MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

.....*.....**

NGO VAN BINH

RESEARCH SOLUTIONS TO IMPROVE THE EFFICIENCY OF INDOOR POSITIONING BASED ON WIRELESS SIGNAL DATA

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM

Major code: 9 48 01 04

Ha Noi – 2023

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy Science and Technology

Supervisors:

Supervisor 1: **Dr. Hoang Do Thanh Tung**
Supervisor 2: **Assoc.Prof.Dr. Nguyen Thanh Hai**

Referee **1**: …….
Referee **2**: …….
Referee **3**: …….

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at ……. hour……, date…… month…….year 2023

The dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library

INTRODUCTION

1. The ugency of the thesis

* Practical aspect: The need to build Indoor Positioning Systems (IPS) has increased significantly and attracted much attention in recent years due to its commercial value as well as its application. IPS provides a variety of indoor location-based services such as rescue, rescue, pathfinding, marketing, etc. in large space areas. With diverse service types, the revenue of the Indoor Locationbased Services (ILBS) market is increasing. According to marketsandmarkets.com1 the market revenue in 2022 is $8.7 million and with a compound annual growth rate of 22.4%, by 2027 the revenue is expected to reach $24 million. Besides, the number of people using smartphones is increasing day by day. According to the statistics of statista.com2, the number of smartphone users worldwide in 2022 is more than 6.5 billion people, estimated to be more than 6.8 billion people in 2023. Statistics have shown research About indoor location positioning is essential for developing applications that provide indoor location-based services intuitively.

* *Scientifically:* Outdoor navigation systems often use satellite signals for positioning, such as the Global Positioning System (GPS). GPS provides good positioning performance and can accurately locate objects from 1-5m. However, GPS signal cannot penetrate well in indoor environment resulting in reduced positioning accuracy, so many other wireless signals such as ultrasonic waves, ultra-wideband, Bluetooth, Zigbee and WiFi have been adopted. Research use for indoor navigation system. In these wireless standards, WiFi has lower positioning accuracy than some other technologies such as ultrasonic waves, broadband. However, WiFi-based positioning system has many advantages such as low cost, no need for additional hardware, high scalability, and can locate objects with reasonable deviation distance, and With its high ability to transfer data between devices and relatively little influence from external factors, WiFi can provide many opportunities to improve accuracy. Moreover, WiFi is becoming more and more popular, most of the current mobile devices of users such as phones, computers, smart watches are WiFi enabled and the infrastructure for using WiFi networks has also developed continuously. customary. Therefore, of these wireless standards, WiFi is the most popular and relevant wireless standard, has become one of the ideal candidates for indoor positioning and is the most widely researched technology. . Therefore, it is possible to build an indoor positioning system based on WiFi wave data (with reasonable accuracy) without additional infrastructure.

There are many techniques and methods of indoor positioning based on WiFi wave data, including: Time of Arrival (ToA), Angle of Arrival (AoA), Time difference of arrival (Time Difference). of Arrival- TDoA), Proximity and FingerPrinting. In particular, compared to other methods, the fingerPrinting method is relatively simple, easily integrated with smart devices, taking advantage of the support from the existing wireless infrastructure (WiFi Access Point, mobile phones,...) without additional hardware. The accuracy and performance of fingerPrinting are still affected by indoor obstacles, but it can still estimate the object position quite accurately with an acceptable deviation distance. Therefore, the fingerPrinting method is a more convenient method and can be applied to the indoor positioning problem based on WiFi data.

From the above reasons, the thesis chooses the research topic: "**Research solutions to improve the efficiency of indoor positioning based on wireless Signal data**". With the task of finding effective solutions to improve the performance and location accuracy of IPS by WiFi's RSS-based fingerPrinting method, contributing to building useful indoor location-based service for user.

The most significant challenge of the fingerPrinting method is the instability of RSS. The cause of the instability of RSS is the receiver itself, the transmitter and the obstructions in the house. Devices and obstructions in addition to attenuating the signal also cause multipath effects. These two factors increase the computational cost, reduce processing speed, reduce performance and especially decrease the positioning accuracy of fingerPrinting method. Although various sampling methods have been used to remove noisy RSS feeds, these values persist regardless of the collection method used. Therefore, many studies and applications have been carried out to improve the efficiency and positioning accuracy of the fingerprinting method.

The first research direction can be mentioned is the selection of APs. The AP is selected based on the RSS value. However, after selecting the APs according to their method, the studies all ignored the remaining APs. This approach can cause some APs to be "misclassified", because also due to multipath effect and signal degradation, RSS values of the same AP are received at the same location at different times. may be different. Therefore, the method of selecting the AP so as not to "miss" the RSS value is a challenge.

The research direction using clustering method has also been interested and implemented by many research groups, resulting in increased speed and accuracy of positioning. However, due to the multipath effect and signal attenuation, and according to the study of Torres-Sospedra et al., the use of the method of comparing the RSS obtained at the location to be located with the center of the clusters to determine clustering can lead to choosing the wrong cluster. Therefore, if there is a suitable cluster selection method, it is possible to estimate the location more accurately.

One of the other popular approaches that many domestic and foreign research groups focus on is the use of machine learning-based fingerPrinting. In addition to some algorithms such as PCA (Principle Component Analysis), KPCA (Kernel Principal Component Analysis) used to reduce features, reduce data dimensions, other algorithms such as KNN, SVM, RF... are used for prediction. location. Recently, the solution using the Ensemble Learning model (ELM) has also been applied. Overall, the research results show that machine learning algorithms have helped the navigation system estimate the location more accurately and can be flexibly applied to many different environments. Although the ELM model has combined many algorithms and has given better positioning efficiency than the basic models, the ELM model still has the possibility of overfitting and the operation of the ELM model can also be ignored. through the strengths of each algorithm. Therefore, building a machine learning model that can take full advantage of the algorithms, reduce the risk of overfitting, and increase the positioning quality for indoor navigation systems is still a challenge.

2. **Objectives of the thesis**

With the research task to get effective indoor positioning solutions, the thesis

sets out the research objective: how to increase the ability to determine the indoor location effectively and accurately. To achieve this goal, based on the analysis of related studies, the thesis offers two solutions:

a. The first solution: Improve the accurate location prediction of the traditional fingerPrinting method by transforming the RSS values through the Access Point (AP) selection method and the cluster selection method.
b. Second solution: Increase the efficiency and accuracy of the machine learning-based fingerPrinting method by a two-phase machine learning model, in which the training results of the previous stage are used to generate training data for the previous stage. second paragraph.

3. The main research contents of the thesis

a. Researching AP selection methods, RSS vector clustering, location clustering and cluster selection methods.
b. Studying machine learning models, with an emphasis on studying machine learning models that integrate multiple machine learning models simultaneously.
c. Build and execute a realistic indoor positioning environment on a single site. Install, test, evaluate the proposed methods of the first solution on the self-built environment.
d. Install, experiment, evaluate the proposed machine learning model in the second solution on a multi-court, multi-tier public dataset and compare it with other publications on the same dataset.

CHAPTER 1: OVERVIEW OF INDOOR LOCATION SOLUTIONS BASED ON WIRELESS DATA

1.1 WiFi-based indoor positioning technologies

GPS is the most popular and widely used outdoor navigation tool, GPS requires Line-Of-Sight (LOS) between satellites and handheld devices. However, obstructions (such as ceilings and walls) cause GPS to degrade due to signal reflection and signal degradation. This results in GPS being inefficient and almost unsuitable for indoor positioning. There are various wireless technologies used in place of GPS for indoor positioning. Among them, commonly used wireless technologies include: Radio Frequency Identification (RFID), Ultra Wide Band UWB, Bluetooth, ZigBee and WiFi. Radio-frequency identification (RFID) systems are capable of indoor positioning and tracking, but RFID implementation is difficult because it is not supported on user mobile devices. Ultra-wideband (UWB) technology is attractive because of its lack of interference, its ability to penetrate materials, and its low sensitivity to multipath effects. However, the slow progress of UWB standardization and high costs limit its use in consumer products and mobile devices. Bluetooth positioning has the advantages of simple, low energy consumption, fast connection speed, high transmission speed, stable and safe signal, but still has high positioning error due to multipath phenomenon in the network. indoor environment. Zigbee is a short-range communication protocol with low power consumption and low cost, but limited in positioning range, large error and poor anti-interference ability. Compared with other wireless technologies, WiFi-based positioning system has many advantages such as low cost, high scalability, ability to locate with reasonable error and ability to improve accuracy. WiFi networks are ubiquitous and infrastructure is constantly evolving, making it an ideal candidate for indoor

positioning and the most widely researched technology.

