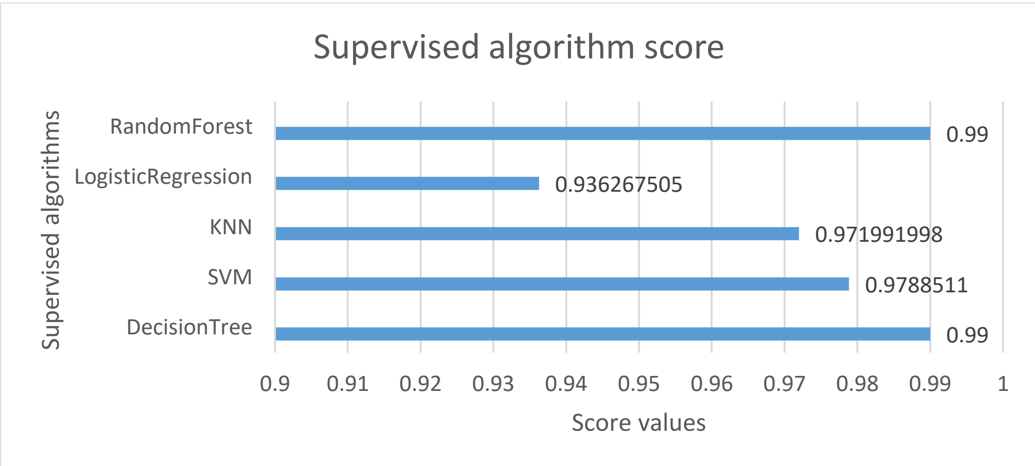


Figure 29 : Different types of Supervised algorithm scores of RICM

silhouette_score	calinski_harabasz_score	davies_bouldin_score
0.118326	427.625497	2.302213

Figure 30 : Three main internal validation scores of RICM

5.4. Feature analysis of obtained clusters

Dominant AP feature's Venn diagram with respect to regions have been plotted in the Figure where numbers are number of AP features in each region and the result is shown in Figure 31 and Figure 32.

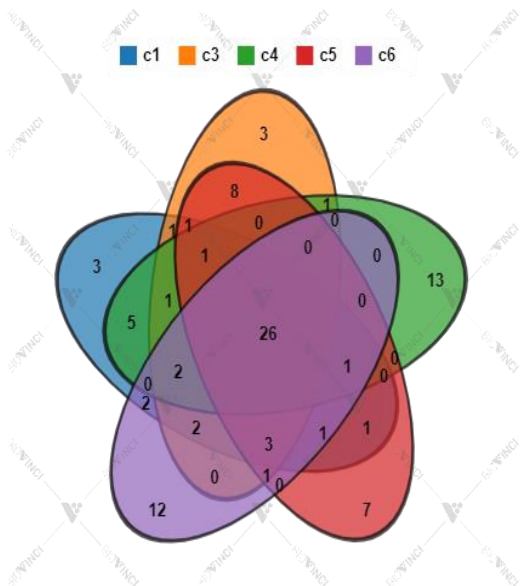


Figure 31: Dominant features of 5 clusters clusters are intersecting with each other.

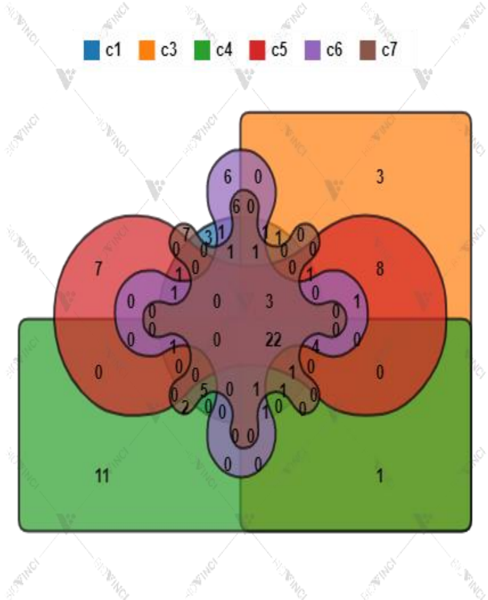


Figure 32. Dominant features of 6 clusters are intersecting with each other.

Dominant AP feature's Bar diagram using values of Standard deviation values has been computed in the following diagram where the cluster five have the most dominant features. The result is shown in the Figure 33 and Figure 34.

