PURPOSE-LED
PUBLISHING™

PAPER • OPEN ACCESS

Geomagnetic field fingerprint from wristband and smartphone clusterization for indoor localization

View the article online for updates and enhancements.

Geomagnetic field fingerprint from wristband and smartphone clusterization for indoor localization

F A Ahdiy[1] and I N Yulita[1]

[1]Universitas Padjadjaran, Jl. Raya Bandung Sumedang KM.21, Hegarmanah, Jatinangor, Sumedang, West Java, Indonesia, 45363

fatanaufa97@gmail.com

Abstract. Nowadays, many wearable and smartphones have their built-in sensors that can be used to obtain the information about location to make an Indoor Localization Systems (ILS). It could be helpful on developing Smart City Systems, where ILS can handle the information collection about space and placement inside a room or building. In this research, we used the datasets from UCI Machine Learning Repository, which consists data of geomagnetic field fingerprint captured by both smartphone and smartwatch sensors from certain locations with timestamps provided, and then followed by clusterization processes to determine whether each device could effectively used for ILS based on radiation emitted in the same environment. For the methods, we used K-means (for k=2 to k=10) and Hierarchical Clustering. After several experiments, hierarchical clustering works best to cluster the data since the outliers problem doesn't occurred in this algorithm, and every instance get clustered with high homogenity for each member and every clusters differs enough in characteristics. Newer device with more sensor needed for next research.

1. Introduction

Smart city has been a hot issue today. As the smart city concepts sprouted massively around the world, the needs of newer and advanced technology to accomplish most of the concepts becomes clearly mandatory, and the proof of it could be seen at mobile devices. The development of mobile devices has been increased significantly in last decade. There are a lot of technology enhancements and innovations implemented in today's mobile devices structures. The improved technology leverages the human habit in using mobile devices. For example, the uses of smartphone today are not limited to make a call and send a message. People can use their smartphone to play games, browse the internet, access the social media, etc. One of the uses of smartphone that can easily found in ever day life is navigation.

Almost every smartphone produced in last decade have Global Positioning System (GPS) built in their machines for navigation purpose. GPS is useful when we need a guidance to a specific location. But GPS has some sort of weaknesses. Its accuracy to navigate in indoor or underground environment could be decreased because of satellite signal blockage. It does have an impact to create an Indoor Localization System (ILS). The Indoor Localization Systems (ILSs) is a system that could be used in Ambient Assisted Living (AAL) scenarios [1], robotic applications [2], and indoor navigation in large environments such as malls, campuses, and airports [3]. Due to GPS' weakness, the community of

makers started to build alternatives which can be used to solve the indoor localization problem. As one of the available alternatives, some modern ILSs are based on the use of a variety of sensors and devices that are embedded in smartphones [4].

This study presents related researches of indoor localization methods based on geomagnetic field fingerprinting [5] which the radiation record will be clustered to find out whether this method is effective enough or not to solve indoor localization problem. This research uses k-means clustering and hierarchical clustering to find out the best number of clusters with different algorithm. With this research, it is hoped that more detail can be found on which algorithm more suited to solve the problem and how well the clustered data to be used for analyzing the method. This method builds a model based on training data [6]. Furthermore, the model is used to predict test data [7]. With this research, it is hoped that more detail can be found on which algorithm more suited to solve the problem and how well the clustered data to be used for analyzing the method.

2. Method

The process of partitioning or grouping a given set of patterns into disjoint clusters is known as clustering [8]. In this research, we used two kind of clustering algorithm, K-means clustering and hierarchical clustering.

2.1. K-means clustering

Big data sets can be clustered by using K-means clustering algorithm. K data elements are selected as initial centers [9]. The distances of all data elements are calculated by Euclidean distance formula. Continue the process with moving data elements to the appropriate cluster which having less distance to centroids. K-means algorithm is used for clustering. It is a type of unsupervised learning where there is no idea about the class or labels for any data and need to discover the clusters without this information [10].

2.1.1 Algorithm. Algorithm of K-means clustering [11]:
1) Select the number of 'c' cluster centers.
2) Find out the distance between each data point and cluster centers using Euclidean distance formula:

$$d(x,y) = \sqrt{\Sigma(x_i - y_i)} \tag{1}$$

where i = 1 to n.
3) Assign the data point to that cluster center whose distance from the cluster center is minimum as compared to all the cluster centers.
4) Recalculate the new clusters center using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i \tag{2}$$

c_i represents the number of data points in i^{th} cluster
5) Recalculate the distances of each data point with new cluster centers.
6) If there is no reassigning of the data points, then stop, otherwise repeat step 3 onwards.