Therefore, in the thesis, WiFi is the wireless technology of choice for the indoor positioning problem, because it is feasible and has potential, does not require additional infrastructure.

1.2 Overview of indoor positioning methods using WiFi data

1.2.1 Methods

WiFi-based positioning methods can be divided into two categories: methods based on the spatial and temporal attributes of the received signal (Time and Space Attributes of Received Signal (TSARS), also known as methods based on the received signal (TSARS). on the range, and the positioning method is based on the received signal strength (RSS).

Range-based indoor positioning methods include Time of Arrival (ToA), Angle of Arrival (AoA) and Time Difference of Arrival (TDoA) methods. Where, ToA calculates the distance according to the Time of Arrival, TDoA measures the delay time, while AoA measures the angle of the incoming signal sent by different access points (Access Point-AP).

RSS-based positioning technology uses the strength of the received signal to determine the user's location. RSS is the actual signal power strength received at the receiver, usually measured in decibels-milliwatts (dBm) or milliWatts (mW). RSS can be used to estimate the distance between the AP and the receiver. The higher the RSS value, the smaller the distance between the receiver and the AP. There are two main methods of using RSS-based indoor positioning: Proximity and Fingerprinting.

1.2.2 Evaluation of methods

The advantages and disadvantages of methods based on the results of analysis, assessment of complexity and environmental impact are summarized in Table 1.1.

From the analysis and statistics of the advantages and disadvantages of each positioning method, it can be seen that FingerPrinting is one of the simplest, most feasible and most widely used indoor positioning methods in the world. lots of research as well as practical applications. FingerPrinting is also the method of choice for PhD student to research and develop solutions to increase the efficiency of the indoor positioning system.

Table 1.1: Summary of advantages and disadvantages of indoor positioning methods

Methods	Advantages	Disadvantages
ToA	Provides high accuracy in LoS environment; Pretty simple algorithm	Requires time synchronization between APs and receivers often require additional hardware. Positioning performance drops with complex indoor environments that do not guarantee LoS
TDoA	Provides high accuracy in LoS environment; Pretty simple algorithm	Requires time synchronization between APs often require additional hardware. Positioning performance drops with complex indoor environments that do not guarantee LoS
AoA	Provides high accuracy in LoS environment	Additional components may be required complex hardware such as directional antennas; requires relatively complex algorithms. Loss of performance in complex environments is not guaranteed LoS

Proximity	Simple algorithm does not need additional hardware requirements	Low precision, rated performance taste reduced with complex indoor environment.
FingerPrinting	No additional hardware required; little affected by the impact of the environment; acceptable accuracy; Does not require the location of the AP	There are many algorithms that use divisors number of positions. Database preparation is time-consuming and labor-intensive but may have to change as the number and location of APs change

1.3 Indoor positioning using fingerPrinting

1.3.1 Architecture of indoor location positioning system using fingerprinting method

The WiFi RSS-based fingerPrinting indoor positioning system is divided into two phases, the offline data collection phase and the online matching phase as shown in Figure 1.1. In which: Offline phase: At each predefined reference point (Reference Point-PR) on the positioning map, the strength of the received signal (RSS) of neighboring APs is collected, they form The RSS vector of the position with the elements of the vector follows the same order of the AP sequence. The RSS vectors, together with the locations stored together form the fingerPrinting database (signal map); Online phase: By comparing and matching the online RSS vector obtained at the device's location with the RSS vectors in the fingerPrinting database using a prediction algorithm, we can estimate the device's location.

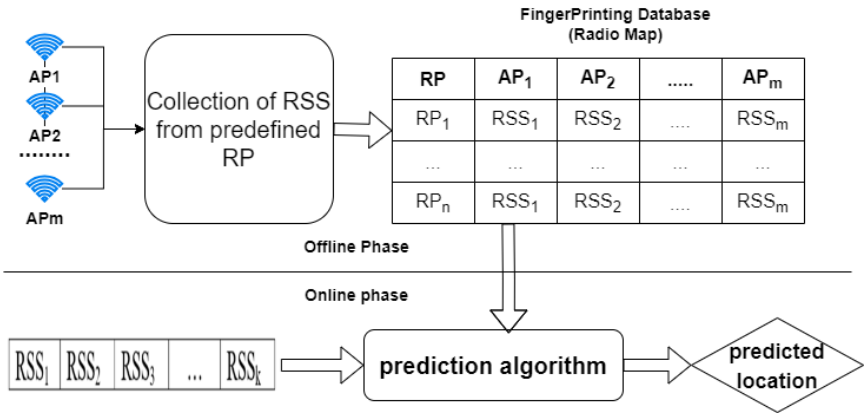


Figure 1.1: Architecture of indoor positioning system by fingerPrinting

1.3.2 Fingerprinting database

After the construction process, we get the fingerPrinting database as shown in Figure 1.1. In particular, the fingerPrinting database includes many fingerPrinting, each fingerPrinting of WiFi signal consists of three elements: location, unique address or MAC address of AP (APid) and RSS vector with components in order. of the AP sequence received at the corresponding position. For each sampling, with a total AP of m, fingerPrinting at the ith RP is defined in Equation (1.1):

$$f_i = [(ViTri_i), RSS1, RSS2, \ldots, RSS_m] \quad (1.1)$$

In which, the RSS value of the AP that is not detected at the RP will be set to the default value (usually 100). The fingerPrinting database is obtained from n structured locations in (1.2).

$$D_n(F_i) = \{f_{i1}, f_{i2}, ..., f_{ik}\} \quad (1.2)$$

1.4 Factors affecting positioning quality of indoor navigation system by fingerPrinting

Static and dynamic obstructions along with transmitter and receiver elements can degrade the signal. In addition, static obstructions (such as windows, doors, walls, objects ...) exist in the indoor space along with the movement of people, the closing and opening of doors, making the signal Signals are transmitted over different paths, causing the signal to reach the receiver at different times, resulting in possible signal overlap. This phenomenon is called the multipath effect.

Since fingerPrinting relies on RSS to estimate the user's location, multipath effects and signal degradation have significant consequences for indoor positioning, not only expensive storage costs but also computational costs The increase leads to slow processing speed, especially the decrease in the efficiency and accuracy of the navigation system. Therefore, it is very valuable to improve the quality and efficiency of RSS while increasing the accuracy and performance of the navigation system.

1.5 Methods to increase the efficiency and positioning accuracy of fingerPrinting

1.5.1 Selecting AP method

The FingerPrinting method uses all the RSS feeds from the access point to determine the location, but with too many RSS feeds, the multipath effect reduces accuracy and increases the system burden. Most solutions choose the AP based on the size of the RSS, because the AP with the strongest RSS gives high accuracy. Feng Chen et al selected the strongest AP in the online phase, and used Fisher's criterion in the offline phase. The MaxMean algorithm sorts the average RSS measurements from multiple APs and selects the strongest AP to locate. Another study divided APs into different RSS thresholds and selected the same AP with the highest threshold during the online period. The residual ranking algorithm selects less sensitive APs and discards APs that are less likely to appear in FingerPrinting. Cluster-based approach, optimal group selection based on common information among APs. Simple RSS-based AP selection method but ignores the rest of APs, but due to multipath effect, same AP at different times can have different RSS value. This means, at the time of sampling, the AP may be close but the RSS is low. Therefore, it is necessary to research a solution to select APs without "wasting" APs.

1.5.2 Clustering method

Two commonly used clustering methods are K-means and equivalence propagation clustering (APC). Swangmuang uses K-means and increases positioning speed by 50%. Seyed Alireza Razavi applies K-means and has reduced computation time. Abdullah modifies the K-means by Bregman

divergence and the resulting mean error reduction. Torres-Sospedra et al improved K-means by combining the selection of the strongest AP, better positioning was increased. Boyuan Wang combines RSS and location in K-means to improve accuracy. Andrei Cramariuc et al use K-means and APC, with APC having lower computational complexity, but not as accurate as K-means. Chen Feng et al applied APC and reduced the mean error. Zengshan Tian et al applied location-based APC clustering and the mean error was also reduced. Pejman uses fingerPrinting database clustering based on RSS and waypoints, increasing prediction performance. Jingxue Bi et al apply APC in both stages to increase accuracy. Limin Wang et al improved APC by assessing data density. Genming Ding et al used an artificial neural network with a clustered model using APC. Both studies reduced positioning time and error. Clustering methods have contributed to acceleration and improved positioning, but multipath effects and signal degradation can cause the RSS value to change at the same location at different times. Therefore, cluster selection by comparing the RSS value obtained at the online stage with the cluster center can lead to confusion about the cluster center, especially when the actual location of the object lies between two or more clusters. . In this case, if the RSS online value is changed, the distance between the RSS online value and the cluster center will also change, leading to the wrong cluster selection. Therefore, the cluster selection method needs to be improved to ensure better positioning accuracy and quality.

