BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

……..…*……**

NGÔ VĂN BÌNH

NGHIÊN CỨU CÁC GIẢI PHÁP ĐỊNH VỊ TRONG NHÀ HIỆU QUẢ DỰA TRÊN DỮ LIỆU SÓNG KHÔNG DÂY

Ngành: Hệ thống thông tin

Mã số: 9 48 01 04

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2023

Công trình này được hoàn thành tại: Học viện khoa học và Công nghệ-Viện Hàn lâm Khoa học và Công nghệ Việt nam

Người hướng dẫn khoa học học 1: TS. Hoàng Đỗ Thanh Tùng
Người hướng dẫn khoa học học 2: PGS.TS. Nguyễn Thanh Hải

Phản biện 1:…….
Phản biện 2:…….
Phản biện 3:…….

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án tiến sỹ cấp Học viện, họp tại Học viện khoa học và Công nghệ-Viện Hàn lâm Khoa học và Công nghệ Việt nam vào hồi ….. giờ……..ngày……tháng……. năm 2023

Có thể tìm luận án tại:
-Thư viện Học viện khoa học và Công nghệ
-Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Lý do chọn đề tài

**** Về mặt thực tiễn:*** Nhu cầu xây dựng các hệ thống định vị trong nhà (Indoor Positioning Systems-IPS) đã tăng lên đáng kể và thu hút nhiều sự chú ý trong những năm gần đây do giá trị thương mại cũng như ứng dụng của nó. IPS cung cấp nhiều dịch vụ dựa trên vị trí trong nhà như cứu hộ, cứu nạn, tìm đường, tiếp thị ... trong các khu vực có không gian lớn như Hình 1.

Hình 1: *Ứng dụng định vị vị trí trong nhà*

Với các loại hình dịch vụ đa dạng, doanh thu của thị trường dịch vụ dựa trên vị trí trong nhà (Indoor Locationbased Services-ILBS) ngày càng tăng. Theo trang marketsandmarkets.com[1] doanh thu của thị trường năm 2022 là 8,7 triệu USD và với tỉ lệ tăng trưởng luỹ kế hàng năm đạt 22,4% thì đến năm 2027 doanh thu dự kiến đạt 24 triệu USD. Bên cạnh đó, số lượng người sử dụng điện thoại thông minh ngày càng tăng. Theo thống kê của trang statista.com[2], số lượng người dùng điện thoại thông minh trên toàn thế giới vào năm 2022 là hơn 6.5 tỷ người, ước tính năm 2023 là hơn 6.8 tỷ người.Các số liệu thống kê đã cho thấy nghiên cứu về định vị vị trí trong nhà là điều cần thiết để phát triển các ứng dụng cung cấp các dịch vụ dựa trên vị trí trong nhà một cách trực quan.

**** Về mặt khoa học:*** Hệ thống định vị ngoài trời thường sử dụng tín hiệu vệ tinh để định vị, ví dụ như hệ thống định vị toàn cầu (Global Positioning System-GPS). GPS cung cấp hiệu suất định vị tốt và có thể định vị chính xác vị trí đối tượng từ 1-5m. Tuy nhiên, tín hiệu GPS không thể thâm nhập tốt trong môi trường trong nhà dẫn đến giảm độ chính xác định vị, do đó nhiều tín hiệu không dây khác như sóng siêu âm, băng thông siêu rộng, Bluetooth, Zigbee và WiFi đã được nghiên cứu sử dụng cho hệ thống định vị trong nhà. Trong các tiêu chuẩn không dây này, WiFi có độ chính xác định vị thấp hơn một số công nghệ khác như sóng siêu âm, băng thông rộng. Tuy nhiên, hệ thống định vị dựa trên WiFi có nhiều ưu điểm như chi phí thấp, không cần phải bổ sung phần cứng, khả năng mở rộng cao và có thể định vị vị trí đối tượng với khoảng cách sai lệch hợp lý, cùng với khả năng truyền dữ liệu cao giữa các thiết bị và tương đối ít bị ảnh hưởng bởi các nhân tố bên ngoài nên WiFi có thể cung cấp nhiều cơ hội để cải thiện độ chính xác. Hơn nữa, WiFi ngày càng trở lên phổ biến, hầu hết các thiết

[1] `https://www.marketsandmarkets.com/Market-Reports/indoor-location-market-989.html`
[2] `https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/`

bị di động hiện tại của người dùng như điện thoại, máy tính, đồng hồ thông minh đều được kích hoạt WiFi và hạ tầng sử dụng mạng WiFi cũng phát triển liên tục. Do đó, Trong các tiêu chuẩn không dây này, WiFi là tiêu chuẩn không dây phổ biến và phù hợp nhất, đã trở thành một trong những ứng cử viên lý tưởng cho định vị trong nhà và là công nghệ được nghiên cứu rộng rãi nhất. Vì vậy, việc xây dựng hệ thống định vị trong nhà dựa trên dữ liệu sóng WiFi (có thể đạt độ chính xác hợp lý) mà không cần thêm cơ sở hạ tầng là hoàn toàn khả thi.