2.2. Hierarchial Clustering

One method in hierarchical clustering is The Ward method. The Ward method iteratively merges two clusters at a time, making sure the merger will increase the total within-cluster variance by the minimum possible [12].

2.2.1. Algorithm. Algorithm of The Ward method:

1) Set K=N. Each cluster in $S=\{S_1, S_2, \ldots, S_K\}$ is composed of a single different entity, making it a singleton.

2) Merge the clusters S_i and S_j which are the closest as per (1), creating the new cluster $S_{(Si \cup Sj)}$. Remove references to the old clusters S_i and S_j, as well as their centroids $c_{(Si)}$ and $c_{(Sj)}$.

3) Set the centroid of the new cluster to the cluster's center of gravity.

4) Reduce K in 1, if K is still bigger than the desired number of clusters go back to Step 2.

3. Methodology

To find which algorithm has best clustering result in term of likelihood and errors, this paper used Geomagnetic Field Fingerprint dataset from UCI Machine Learning Repository which can be used publicly to be clustered using k-means clustering and hierarchical clustering.

3.1. Dataset description

The mentioned authors used a smartphone and a smartwatch to capture the geomagnetic field fingerprints from an environment by two users. The data acquisition process involved two campaigns performed at the first floor of the Institute of Information Science and Technologies (ISTI), inside the Italian National Council (CNR) building. The data acquisition campaign has been performed by wearing two devices simultaneously: a smartphone and a smartwatch. The used smartphone model is the Sony Xperia M2, while the smartwatch is the LG W110G Watch R. Both devices were running the Android OS with dedicated apps developed to collect the data [13]. Data gathered during the study comprises both physical parameters and Wi-Fi access points information, but this paper won't use the data from Wi-Fi access point. For the details about the dataset acquisition, please kindly refer to the related paper because this paper will use only the dataset without detailing the origin of the data. The dataset contains:

- Timestamp;
- AccelerometerX, AccelerometerY, AccelerometerZ;
- MagneticFieldX, MagneticFieldY, MagneticFieldZ;
- Z-Axis Angle (Azimuth), X-Axis Angle (Pitch), Y-Axis Angle (Roll);
- GyroscopeX, GyroscopeY, GyroscopeZ.

There are four datasets, with each dataset record the different data. First dataset recorded the first campaign using phone, second dataset recorded the first campaign using watch. The first campaign was performed by one person using phone and watch at the same time. Third dataset recorded second campaign using phone, and forth dataset record the second campaign using watch. The second campaign was performed by one person using phone and watch at the same time. First campaign took the place first, later followed by second campaign, so each instance in their respective dataset have unique timestamp.

3.2. Preprocessing mechanism

Every record in the dataset are numerical. Total attribute for each dataset is 13 attributes, with the details as mentioned in previous section. Each dataset has similar parameters (attributes), but the smartphone didn't record the gyro data. Table 1 shows the parameters collected for each device, and table 2 shows the characteristic of the dataset.

Table 1. Parameters collected for each device.

Parameter	Phone	Watch
Accelerometer	V	V
Geomagnetic	V	V
Gyroscope	-	V
Orientation	V	V

Table 2. Main characteristics of datasets.

File Name	Number of Campaign	Device	Number of Attributes	Number of Instances	Missing Value	Type
measure1_smartphone_sens	1	Phone	13	18,354	No	Multivariate
measure1_smartwatch_sens	1	Watch	13	58,374	No	Multivariate
measure2_smartphone_sens	2	Phone	13	17,786	No	Multivariate
measure2_smartwatch_sens	2	Watch	13	58,374	No	Multivariate

Because of different number of instances with phone device between first campaign and second campaign, feature selection is performed to make both campaigns have same number of instances. Due the lack of hardware resources which has an impact on processable dataset size, every dataset must be uniformed in number of instances. As the result, the number of instances in first phone campaign is reduced by 568, and the number of instances in both watch campaigns is reduced to 17786. Detail can be found in table 3.

Table 3. Feature selection result on dataset.