1.5.3 FingerPrinting method based on machine learning algorithm

FingerPrinting databases are often large with many records and data fields. To speed up processing and improve positioning, many machine learning algorithms (Machine Learning-ML) have been applied. Machine learning has the ability to learn and identify patterns in data, based on the learning process to make decisions for new data. With machine learning-based fingerPrinting, the machine learning model is trained to find the relationship between the RSS vector and the location. When applying the model to the RSS vector at the online stage, the positioning accuracy and efficiency increase significantly.

1.5.3.1. The fingerPrinting method is based on an independent machine learning model

Machine learning algorithms have made a significant contribution to solving the localization problem in home based method fingerPrinting. KNN has been used very early and has shown superior performance compared to fingerPrinting. SVM is also applied and gives the exact same positioning result as KNN. RF is used in spaces with no walls or obstructions, and has greatly improved accuracy and execution time. LR and its variants also gave good results and improved accuracy compared to fingerPrinting. In addition, algorithms such as DNN and LightGBM have also been applied and provide higher performance in positioning. Other algorithms such as LDA and NB (Naive Bayes) have also been tested and give quite good positioning results. In general, applying machine learning algorithms has improved positioning accuracy and improved system performance compared to the traditional fingerPrinting method.

Each algorithm has its own advantages and limitations, and the choice of algorithm depends on the requirements of the problem and the data. However, if only one algorithm is used in the navigation system, the capabilities of other

algorithms can be overlooked. Therefore, many research groups have used the Ensemble Learning model (ELM) to better take advantage of the algorithms and increase the positioning efficiency of the system.

1.5.3.2. The fingerPrinting method is based on associative machine learning models

Ensemble Learning Model (ELM) consists of a set of models are combined to form a stronger model. The main idea of Ensemble Learning is to combine the predictions of many different models to produce a final prediction with higher accuracy. Specifically, the combination of DNN and KNN in one study gave better results with an error of 1.39m to 1.5m. Using the Ensemble Learning (ELM) model has also yielded remarkable results, with a deviation of about 4m in 80% of the tests and an RMSE of 8.79m and 8.83m for the X and Y axes. developed an ELM model based on KNN, DNN, RF and SVM, and the "voting" results of the models predicted the position with a bias of 1.1 in 60.38% of the trials. However, although these methods have improved the accuracy and performance of the model, some challenges still exist. A common problem is the possibility of overfitting when training models on the same data set. In addition, weighting or using a mechanism ("voting") the prediction results of the baseline models can reduce the reliability of the final prediction. To solve these problems, it is necessary to build new models that are able to limit overfitting and improve efficiency through the training results of the base models.

Conclusion of Chapter 1. In Chapter 1, the thesis first presents an overview of the indoor positioning problem based on wireless wave data and the problems of the problem. Next, popular wireless technologies used in indoor positioning problems are introduced, after evaluating and comparing technologies, WiFi is the most suitable technology. The indoor positioning system based on WiFi wave data can be implemented by many different techniques and methods. Among them, the fingerPrinting method is the most appreciated due to its low cost, suitable for indoor environment, ease of implementation and acceptable accuracy. However, the fingerPrinting method faces two challenges that reduce the system's positioning accuracy and efficiency, namely multipath effect and wave signal attenuation. To increase the quality and positioning performance of the fingerPrinting method, many solutions have been proposed by many research groups.

CHƯƠNG 2: AP SELECTION METHODS AND FINGERPRINTING DATABASE CLUSTERING

2.1 Problem

In buildings and commercial centers, it has become common to equip many WiFi APs to ensure the quality of Internet access. However, increasing the number and density of APs also poses challenges for indoor positioning using the WiFi RSS-based fingerPrinting method. The first problem is that the multipath phenomenon significantly affects the positioning quality. Many studies have investigated the required number of APs and suggested how to select APs based on the value of RSS to increase the location quality. However, the effects of multipath effects and signal attenuation can change the RSS value of the same AP at the same location resulting in several APs being mistakenly rejected. Therefore, the thesis proposes a new AP selection method to reduce the possibility of mistakenly rejecting the AP and the impact of multipath effect,

thereby increasing the accuracy. The second problem is that the size of the fingerPrinting database increases with the number of APs, which increases the computation cost and reduces the positioning speed. Clustering method has been applied to solve this problem. However, there is still the problem of cluster selection in the online phase and in the experimental results of the thesis's AP selection proposal, the phenomenon of some location prediction results being "jumped" too far. Therefore, the thesis proposes a new cluster selection method to overcome the location deviation and improve the quality of the location as well as to overcome the problem of the proposed AP selection.

2.2 Proposed method to select AP

The proposed AP selection method is based on two factors: (1) Select the AP with the most feasible RSS value for the positioning process. (2) Use APs with the strongest RSS value for better accuracy. However, multipath effects and signal degradation make it difficult to distinguish RSS values and can lead to incorrect selection. At the same time, the AP selection method only focuses on the N APs with the highest RSS values, ignoring other RSS values and possibly causing loss of important information. Therefore, the thesis proposes a new AP selection method at the online stage. Figure 2.1 shows the implementation flowchart of the proposed AP selection method. The implementation steps are shown in Algorithm 2.1. The algorithmic complexity of the method will increase rapidly with the value of k because the number of triangles generated is C(k, 3) = k! / (3! * (k - 3)!). Therefore, NCS recommends using a minimum of 3 RSS numbers and a maximum of 5.

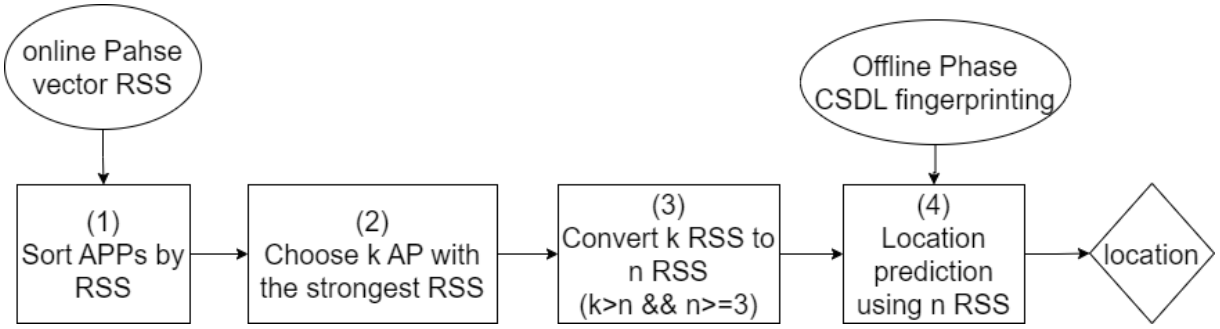


Figure 2.1. Flowchart of the proposed AP selection method

Algorithm 2.1: : Algorithm to locate by APs with the strongest RSS.

Input: $R \rightarrow \{RSS1, RSS2, ..., RSS_m\}$ (m RSS values obtained from m APs at unknown location)
Output: V : Predicted position.
begin
Step 1: Choose the most strongest RSS
Sort R in descending order;
$R_k \rightarrow \{RSS1, RSS2, ..., RSS_m\}$; (k largest RSS value from R)
Bước 2: Convert the Rk to Rn contain new RSS
Initialize n is the number of RSS needed to predict the location.
while $k >= n$ **do**
S_t= set of t triangles generated from k RSS in R_k;
$P \rightarrow 0/$; (set of the centroids of the triangle)
for i = 1 to t **do**
P:$P \cup$ he ith triangle's centroid in S_t
end
Sort P_t values in descending order

k'= k-1

$R_k' \rightarrow P_t$; (k' first element in P_t)

$R_k \rightarrow R_k'$

end

Bước 4: Calculate the location to be located.

Determine the location to be located with the new RSS set in R_k ; (k=n)

$V \rightarrow$ Predicted position;

Return V ;

end

2.3 Proposing cluster selection method

In this part, the thesis proposes a cluster selection method, which combines the cluster selection method by online RSS with the KNN algorithm. The operation flow chart of the method is shown in Figure 2.2. In the steps of implementing cluster selection method, the replacement of out-of-cluster locations with cluster neighbors aims to pull k locations closer together, then the ability to predict locations can be more accurate. as remote locations can cause the predicted location to shift away. Besides, replacing the location is essentially changing the RSS value, which can also limit the impact of multipath effects and signal degradation. The implementation process of the proposed method is shown in Algorithm 2.2.