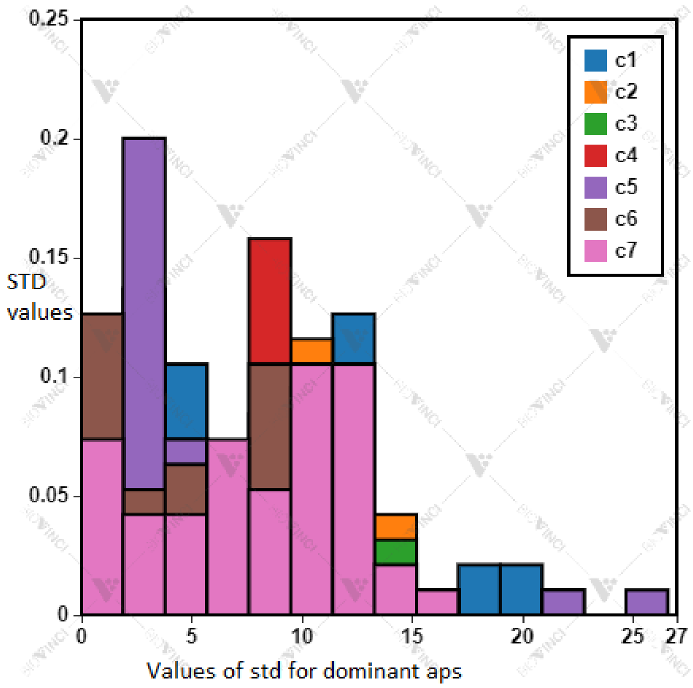


Figure 33 : Dominant AP feature's Bar diagram using values of Standard deviation values for each cluster

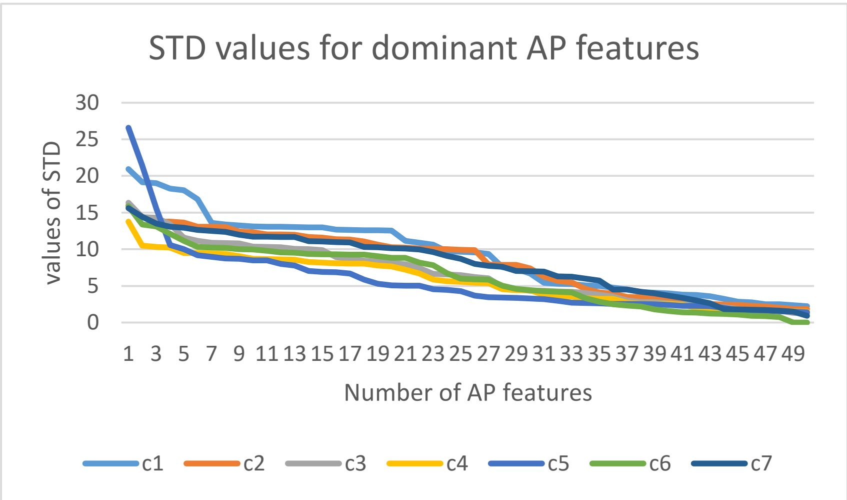


Figure 34 : STD values of Dominant AP features for each clusters

Chapter 6

Conclusions and Future work

A Rank-Based Iterative Clustering (RICM) method is proposed in this work to improve clustering solutions for indoor localization over existing traditional techniques. The combined decision of different clustering algorithms are considered, and that too priority wise (based on the rank of individual algorithms), which minimizes the chance of including two or more data samples in same cluster which are physically located far away. This makes the proposed approach suitable for real-time environments. We have mapped the clusters obtained from RICM method for the 4^{th} floor dataset of JUIndoorLoc, to the original floor map which gives a visualization of this minimization. The validation of the individual clustering algorithms is computed using some internal validation scores. Different supervised classifiers are trained on the clustered dataset obtained as output of the proposed unsupervised approach, and tested with random data samples, which results in a maximum accuracy in a range of 97 to 99 %.

In future, we will focus on optimizing the number of target clusters and improving the feature sets for individual cluster regions to make the localization more efficient. The effect of different devices on clustering will also be investigated.

References

1. Saha, Arka & Sadhukhan, Pampa. (2015). A Novel Clustering Strategy for Fingerprinting-based Localization System to Reduce the Searching Time. 10.1109/ReTIS.2015.7232937.

2.Berz, Everton & Tesch, Deivid & Hessel, F.. (2015). RFID indoor localization based on support vector regression and k-means. 2015. 1418-1423. 10.1109/ISIE.2015.7281681.

3.Gomes, Rahul & Ahsan, Mostofa & Denton, Anne. (2018). Random Forest Classifier in SDN Framework for User-Based Indoor Localization. 10.1109/EIT.2018.8500111.

4.Xue, Jianzhe & Liu, Junyu & Sheng, Min & Shi, Yan & Li, Jiandong. (2020). A WiFi fingerprint based high-adaptability indoor localization via machine learning. China Communications. 17. 247-259. 10.23919/J.CC.2020.07.018.

5. Li, Peihao & Yang, Xu & Yin, Yuqing & Gao, Shouwan & Niu, Qiang. (2020). Smartphone-Based Indoor Localization With Integrated Fingerprint Signal. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2974038.

6. Zafari, Faheem & Papapanagiotou, Ioannis & Devetsikiotis, Michael & Hacker, Thomas. (2017). An iBeacon based Proximity and Indoor Localization System.