Number of Campaign	Device	Number of Instances (Original)	Number of Instances (After feature selection)	Number of reduce
1	Phone	18,354	17,786	568
1	Watch	58,374	17,786	40,588
2	Phone	17,786	17,786	0
2	Watch	58,374	17,786	40,588

Next process linked to dataset multivariate characteristic. Each attribute category (i.e. Accelerometer, including its X, Y, and Z) have their own measurement unit. This raise a new problem as the dataset couldn't be processed to the next step (clusterization) without standardize/transform it first. Assume the dataset has gaussian distribution (bell curve), then the data should be transformed into z_score which being calculated by equation below:

$$z_{score} = \frac{X - \bar{X}}{SD_X} \tag{3}$$

$$SD_X = \frac{(X - \overline{X})^2}{n - 1} \tag{4}$$

X represent the original score, $\bar{X}$ represents the mean of sample distribution, n represents the number of samples, and SD_X represents the standard deviation. Then, the dataset value for each instance being replaced with the new z_{score}. This step is performed to every dataset available for this paper. Table 3 shows the example of record in original dataset, and table 4 shows the example of record in dataset after being standardized.

Table 4. Smartphone sensors table - example of data entry.

Timesta mp (Unix)	Accelero meterX	Accelero meterY	Accelero meterZ	Magneti cFieldX	Magneti cFieldY	Magneti cFieldZ	Z-AxisAngl e(Azimut h)	X-AxisAngl e(Pitch)	Y-AxisAngl e(Roll)	GyroX	GyroY	GyroZ
1422628 680942	0.209	3.566	8.956	-1.024	3.676	-46.664	1.436	-21.878	1.176	-0.005	0.028	0.009

Table 5. Smartphone sensors table - example of data z_score.

Timesta mp (Unix)	Accelero meterX	Accelero meterY	Accelero meterZ	Magneti cFieldX	Magneti cFieldY	Magneti cFieldZ	Z-AxisAngl e(Azimut h)	X-AxisAngl e(Pitch)	Y-AxisAngl e(Roll)	GyroX	GyroY	GyroZ
1422628 680942	-0.33258	3.34721	2.12507	0.18274	-1.21861	-2.27134	-1.55757	-2.6589	-0.30323	0.01419	0.07352	0.04318

3.3. Clusterization mechanism

After the dataset has been standardized, the next process is clustering. The algorithms used are K-means clustering and hierarchical clustering, both performed with python-based software in a machine with this specification:

Table 6. Machine specification.

| Name	Value
Processor	Intel® Core™ i7-5500U CPU @ 2.40GHz (4 CPUs), ~2.4GHz
Memory	8192MB RAM
Operating System	Windows 10 Education 64-bit (10.0, Build 17134)

K-means clustering performed on all dataset, followed by hierarchical clustering on the same subjects. The analysis will be based on result, excluding the performance matter as time complexity, process time, etc.

4. Results

As mentioned before, this research using K-means clustering and hierarchical clustering.