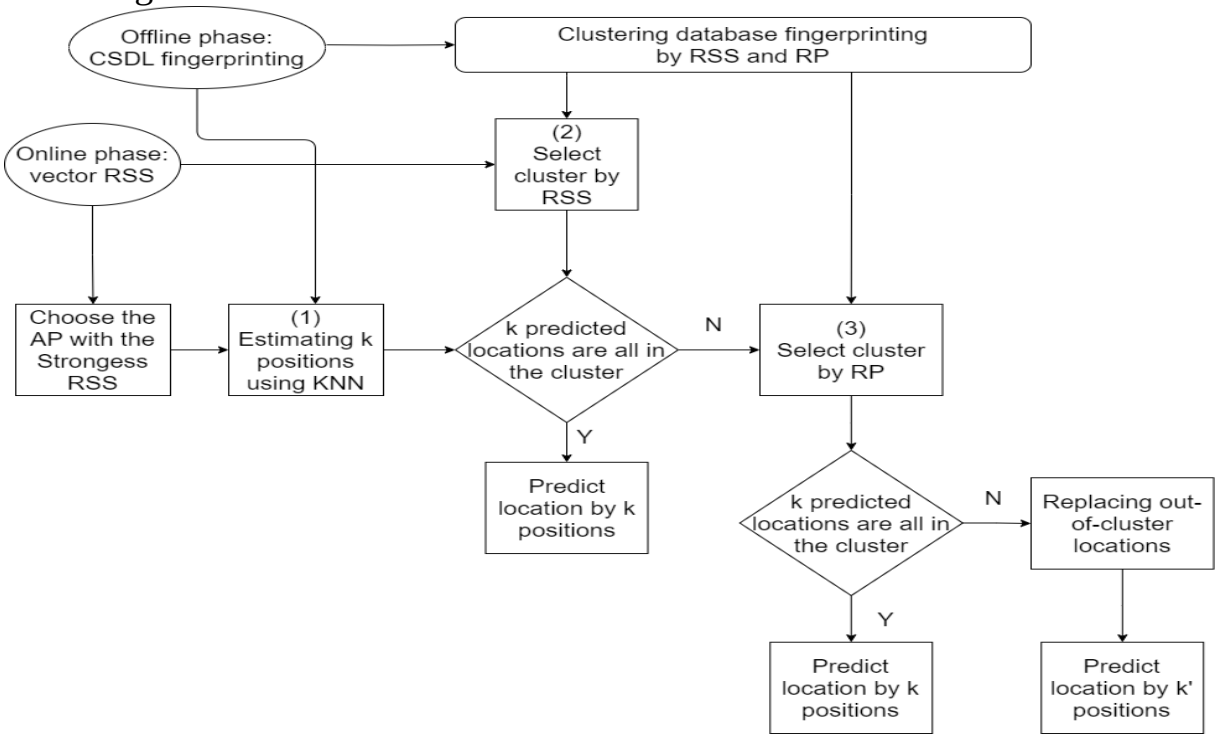


Figure 2.2: Flowchart of the proposed cluster selection method

Algorithm 2.2: Algorithm cluster selection.

Input: C_n=($C1$,$C2$,...,C_n); n clusters have been pre-generated in the offline phase

$R_m \rightarrow \{RSS1, RSS2, ..., RSS_m\}$ m RSS value obtained from unknown location

Output: V : Positioning position.

begin

Step 1: Calculate k locations and select clusters

$P_k \rightarrow \{P1, P2, ..., P_k\}$ k "neighborhood" positions from KNN equal m' RSS selected from m RSS;

Cluster selection using RSS in R_m

Step 2: Check k locations present in cluster

if (k positions in the cluster) **then**

$V \rightarrow$ Location predicted by list of positions of P_k
Return V ;
end
Step 3: Select cluster by location
Select cluster by location equal to k positions of P_k
if (k positions in the cluster) **then**
$V \rightarrow$ Location predicted by list of positions of P_k
Return V ;
end
Bước 4: Find the cluster containing the most positions in P_k and replace the position
max=0; $C_{max} \rightarrow 0$/
for i = 1 to n **do**
temp= number of positions of P_k in C_i;
if max<temp **then**
max=temp;
$Cmax \rightarrow Ci$;
end
end
Replace positions not in C_{max} with locations adjacent to those of P_k present in C_{max}
P_k' : new location set
Bước 5: Locate using a list of positions of P_k'
$V \rightarrow$ *Predicted position;*
Return V ;
end

2.4 Results and evaluation of AP . selection method

2.4.1 Experimental content and scenarios.

The thesis conducts experiments and compares two AP selection methods: AP selection method based on the largest RSS value and AP selection method proposed in the thesis. The AP selection method based on the largest RSS value will select the n largest RSS values, while the proposed AP selection method will select the m strongest RSS values (where m > n) and convert it to n. new RSS value. The value of n in the test is 3. NCS and team have conducted experimental scenarios based on the daily movement of users, there are 5 migration scenarios shown in Figure 2.3, including: straight sideways, walking straight along, take a 90-degree turn to the right, take a 90-degree turn to the left, go diagonally. A total of 250 samples were recorded for all 5 migration scenarios.

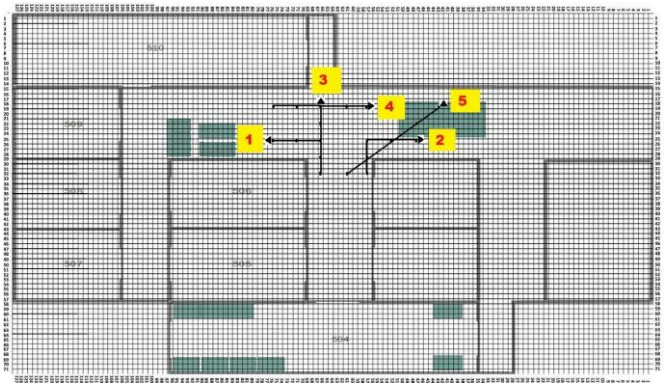


Figure 2.3: Proposed test scenario to choose AP

2.4.2 Experimental results and evaluation

Experimental results of the methods are conducted according to each migration scenario. A total of 250 experiments were performed. The following is the summary and evaluation results.

The experimental results of the two methods are evaluated based on the mean position deviation across the scenarios. Table 2.1 shows the mean position deviation of the AP selection method based on the strongest RSS value, while Table 2.2 shows the mean position deviation of the proposed AP selection method. The results show that the average position deviation of the two methods across all scenarios is 3.23m and 2.46m, respectively. This shows that the proposed AP selection method reduces the average deviation by about 24% compared to the AP selection method based on the strongest RSS value.

Table 2.1: Average position error of the method of selecting the AP with the strongest RSS

Scenario	Error (X)	Error (Y)	Average position error (m)
1	9.64	7.93	2.98
2	10.04	10.73	3.24
3	7.33	12.59	2.92
4	15.82	8.59	4.26
5	8.44	10.20	2.77
		Average position error	**3.23**

Table 2.2: Average position error of the proposed APP selection method

Scenario	Error (X)	Error (Y)	Average position error (m)
1	6.27	10.19	2.53
2	4.81	7.80	1.92
3	5.46	12.50	2.64
4	7.33	16.16	3.32
5	5.60	6.84	1.87
		Average position error	**2.46**

The experimental results along with the evaluation of the results between the two methods of selecting APs based on the strongest RSS value and the method of selecting APs based on RSS value variations have proved the feasibility of the proposed method in this study. thesis, and the ability to improve the positioning quality of the fingerPrinting method. However, in the experimental process, the proposed method still has some cases where the predicted position has a deviation greater than 4m compared to the actual position as shown in Table 2.3. Therefore, in order to solve this problem and improve the accuracy of the positioning process, the thesis has studied the clustering method and proposed a corresponding clustering method. Hopefully, this method will solve the problem of large deviation in experimental results and improve the accuracy of the positioning process.

Table 2.3: Statistics on the number of position errors of the proposed AP selection method

Scenario	*Errorr*			
	>=4m	>=5m	>=6m	>=7m
1	2	2	0	0
2	0	0	0	0
3	3	1	0	0

2.5 Results and evaluation of cluster selection method.

In the first phase, the thesis tests both k-means and APC clustering methods (equivalent spread clustering) to select a clustering method suitable for the built environment. Based on the experimental results, the thesis chooses the APC method as the clustering method for the next experiments.