7. Yang, Bo & Guo, Luyao & Guo, Ruijie & Zhao, Miaomiao & Zhao, Tiantian. (2020). A Novel Trilateration Algorithm for RSSI-based Indoor Localization. IEEE Sensors Journal. PP. 1-1. 10.1109/JSEN.2020.2980966.

8. Palacios, Joan & Bielsa, Guillermo & Casari, Paolo & Widmer, Joerg. (2019). Single– and Multiple–Access Point Indoor Localization for Millimeter Wave Networks. IEEE Transactions on Wireless Communications. PP. 1-1. 10.1109/TWC.2019.2899313.

9. Nazemzadeh, Payam & Fontanelli, Daniele & Macii, D.. (2013). An indoor position tracking technique based on data fusion for ambient assisted living. 7-12. 10.1109/CIVEMSA.2013.6617387.

10. Razavi, Seyed Alireza & Valkama, Mikko & Lohan, Elena Simona. (2015). K-means Fingerprint Clustering for Low-Complexity Floor Estimation in Indoor Mobile Localization. 10.1109/GLOCOMW.2015.7414026.

11. Zhong, Yazhou & Wu, Fei & Zhang, Juan & Dong, Bo. (2016). WiFi indoor localization based on K-means. 663-667. 10.1109/ICALIP.2016.7846667.

12. Zhang, Chenbin & Qin, Ningning & Xue, Yanbo & Yang, Le. (2020). Received Signal Strength-Based Indoor Localization Using Hierarchical Classification. Sensors. 20. 1067. 10.3390/s20041067.

13. Malekzadeh, Parvin & Salimibeni, Mohammad & Atashi, Mohammadamin & Barbulescu, Mihai & Plataniotis, Konstantinos & Mohammadi, Arash. (2019). Gaussian Mixture-based

Indoor Localization via Bluetooth Low Energy Sensors. 1-4. 10.1109/SENSORS43011.2019.8956950.

14. Liu, Minxiang & Leung, Henry & Maddumabandara, Achintha. (2015). Experimental Evaluation of Indoor Localization Using Wireless Sensor Networks. IEEE Sensors Journal. 15. 1-1. 10.1109/JSEN.2015.2438193.

15. Improving UWB Based Indoor Positioning in Industrial Environments through Machine Learning by Sivanand Krishnan, Member, IEEE, Rochelle Xenia Mendoza Santos, Enhao Ranier Yap and May Thu Zin

16 Krishnan, Sivanand & Santos, Rochelle & Yap, Enhao & Zin, May. (2018). Improving UWB Based Indoor Positioning in Industrial Environments Through Machine Learning. 1484-1488. 10.1109/ICARCV.2018.8581305.

17. Hou, Yanbin & Yang, Xiaodong & Abbasi, Qammer. (2018). Efficient AoA-Based Wireless Indoor Localization for Hospital Outpatients Using Mobile Devices. Sensors. 18. 3698. 10.3390/s18113698.

18. Wen, Kai & Seow, Chee Kiat & Tan, Soon. (2020). An Indoor Localization and Tracking System Using Successive Weighted RSS Projection. IEEE Antennas and Wireless Propagation Letters. PP. 1-1. 10.1109/LAWP.2020.3011993.

19. Zhao, Kun & Zhao, Tiantian & Zheng, Zhengqi & Yu, Chao & Ma, Difeng & Rabie, Khaled & Kharel, Rupak. (2020). Optimization of Time Synchronization and Algorithms with TDOA Based Indoor Positioning Technique for Internet of Things. Sensors. 20. 10.3390/s20226513.

20. https://programmer.ink/think/k-means-clustering.html

21. Nazemzadeh, Payam & Moro, Federico & Fontanelli, Daniele & Macii, David & Palopoli, Luigi. (2015). Indoor Positioning of a Robotic Walking Assistant for Large Public Environments. IEEE Transactions on Instrumentation and Measurement. 64. 1-1. 10.1109/TIM.2015.2437637.

22. Zhu, Xuanyu. (2020). Indoor Localization Based on Optimized KNN. Network and Communication Technologies. 5. 34. 10.5539/nct.v5n2p34.

23. Chriki, Amira & Touati, Haifa & Snoussi, Hichem. (2017). SVM-based indoor localization in Wireless Sensor Networks. 1144-1149. 10.1109/IWCMC.2017.7986446.

24. Mehmood, Hamid & Tripathi, Nitin & Tipdecho, Taravudh. (2010). Indoor Positioning System Using Artificial Neural Network. Journal of Computer Science. 6. 10.3844/jcssp.2010.1219.1225.

25. Roy, Priya & Chowdhury, Chandreyee & Ghosh, Dip & Bandyopadhyay, Sanghamitra. (2019). JUIndoorLoc: A Ubiquitous Framework for Smartphone-Based Indoor Localization Subject to Context and Device Heterogeneity. Wireless Personal Communications. 106. 10.1007/s11277-019-06188-2.