4.1. K-means clustering phase

K-means clustering performed from k=2 until k=10, which the goal is to find the best K value and how many clusters build with their respective K values. To determine the best K value, elbow point method is used to count the average value of the closest distance between each data in the cluster and their centroid, and then compare each K values.
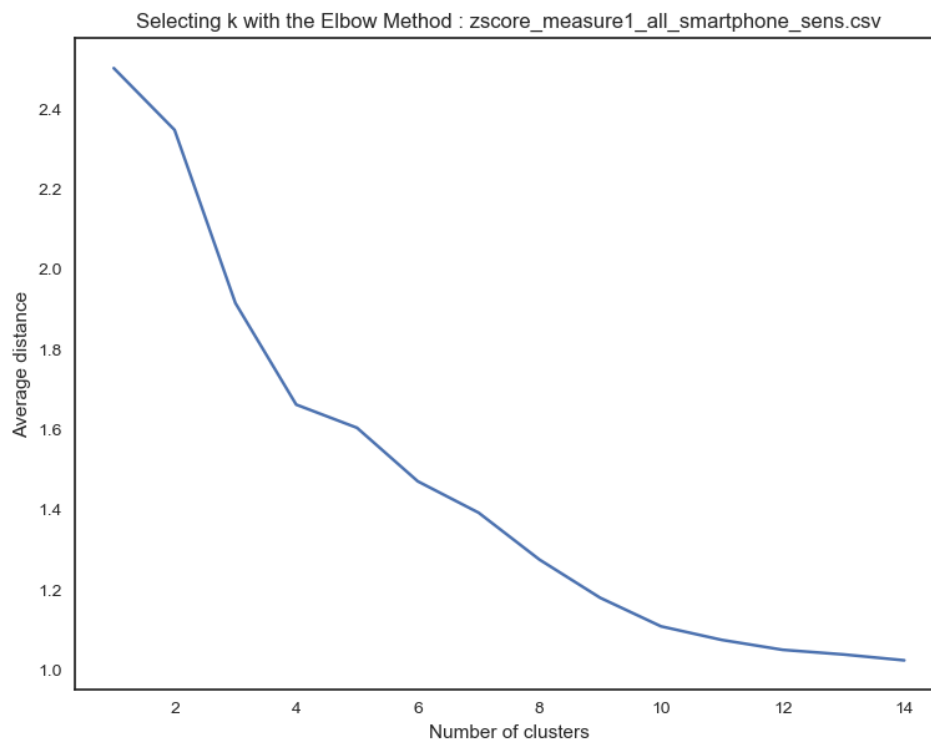
a
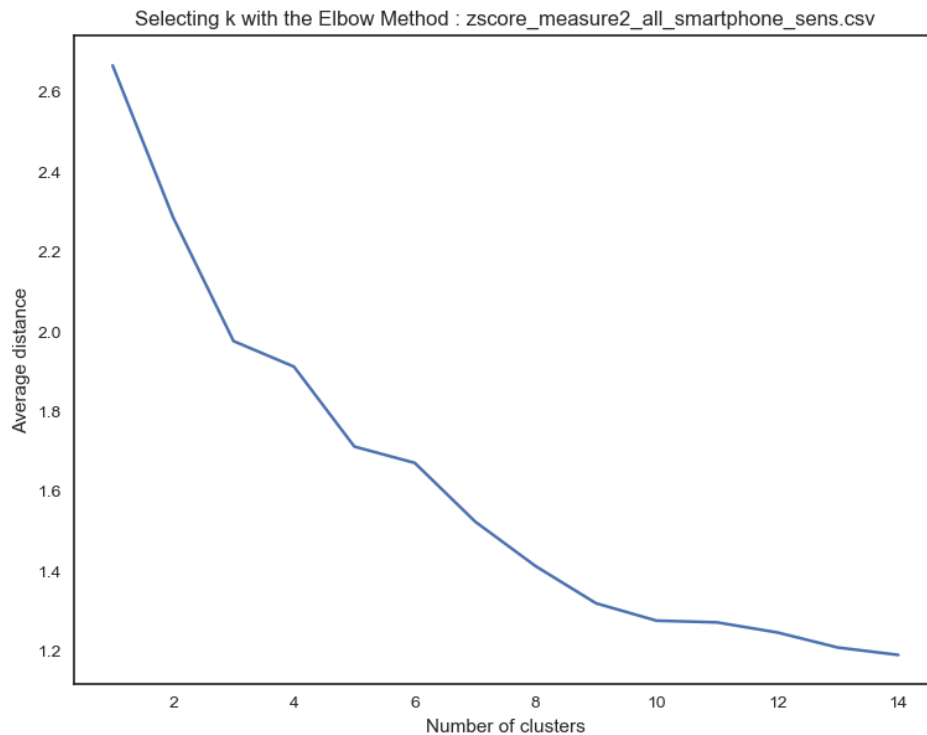
b
Figure 1. Elbow point graphics of smartphone radiation on (a) first campaign; (b) second campaign.