2.5.1 Experimental content and scenario

The method was tested in two different areas on the map. The areas and direction of movement are shown in Figure 2.4. The reason for the division into two areas is because the location map is uneven and the AP distribution is also uneven, which leads to different RSS quality in the regions. The input of the KNN algorithm is still the proposed AP selection method with the selected number of RSS as 4RSS.

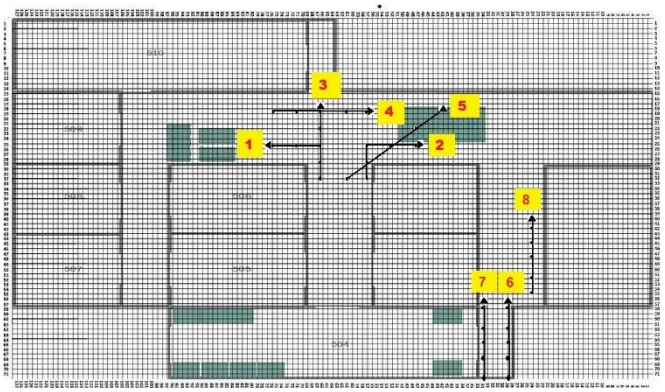


Figure 2.4. Proposed test scenario to select cluster

2.5.2 Experimental results and evaluation.

Table 2.4 shows the results of locating region 1 with scenarios from 1 to 5. Table 2.5 shows the results of region 2 of scenarios 6 to 8. Experimental results on the two regions give very different results, at zone 1 with scenarios from 1 to 5, the average deviation between predicted and actual location is 4.08m, but with zone 2 from scenario 6 to 8 the average deviation decreases by nearly 2m to 2.18m . With the indoor positioning problem, the difference of 2m is not small. This difference is explained by the uneven distribution on the map both in terms of the map and the APs (notice that partition 2 is placed by the group with 6 more fixed APs). Comparing the results with the proposed AP selection, the location quality when using clustering in region one with scenarios from 1 to 5 is reduced, the average deviation when not applying clustering is 2.46m, after applying clustering increased to 4.08m. Partition 2, with scenarios 6 to 8 looks better with an average error of 2.18m. However, because the model of positioning by APs with the strongest RSS is not tested on this partition, there is no basis for comparison.

There are many reasons that lead to the proposed method failing to meet expectations, including the map is not large enough, the data collection locations only focus on the corridors leading to uneven distribution, the number of APs as

well as the number of APs. can cause clustering, cluster selection is not as expected.

Table 2.4: Result zone 1, scenarios 1 to 5

Scenario	Error (X)	Error (Y)	Average position error (m)
1	2.58	3.14	4.27
2	1.58	2.53	3.21
3	2.27	4.18	5.10
4	2.29	3.98	4.97
5	1.69	1.90	2.86
Average position error			**4.08**

Table 2.5: Result zone 2, scenarios 6 to 8

Scenario	Error (X)	Error (Y)	Average position error (m)
6	1.73	0.51	1.93
7	1.59	0.44	1.68
8	1.84	1.68	2.92
Average position error			2.18

Conclusion of Chapter 2, In Chapter 2, the thesis presents two methods of data processing in the online phase to overcome the impact of multipath effect, signal degradation on RSS to increase positioning accuracy. The methods have been tested in the real environment and elaborately built by NCS and the research team. Among the two proposed methods, the results of AP selection method show the feasibility of the method. The clustering method has not achieved the expected results, but it helps to further confirm that the lack of data, the uneven distribution of RPs and APs is the cause of the reduction in location quality and is detrimental to the method. clustering.

CHƯƠNG 3: TWO-PHASE MACHINE MODEL

3.1 Problem

Each machine learning algorithm carries its own advantages over the others. Therefore, combining different machine learning algorithms can create a comprehensive solution for a particular application. By merging information from different machine learning algorithms, Convolutional Machine Learning Models (ELMs) can improve overall system accuracy and performance compared to models of individual algorithms. . ELM modeling focuses on combining the predictions of individual models to produce the final prediction. While each submodule in ELM may have its own tendency to overfit the data. When submodules have this tendency, the association model can be affected and inherit these undesirable characteristics. This leads to the combined model also overfitting the training data and making it difficult to make good predictions on the new data.

In this chapter, the thesis proposes a two-phase machine learning model. Instead of aggregating the predictions of the individual models to generate the final prediction like ELM, the two-phase machine learning model merges the training results from the individual models in the first phase, taking advantage

of the use the diversity and difference between models to generate training data for the next stage. The two-phase model is capable of providing continuous training and enhancing the efficiency and accuracy of location prediction. In addition, the use of training data generated from various models in stage one reduces the likelihood of overall model overfitting.

3.2 Proposed model

In this part, the thesis proposes a two-phase training model with the aim of increasing the diversity and accuracy of the training data for the second-phase model. The two-phase model training method takes advantage of the diversity of the models in the first stage and combines their results to generate diverse training data and provide more accurate predictions for the model. phase two. This reduces the possibility of overfitting and provides a model with better predictability and generalization over new data.

The proposed model of the thesis is shown in Figure 3.1. The two-phase model training process has been shown in Figure 3.2, where Yˆ1 Yˆ2, ... and Yˆn are the prediction results of n models in the first stage, these results will be used together. with the testing dataset to generate the training data set for the algorithm in the next stage. Yˆf is the end result of the second stage. The detailed training process of the model is presented in Algorithm 3.1 with computational complexity O (Max (‖Di‖) ∗ m ∗ n).

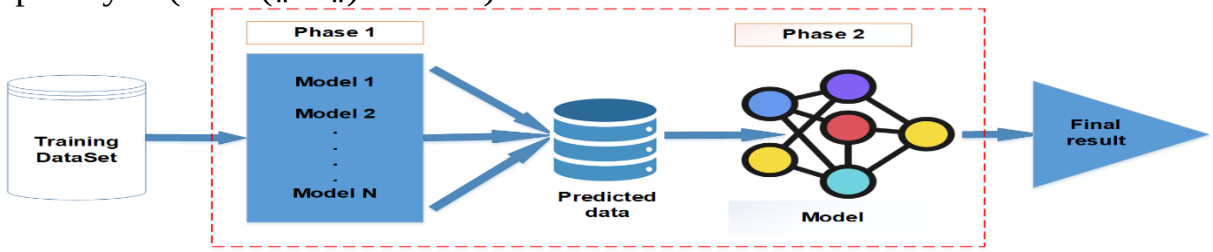


Figure 3.1: Two-phase training model

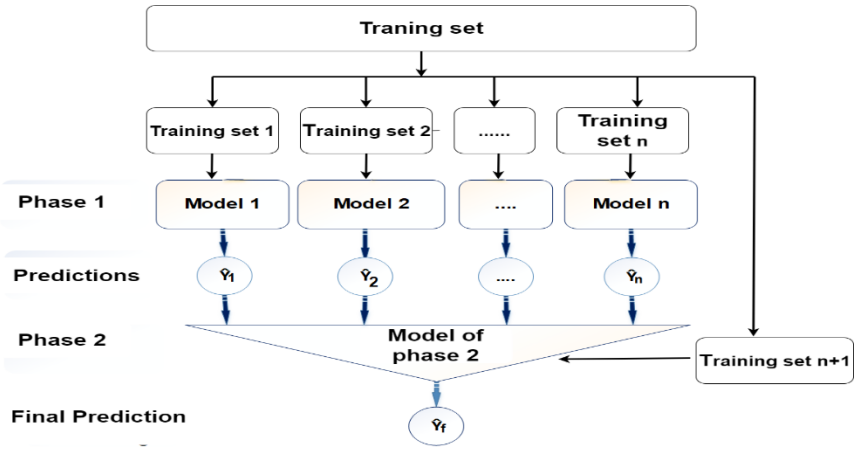


Figure 3.2: model training process

Algorithm 3.1: Two-phase model training algorithm

Input: $D \rightarrow \{x_i, y_i\}^m, x_i \subset X, y_i \subset y$ Where X is the set of features, y is the set of labels, m is the number of rows in the data set.