Figure 1 shows each point for each K value (represented by X-axis) from 1 to 14 from smartphone campaigns, while figure 2 shows the same from smartwatch campaigns. To find which one is the elbow point, find the farthest line between the point and the line between first cluster point and last cluster point.
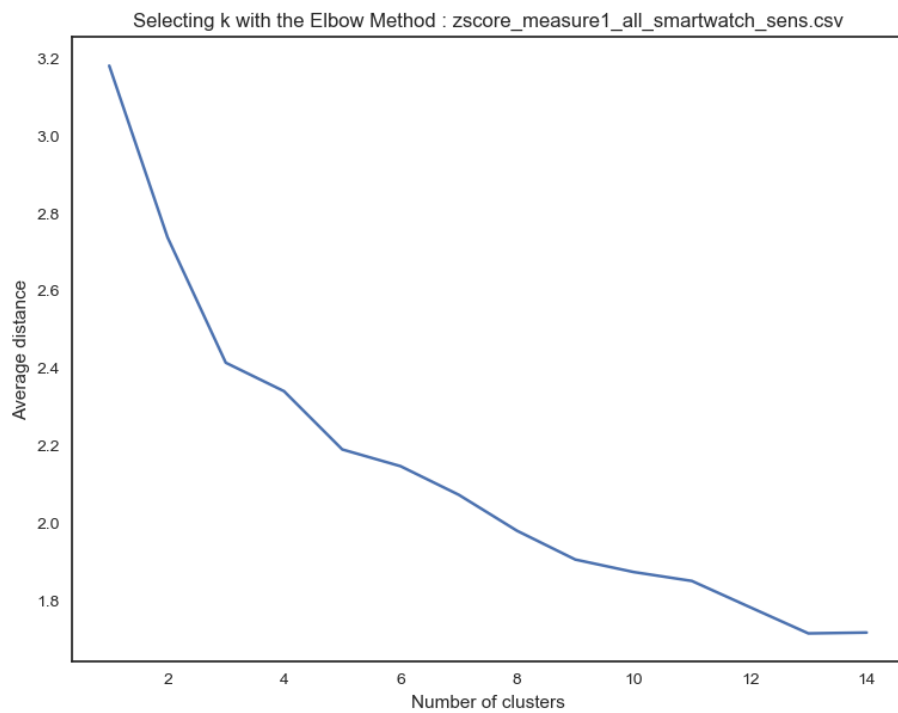
a
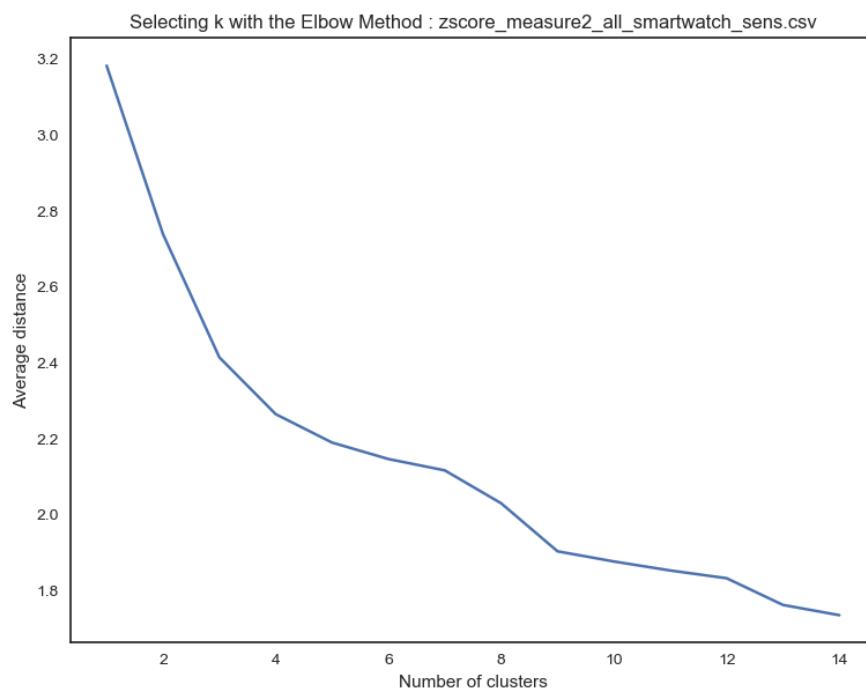
b

Figure 2. Elbow point graphics of smartwatch radiation on (a) first campaign; (b) second campaign.

From the figure above, it can be concluded the elbow point for figure 1(a) is located on number 4, figure 1(b) is located on number 5, figure 2(a) is located on number 3, and figure 2(b) is located on number 3. They become our K values for each respective dataset. After the best K value has been determined, continue with finding centroid value of each clusters available.

Table 7. Centroid values from clustered first phone campaign data with K-means clustering

Cluster	AccX	AccY	AccZ	MFX	MFY	MFZ	ZAA	XAA	YAA	GX	GY	GZ
A	-0.084	0.498	-0.188	0.274	-1.128	0.826	0.124	-0.411	-0.067	0	0	0
B	0.004	0.042	0.154	1.031	0.665	-0.674	1.377	-0.077	0.005	0	0	0
C	-0.064	-5.437	-5.447	-0.037	2.167	3.333	-0.543	6.156	-0.226	0	0	0
D	0.060	-0.126	0.287	-0.824	0.282	-0.317	-0.908	0.053	0.054	0	0	0

Table 8. Centroid values from clustered second phone campaign data with K-means clustering

Cluster	AccX	AccY	AccZ	MFX	MFY	MFZ	ZAA	XAA	YAA	GX	GY	GZ
A	0.085	0.715	-0.518	0.303	-1.104	1.008	0.204	-0.706	0.096	0	0	0
B	-9.540	-5.185	-5.639	2.340	-0.969	1.511	0.115	4.502	-10.94	0	0	0
C	0.094	-1.034	0.723	-0.821	-0.129	0.144	-0.632	0.960	0.086	0	0	0
D	0.107	-0.391	0.276	1.057	0.825	-0.647	1.312	0.367	0.097	0	0	0
E	-0.150	0.541	-0.210	-0.864	0.565	-0.666	-1.154	-0.450	-0.123	0	0	0

Table 9. Centroid values from clustered first watch campaign data with K-means clustering

| Cluster	AccX	AccY	AccZ	MFX	MFY	MFZ	ZAA	XAA	YAA	GX	GY	GZ
A	-0.037	-0.335	-0.584	-0.250	0.144	1.046	-1.469	0.485	-0.002	-0.131	0.028	0.0002
B	-0.707	2.230	1.560	0.557	-0.992	-1.143	-0.088	-2.572	-0.588	0.007	-0.058	-0.062
C	0.108	-0.484	-0.272	0.223	0.465	0.745	1.087	0.331	0.086	-0.009	-0.012	-0.002

Table 10. Centroid values from clustered second watch campaign data with K-means clustering

| Cluster	AccX	AccY	AccZ	MFX	MFY	MFZ	ZAA	XAA	YAA	GX	GY	GZ
A	-0.037	-0.335	-0.584	-0.250	0.144	1.046	-1.469	0.485	-0.002	-0.131	0.028	0.0002
B	0.108	-0.484	-0.272	0.223	0.465	0.745	1.087	0.331	0.086	-0.009	-0.012	-0.002
C	-0.707	2.230	1.560	0.557	-0.992	-1.143	-0.088	-2.572	-0.588	0.007	-0.058	-0.062

K-means clustering phase is completed, followed by hierarchical clustering phase.

4.2. Hierarchical clustering phase
Hierarchical clustering differs with K-means clustering in determine the number of clusters. K-means does the calculation based on determined number of clusters before the process started. Hierarchical clustering does the opposite. This algorithm doesn't need any number of clusters declared before to start the calculation. Hierarchical clustering will cluster the data, whether bottom-up or top-down, and then find the best number of clusters by itself.
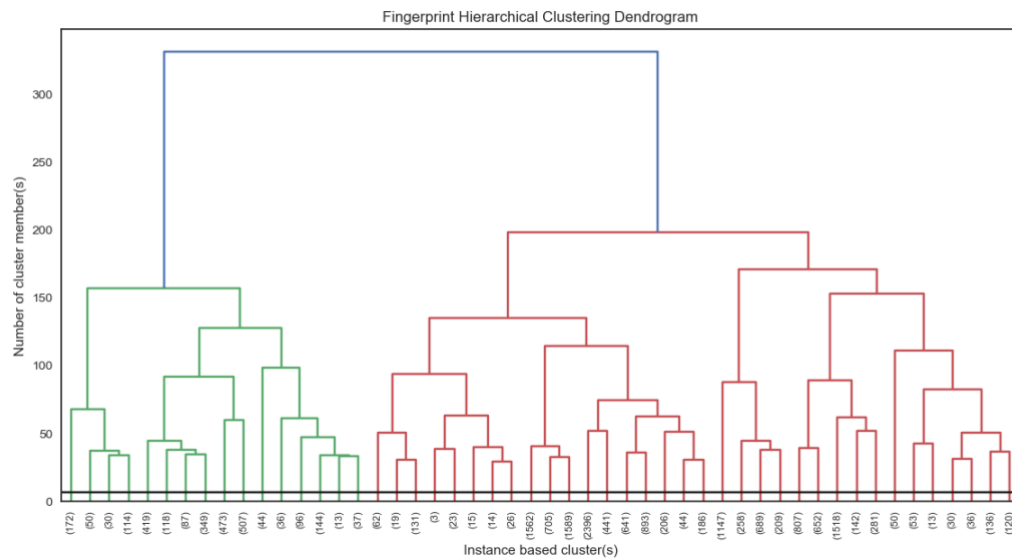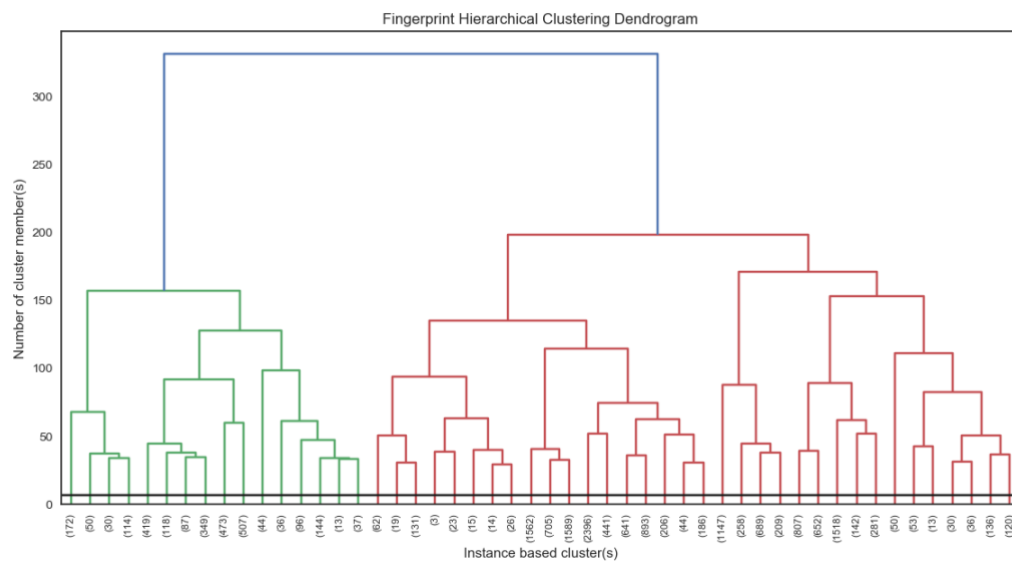
a