Dữ liệu ra: $\hat{Y}f$

begin

Step 1:

Initialize$\{M1, M2, ..., M_n\}$; n machine learning algorithms for the first phase

Divide D into subsets $\{D1, D2, ..., D_n, D_{n+1}\}$; n+1 subset of D

$D \rightarrow 0/$; Training data set of the second phase

Step 2: Train using the algorithms of the first phase

for $i = 1$ to n **do**

$X_i^{train}, y_i^{train}, X_i^{test}, y_i^{test} \rightarrow D_i$; Divide D_i nto training and test sets

$Model_0 \rightarrow train\ M_i, X_i^{train}, y_i^{train}$; M_i Model

$\hat{Y}i \rightarrow Model_0(X_i^{test})$; rediction result of $Model_0$

$D_i' \rightarrow X^{test}, \hat{Y}$; Combined data for stage two

$D' \leftarrow D' \cup D'_i$ **end**

Step 3: Training by algorithm of the second stage Initialization:: $M_{Combine}$;

$Model1 \rightarrow train(M_{Combine}, D')$; Train the model in the second phase

$\hat{Y}f \rightarrow Model1\ (D_{n+1})$; prediction results $Model1$

end

3.3 Experimental environment and positioning problem

3.3.1 Experimental data set

The two-phase machine learning model is tested on the UJIIndoorLoc dataset, which is a multi-building, multi-storey dataset used by many research groups and is suitable for the problem in chapter 3 of the thesis. The UJIIndoorLoc dataset was made by a research team from the Jaume I University in Spain. The indoor navigation system of this University is built on 3 buildings, each building has 4 or 5 floors, total area 108,703m^2. UJIIndoorLoc has a total of 21,049 samples, of which 19,938 are for training dataset and 1,111 samples for validation Dataset**.**

3.3.2 Positioning problem

The UJIIndoorLoc dataset represents a multi-building, multi-story indoor positioning environment. Therefore, the indoor positioning problem is solved by the proposed thesis model which is stated as follows: For an indoor positioning system consisting of B buildings, each building consists of F floors. In each floor are installed many APs. Let api be the RSSI value received from APi at a sampling point in building Bi and at layer Fj. If the total number of APs present in all buildings is N, then each sampling we get a feature vector like Equation (3.1).

$$f_i = (ap_1, ap_2, \ldots, ap_i, \ldots, ap_N) \tag{3.1}$$

where api = −104 to 0 and api = 100 if no signal. The feature vector fi has a corresponding label of latitude and longitude (denoted by xi and yi), a building specifying bt, and a floor ft defined. After sampling at all the reference points we have a database D containing feature vectors along with their corresponding labels as Equation (3.2).

$$\mathscr{D} = \begin{bmatrix} (a_1, x_1, y_1, b_{t1}, f_{t1}) \\ \cdots\cdots\cdots \\ (a_i, x_i, y_i, b_{ti}, f_{ti}) \\ \cdots\cdots\cdots \\ (a_N, x_N, y_N, b_{tN}, f_{tN}) \end{bmatrix} \tag{3.2}$$

For training, we know the intensity value of N RSS and the corresponding label, for example (a_1, x_1, y_1, b_{t1}, f_{t1}). For forecasting, we know the RSS values for (a2), and the corresponding label estimate is (x_2, y_2, b_{t2}, f_{t2}). Thus we have the data set D = {X, Y } , where the set X = [(f_1, f_2, ..., f_N)] is the set of features and Y = [(x_1, y_1, b_{t1}, f_{t1}),...,(x_N, y_N, b_{tN}, f_{tN})] is the set of corresponding labels. In which, the problem needs to determine which building the user/device is in, which floor (building-floor) based on the building labels Bi and floor Fj and in which position based on the longitude and longitude labels and

latitude. In the UJIIndoorLoc dataset, buildings Bi and Fj floors contain discrete values, and longitude and latitude (xi,yi) contain continuous values. Therefore, based on the data properties of the labels, the thesis builds two models: a classification model that implements the building-floor prediction problem and a regression model that implements the location estimation problem.

3.4. The two-phase classification model predicts the funeral

3.4.1 Building and proposing a two-phase classification model to predict the funeral hall

3.4.1.1. Model building

Based on the results of the study of machine learning algorithms in Chapter 1, NCS has selected a number of classification algorithms to choose the best algorithms for the first stage of the model. The algorithms include LR, LDA, KNN, CART, GB and SVM and the operating procedure is shown in Figure 3.3.

The performance of the aggregated independent models is more clearly demonstrated through the macro averages indicator. Table 3.1 shows the macro averages. The indexes of the SVM, KNN and LR models are all higher than the remaining models. The index of LR is only slightly higher than that of CART, but according to existing studies, LR has many advantages over CART and to reduce the load on the system, the thesis only chooses the LR algorithm.

The correct floor-to-floor prediction ability of the models is shown in Table 3.2. Again, the SVM, KNN and LR models have a better ability to predict the correct floor than the CART, LDA and NB models. Summarize the performance comparison results and the correct prediction results of the 3rd-floor building algorithms LR, KNN and SVM are selected for the first stage of the model. In the second stage, the NCS chooses the Logistic Regression (LR) algorithm. Based on these results, the two-phase classification model for building-floor prediction is proposed by the thesis in the next section.

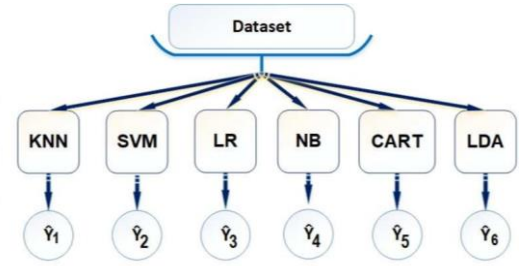


Figure 3.3: Execution of building-independent classifier models

Table 3.1: Summarizing the performance of independent models predicting floors using Macro averages

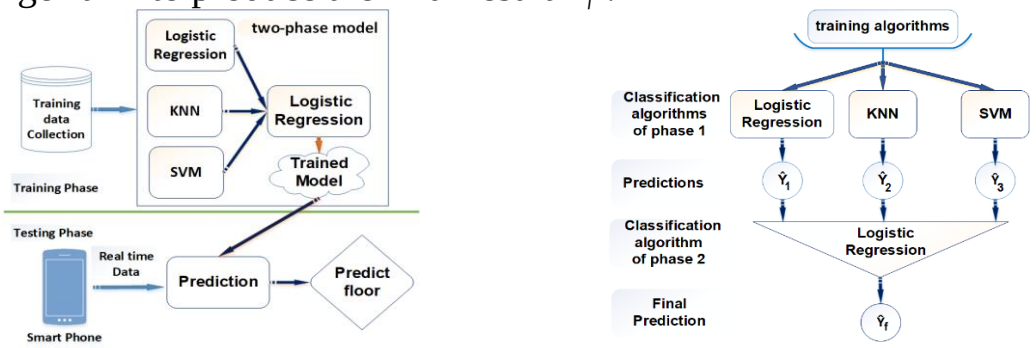
Macro averages	**SVM**	**KNN**	**LR**	**CART**	**LDA**	**NB**
Precision	**98.43**	**97.71**	**96.62**	96.50	94.42	63.70
Recall	**98.47**	**97.98**	**96.69**	96.71	94.26	55.37
F1 score	**98.45**	**97.83**	**96.65**	96.60	94.33	47.42

Table 3.2: Correct building-floor prediction results and execution time of independent models

	SVM	**KNN**	**LR**	**CART**	**LDA**	**NB**
Accuracy	**98.57**	**97.93**	**96.86**	96.76	94.66	49.09
Time (s)	7.95	0.04	3.19	0.47	1.21	0.67

3.4.1.2. Proposing a two-phase classification model for building-floor prediction

The two-phase classification model predicts the building-story along with its operation. It is shown in Figure 3.4. Where Figure 3.4a shows a two-phase model. Figure 3.4b shows the execution between two phases of the model, where $\hat{Y}_1$ $\hat{Y}_2$ and $\hat{Y}_3$ are the prediction results of the first stage, this result set is combined with the testing dataset to generate the training data. Train the LR algorithm to produce the final result $\hat{Y}_f$.



(a)Model structure (b) Model training process

Figure 3.4. Two-phase classification model for building-floor prediction

3.4.2 Experimental results and evaluation of the two-phase classification model for building-floor prediction

The performance and correct prediction results of the proposed model are clearly shown in Table 3.3. The parameters in Table 3.3 represent the performance and accuracy results. These evaluation indicators show that the proposed model predicts the location by layer with high efficiency and can correctly predict the floor with the rate of 98.73%.

The two-phase classifier model predicts the building-floor with high efficiency and high correct rate prediction. However, to assess real improvement, it is necessary to compare the results with independent models.

Table 3.3: Performance and correct prediction results of the proposed building-floor prediction model

	Macro avg Precision	Macro avg Recall	Macro avg F1-Score	Accuracy	Time(s)
proposed model	98.71	98.61	98.66	98.73	99.31

Table 3.4 shows the comparison of the correct building-to-floor prediction performance and results of the building-to-floor prediction model with the independent models. The results show that, in terms of performance, all the Precision, Recall, and F1-Score indexes of the proposed model are slightly better than the independent models. This indicates that the continuous training approach of machine learning models, in which the former provides data for the latter, has been successful and feasible for the tangent prediction problem.

Table 3.4: Comparison of performance and prediction results of the proposed model and the independent building-floor prediction models

	precision	recall	f1-score	accuracy

LR	96.62%	96.69%	96.65%	96.86%
KNN	97.71%	97.98%	97.83%	97.93%
SVM	98.43%	98.47%	98.45%	98.57%
proposed model	**98.71%**	**98.61%**	**98.66%**	**98.73%**

3.5. The two-phase regression model estimates the position

3.5.1 Building and proposing a two-phase regression model for location estimation

3.5.1.1. Building and proposing a regression model to estimate longitude

Algorithms used to select the best algorithms for the first stage of the longitude estimation regression model include SVM, ExtraTree, GB, KNN, RF and LightGBM regression algorithms as shown in Figure 3.5.

The performance results of the models are shown in Table 3.5. With this result, in the longitude estimation two-phase regression model, NCS chooses ExtraTree, KNN, RF and LightGBM regression algorithm for the first stage and Linear Regression algorithm for the second period.

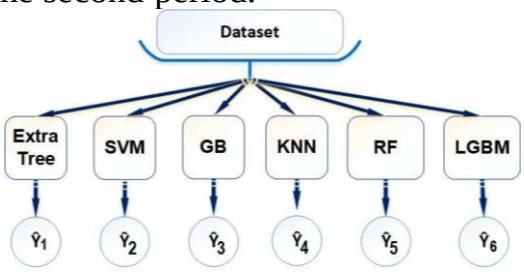
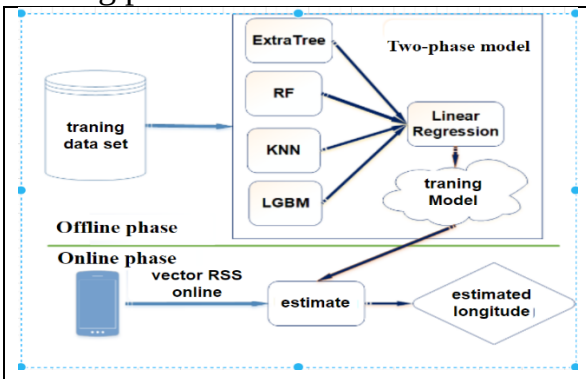
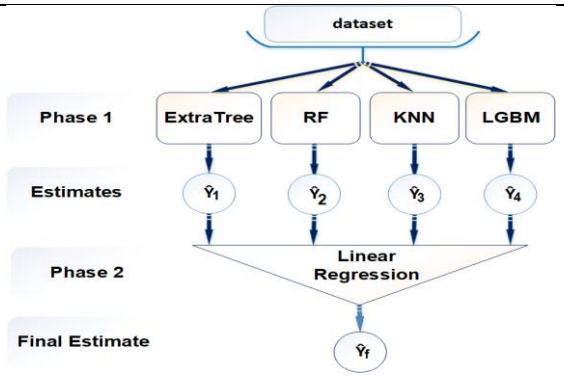


Figure 3.5: Process of implementing independent regression models for longitude estimation

The two-phase regression model for longitude estimation proposed by the thesis is shown in Figure 3.6. In which 3.6a represents the model and 3.6b represents the training process of the model.



(a) Model structure



(b) Model training process

Figure 3.6. Two-phase regression model for longitude estimation

3.5.1.2. Building and proposing a regression model to estimate latitude

Similar to when building a longitude estimation regression model, the thesis also tests independent models with regression algorithms such as the longitude estimation model building part and the algorithms are still the same as in the previous section. Figure 3.5. The test results of independent regression models are shown in Table 3.6.

Table 3.6: Performance and bias of independent regression models estimating latitude

	SVM Regressor	**ExtraTree Regressor**	GB Regressor	**KNN Regressor**	**RF Regressor**	**LightGBM Regressor**
R2-Score(%)	96.1	**98.6**	95.5	**99.3**	**99.4**	**98.8**
MSE(m)	175.2	**54.4**	200.5	**31.03**	**24.8**	**52.2**
MAE(m)	8.32	**2.75**	10.50	**2.55**	**2.18**	**4.61**
Time(s)	66.35	**0.38**	9.5	**0.027**	**37.8**	**0.32**

According to this result, the thesis chooses ExtraTree, KNN, RF and LightGBM algorithms for the first stage and Linear Regression algorithm for the second stage. It is easy to see that the latitude estimation regression model has the same structure as the longitude estimation regression model shown in Figure 3.6.

3.5.2 Results and evaluation of the two-phase regression model for position estimation

3.5.2.1. Results and evaluation of two-phase regression model longitude estimation

The performance and estimation results of the longitude estimation model are shown in the table 3.7. With this result, the R2-score value of 99.621% indicates that the model has successfully captured 99.621% of the variability in the target variable (longitude) using the features (RSS vector). This shows that the model fits the data well and can make accurate predictions about the new data. The MAE index value of 2.7m indicates that, on average, the model's predictions deviate by about 2.7m from the true longitude value.

Table 3.7: Performance and estimation results of the longitude estimation regression model

	R2-Score(%)	**MSE(m)**	**MAE(m)**	**Time(s)**
proposed model	99.621	59.32	2.70	165.00

Table 3.8 compares the performance and the difference between estimated longitude and real longitude of the proposed longitude estimation model with independent models. The performance shown in Table 3.8 shows that the proposed model fits the data better, which means higher performance. The longitude estimation of the proposed model is also less wrong than the independent models, 2.7m compared to the nearest value of 2.73m and the furthest 6m.

Table 3.8: Comparison of performance and estimation results of the proposed model and independent models for longitude estimation

	R2 Score	**MSE (m)**	**MAE (m)**
ExtraTree	99.30%	109.44	3.62
KNN	99.49%	79.39	3.26
RF	99.61%	61.59	2.73
LightGBM	99.28%	112.47	6
proposed model	**99.62%**	**59.32**	**2.7**

3.5.2.2. Results and evaluation of the two-phaseregression model of latitude estimation

Table 3.9 shows the performance and latitude estimation results of the model. This result is somewhat better than the longitude estimation model. In which, the figure of 99.52% of the R2-score represents that the model fits the data well and can make accurate predictions about the new data. With a MAE index of 1.95m, it shows that the estimated latitude deviates from the true latitude of 1.95m, which is smaller when estimating longitude.

The performance and latitude estimation results of the proposed model

compared with the independent models are shown in Table 3.10. Similar to longitude estimation, the proposed model for estimating latitude also has higher performance indicators and a lower deviation between estimated and true latitude than the independent models.

Table 3.9: Performance and estimated results of the latitudinal regression model

	R2-Score(%)	MSE(m)	MAE(m)	Time(s)
proposed model	99.52	21.66	1.95	170.82

Table 3.10: Comparison of performance and estimated results of the proposed model and the independent model for latitude estimation

	R2 Score	MSE(m)	MAE(m)
ExtraTree	98.68%	59.43	2.75
KNN	99.31%	31.04	2.55
RF	99.45%	24.81	2.18
LightGBM	98.84%	52.27	4.62
proposed model	**99.52%**	**21.66**	**1.95**

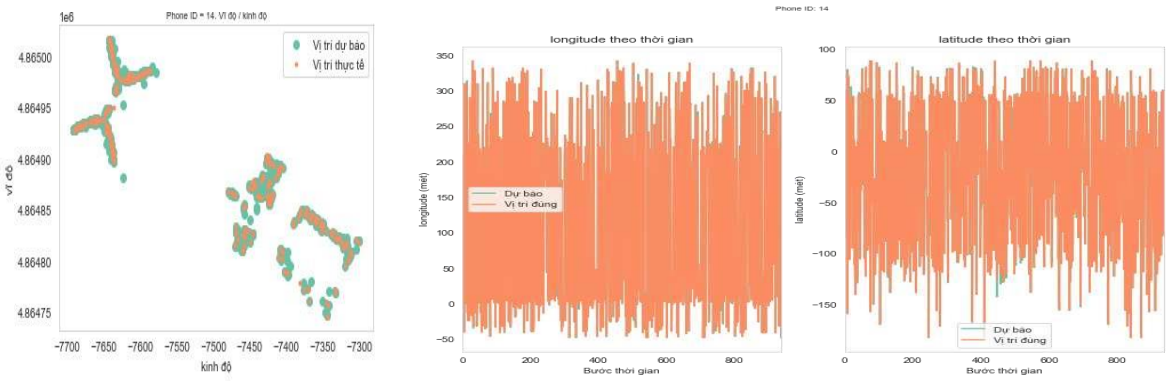
3.5.2.3. Summary of location prediction results

With the MAE of longitude of 2.7m and latitude of 1.95, the average deviation of the estimated location from the actual location calculated by Euclidean Formula (3.3) is 3.3m.

$$MAE_{ViTri} = \sqrt{[(X_2 - X_1)^2 + (Y_2 - Y_1)^2]} \qquad (3.3)$$

3.6 Results and evaluation of the proposed model with real data

In this section, the proposed model is evaluated by the validation dataset. Thesis tested on all phones. However, due to the large number of phones, NCS only selected representative results as phoneID=14. Figure 3.7 shows the deviation of the estimated location from the actual location taken with a phone with PhoneID=14. In which, figure 3.7a shows the position deviation according to the two-dimensional spatial coordinate cross-section of longitude and latitude, on a plane with many overlapping points. Figure 3.7b shows the location deviation results in longitude and latitude according to the movement of the user (phone). Green represents the estimated location. Orange represents the actual location. The colored positions represent a match between the estimated and actual locations that are virtually identical. These images once again confirm the accuracy of the model proposed by the thesis.