b
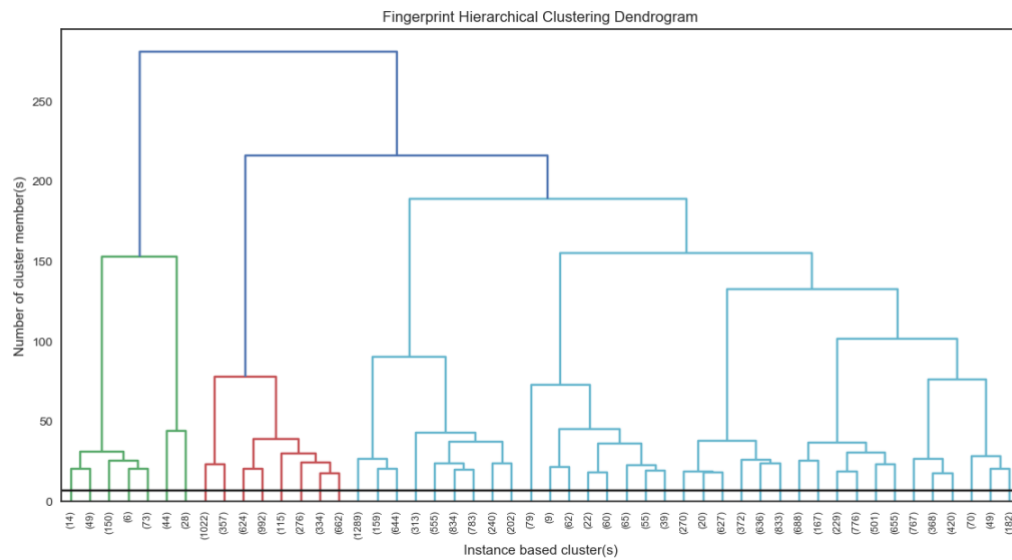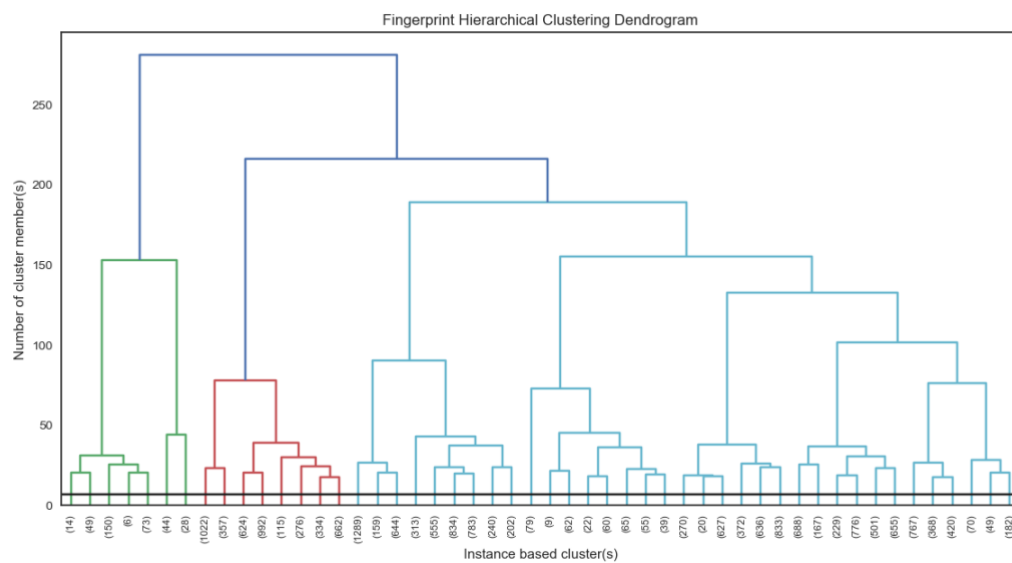
Figure 3. Dendogram of phone radiation on (a) first campaign; (b) second campaign.