(a) cross section phoneID=14 **(b)** Move of phoneID=14

Figure 3.7: Accuracy testing

3.7 Compare the results of the proposed model with the models of other studies

In this section, the thesis compares the experimental results of the proposed model with the results of other studies on the same UJIIndoorLoc dataset.

Table 3.11 shows the research results and the model results of the thesis. Including the results of other studies. According to this result, in terms of stratification prediction, the proposed model ranks second, inferior to that of Lu Yin et al. 0.59%; Lu Yin et al. do not use the EML model. In the group's study, the denoising autoencoder was the main factor to improve the positioning quality. This encoder works to extract key features from sparse RSS data and reduce the effect of noise and outliers before feeding the data into the LightGBM algorithm.

Table 3.11: Comparison of the proposed model results with other studies

Research	Building – Floor prediction	Location Estimation Error (MAE) (m)
Beenish Ayesha Akram et al ;	-	6,46
Shivam Wadhwa et al ;	97,95%	7,93
Gan et al;	95,41%	6,4
Lu Yin et al;	99,32%	96.73%
Charoenruengkit et al;	97%	5,65
proposed model	**98,73%**	**3,3**

Conclusion of Chapter 2. The two-stage model opens up the opportunity to combine training results from multiple individual models, taking advantage of their diversity and differences. This is beneficial in improving the predictability and accuracy of the overall model. Thereby, the model provides a continuous and reinforcement training method, which improves the efficiency and accuracy of location estimation. This has been demonstrated through building-to-floor prediction models and location estimation using latitude and longitude. The experimental results show that the proposed two-stage model is an effective machine learning method.

CONCLUDE

The study " **Research solutions to improve the efficiency of indoor positioning based on wireless Signal data** " is an approach to indoor positioning problem by fingerPrinting method using WiFi wave strength which is highly practical, because services based on on the position of not only growing globally but also gradually developing in Vietnam. Although many studies and solutions have been published recently, there are still many challenges that have not been solved or can be further improved because different indoor environments are different and complex. different, even in the same environment, at different times can have different complexity, due to the change of obstacles. The problem of indoor positioning by fingerPrinting using RSS of WiFi waves still faces two main challenges: multipath effect and signal attenuation.

To solve this problem, the thesis conducts an overview research on technologies, techniques, construction models and solves the problems of indoor positioning problem by fingerPrinting is based on WiFi's RSS feed. From theoretical as well as experimental studies, the thesis has proposed 02 improvements to the traditional fingerPrinting method including: Transforming RSS vector values online with the goal of reducing the impact of the environment on the RSS value. by the AP selection method. Changing the way

to select clusters and handle out-of-cluster locations of the APC clustering method, the changes aim to select the most feasible cluster and ensure the convergence of locations in the cluster, thereby improving the accuracy of the cluster. precise positioning. These two proposals are tested on the environment built by the researcher and the research team to ensure the requirements of the indoor positioning environment. As a result, the first improvement increased the accuracy by 24%, the second improvement did not increase the positioning accuracy because of the uneven distribution of positions and APs as well as the small number of samples in the fingerPrinting database.

In the next section, the thesis applies machine learning to the fingerPrinting method and proposes a two-stage machine learning model to increase the quality and performance of positioning. The proposed model is tested on a multi-court, multi-layer dataset with a large area and number of samples. As a result, the average stratified location prediction model correctly predicted 98.73%. The model estimates the location with an average deviation of 2.7m in longitude and 1.95m in latitude, and the average deviation calculated by the Pythagorean theorem is 3.3m. These results are higher than the results of the basic models and compared with other studies, the results of the thesis are also ranked high. However, the proposed model still has some problems based on the results obtained from the experimental process. First, using a variety of algorithms in the first stage to generate a diverse range of predictions and estimates to improve the overall accuracy of the model can make it difficult to choose a superset. parameters for each algorithm. This is especially important because these hyperparameters can affect the performance of the model. Second, the results of the first stage are used to generate the training data for the second stage, which helps the model understand more complex relationships between features and labels, as well as improves the predictive ability, but combining the predictions of multiple models can lead to increased complexity, computational time, and the risk of overfitting if not done carefully.

The main contributions of the thesis include:

- Proposing to improve the positioning method by AP with the strongest RSS to increase positioning accuracy. As a result, the average deviation between the predicted position and the actual location decreased by 24%.

- Proposing to change the cluster selection method, although the results have not been achieved as expected, but the thesis has learned lessons, in the indoor environment with small scale, the number of locations, the AP is small, the distribution is not Evenly, the clustering and clustering methods may not achieve the set goals and need further improvement.

- Proposing a two-stage training machine learning model with the task of increasing positioning accuracy and performance. This model has shown success through solving two problems of building-floor prediction and location estimation in a building. In which, the building-floor prediction problem is implemented by the classification model, the location estimation problem is solved by two longitude estimation regression models and the latitude estimation regression models. All three models gave better results than the independent models in both model performance and accuracy, demonstrating the feasibility of the two-stage training model. Compared with other models on the same data set, the model's results are also highly appreciated.

The results of the thesis contribute to providing effective solutions to increase the efficiency and quality of indoor positioning by fingerPrinting using

WiFi's RSS, contributing to the development of location-based services. In the future, the thesis continues to expand the research on models that improve performance, positioning accuracy and can be applied to many different indoor environments.

Scalable issues include:

- Proposing to improve the positioning method by AP with the strongest RSS has increased positioning accuracy. However, the algorithmic complexity of the method is still very high up to O(N4), which leads to high positioning time. Therefore, one of the research directions that NCS will continue is to improve the AP selection algorithm so as to reduce the complexity of the algorithm, thereby helping the system to locate the location faster while still ensuring the accuracy of the location. Exactly.
- Continuing to develop the problem of reducing the size and attributes by attribute reduction technique by the reductive finding algorithm using fuzzy distance, the initial test of this direction gave positive results and was published. father at a recent conference.
- Research and apply semi-supervised and unsupervised machine learning algorithms and clustering problems
- Test the two-phase hybrid machine learning model on other datasets to further test the model's performance, quality, and scalability.
- Research and test data preprocessing methods for training data set.
- Upgrading the two-phase associative machine learning model with deep learning algorithms.
- Building realistic indoor positioning models in large buildings, applying modern technologies such as using robots to collect samples and test.

PUBLISHED SCIENCE WORKS

- [CT1] Van-Binh Ngo, Van-Hieu Vu, Do-Thanh-Tung Hoang. "Two-Phase Combined Model to Improve the Accuracy of Indoor Location Fingerprinting", Journal of Com puter Science and Cybernetics, Vol. 38 No. 4 (2022)
- [CT2] Ngo Van Binh, Vu Van Hieu. "An effective WiFi indoor positioning technique using combined machine learning", IT and communication research, development and application works - Journal of Information and Communication, Issue 2, December/ 2022.
- [CT3] Binh Ngo Van, Vuong Quang Phuong, Hoang Do Thanh Tung. "Improve the Fin- gerprinting Algorithm Based on Affinity Propagation Clustering to Increase the Accuracy and Speed of Indoor Positioning Systems", Advances in Intelligent Information Hiding and Multimedia Signal Processing. Smart Innovation, Systems and Technologies (Vol.211. No. 11,2020 Springer) (SCOPUS)
- [CT4] Ngo Van Binh, Vuong Quang Phuong, Hoang Do Thanh Tung. "Design, Construction and Clustering of Sample Datasets for Indoor Navigation Systems". Proceedings of the XX National Conference Some selected issues of Information and Communication Technology, Quy Nhon, November 2017.