a

b

Figure 4. Dendogram of watch radiation on (a) first campaign; (b) second campaign.

From figure 3(a) and 3(b), it can be concluded both phone campaigns have same number of clusters, three clusters to be exact, and the same goes with both watch campaigns which have same number of clusters too, two clusters to be exact. All phases are completed, go the next step is cluster analysis.

5. Discussion and Conclusion
Table 11 shows the summary of the experiment result.

Table 11. Experiment result.

Dataset	Number of clusters (K-means)	Number of clusters (Hierarchical)
Measure1_smartphone_sens	4	3
Measure2_smartphone_sens	5	3
Measure1_smartwatch_sens	3	2
Measure2_smartwatch_sens	3	2

From the result, we can conclude that radiation emitted from the phone and watch were being clustered to many clusters. The cluster itself will be the basis of classification which could lead to the record being classified, for example, as weak radiation, medium radiation, strong radiation, and many more. But clusterization data itself can't be used to make a label as it need to be processed further. The clustered data can be used to determine and predict the label assume the process is continued to classification stage. That conclusion gives stronger basis to continue the research.

6. References

[1]   Barsocchi P, Chessa S, Furfari F, Potorti F. Evaluating Ambient Assisted Living Solutions: The Localization Competition. IEEE Pervasive Computing. 2013;12(4):72-79.

[2]   Longhi S, Siciliano P, Germani M, Monteriù A. Ambient Assisted Living. Cham: Springer International Publishing; 2014.

[3]   Potorti F, Barsocchi P, Girolami M, Torres-Sospedra J, Montoliu R. Evaluating indoor localization solutions in large environments through competitive benchmarking: The EvAAL-ETRI competition. 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN). 2015. p. 1-10.

[4]   Han J, Owusu E, Nguyen L, Perrig A, Zhang J. Accomplice: Location inference using accelerometers on smartphones. 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS). 2012.

[5]   Barsocchi P, Crivello A, La Rosa D, Palumbo F. A multisource and multivariate dataset for indoor localization methods based on WLAN and geo-magnetic field fingerprinting. 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN). 2016.

[6]   Yulita I, Fanany M, Arymuthy A. Bi-directional Long Short-Term Memory using Quantized data of Deep Belief Networks for Sleep Stage Classification. Procedia Computer Science. 2017;116:530-538.

[7]   Yulita I, Fanany M, Arymurthy A. Fast Convolutional Method for Automatic Sleep Stage Classification. Healthcare Informatics Research. 2018;24(3):170.

[8]   Yulita I, Wasito I. gCLUPS: Graph clustering based on pairwise similarity. 2013 International Conference of Information and Communication Technology (ICoICT). 2013.

[9]   Raval U, Janni C. Implementing & Improvisation of K-means Clustering Algorithm. International Journal of Computer Science and Mobile Computing. 2016;5(5): 191-203.

[10]   Oyelade O, Oladipupo O, Obagbuwa I. Application of k Means Clustering algorithm for prediction of Students Academic Performance. arXiv preprint arXiv. 2010;:1002.2425.

[11]   Rana S, Garg R. Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute. 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT). 2016.

[12]   de Amorim R. Feature Relevance in Ward's Hierarchical Clustering Using the L p Norm. Journal of Classification. 2015;32(1):46-62.

[13]   Palumbo F, La Rosa D, Chessa S. Gp-m: Mobile middleware infrastructure for ambient assisted living. 2014 IEEE Symposium on Computers and Communication (ISCC). 2014.