Có nhiều kỹ thuật, phương pháp định vị trong nhà dựa trên dữ liệu sóng WiFi, bao gồm: Thời gian đến (Time of Arrival-ToA), Góc đến (Angle of Arrival-AoA), Chênh lệch thời gian đến(Time Difference of Arrival- TDoA), Tiệm cận và FingerPrinting. Trong đó, so với các phương pháp khác, phương pháp fingerPrinting tương đối đơn giản, dễ dàng tích hợp với các thiết bị thông minh, tận dụng được sự hỗ trợ từ cơ sở hạ tầng không dây hiện có (thiết bị phát WiFi, điện thoại di động,...) mà không cần thêm phần cứng. Độ chính xác, hiệu suất của fingerPrinting vẫn bị ảnh hưởng bởi vật cản trong nhà nhưng nó vẫn có thể ước lượng được vị trí đối tượng khá chính xác với khoảng cách sai lệch chấp nhận được. Do đó, phương pháp fingerPrinting là phương pháp thuận lợi hơn và có thể áp dụng cho bài toán định vị vị trí trong nhà dựa trên dữ liệu sóng WiFi.

Từ những lý do trên, luận án chọn đề tài nghiên cứu: **"Nghiên cứu các giải pháp định vị trong nhà hiệu quả dựa trên dữ liệu sóng không dây"**. Với nhiệm vụ tìm ra các giải pháp hiệu quả để nâng cao hiệu suất, độ chính xác định vị vị trí của IPS bằng phương pháp fingerPrinting dựa vào RSS của WiFi, góp phần xây dựng dịch vụ dựa trên vị trí trong nhà hữu ích cho người dùng.

Thách thức đáng kể nhất của phương pháp fingerPrinting chính là sự không ổn định của RSS. Nguyên nhân gây ra sự không ổn định của RSS là do chính bản thân thiết bị thu, phát và các vật cản trong nhà. Các thiết bị và vật cản ngoài việc làm suy giảm tín hiệu thì chúng còn gây ra hiệu ứng đa đường dẫn. Hai yếu tố này làm tăng chi phí tính toán, giảm tốc độ xử lý, giảm hiệu suất và đặc biệt là suy giảm độ chính xác định vị của phương pháp fingerPrinting. Mặc dù đã có nhiều phương pháp lấy mẫu khác nhau nhằm loại bỏ các RSS bị nhiễu, nhưng các giá trị này vẫn tồn tại bất kể phương pháp thu thập được dùng. Do đó, nhiều công trình nghiên cứu, ứng dụng đã được thực hiện nhằm nâng cao hiệu quả và độ chính xác định vị của phương pháp fingerPrinting.

Hướng nghiên cứu đầu tiên có thể kể đến là lựa chọn các AP. AP được lựa chọn dựa trên giá trị RSS. Tuy nhiên, sau khi chọn ra các AP theo phương pháp của mình, các nghiên cứu đều bỏ qua không sử dụng các AP còn lại. Cách làm này có thể làm cho một số AP bị "loại nhầm", bởi cũng do hiệu ứng đa đường và suy giảm tín hiệu dẫn đến giá trị RSS của cùng một AP thu được tại cùng một vị trí ở các thời điểm khác nhau có thể khác nhau. Do đó, phương pháp chọn AP để không "bỏ sót" giá trị RSS là một thách thức.

Hướng nghiên cứu sử dụng phương pháp phân cụm cũng đã được nhiều nhóm nghiên cứu quan tâm và thực hiện, kết quả tốc độ và độ chính xác định vị đã tăng lên. Tuy nhiên, do hiệu ứng đa đường và suy giảm tín hiệu, và theo nghiên cứu của Torres-Sospedra và cộng sự, việc sử dụng phương pháp so sánh các RSS thu được tại vị trí cần định vị với tâm các cụm để xác định cụm có thể dẫn đến việc chọn sai cụm. Do đó, nếu có phương pháp chọn cụm phù hợp thì có thể ước lượng được vị trí chính xác hơn.

Một trong những phương pháp tiếp cận phổ biến khác được nhiều nhóm nghiên cứu trong và ngoài nước tập trung nghiên cứu là sử dụng phương pháp fingerPrinting dựa trên học máy. Ngoài một số thuật toán như PCA (Principle Component Analysis), KPCA (Kernel Principal Component Analysis) được dùng để giảm đặc trưng, giảm chiều dữ liệu thì các thuật toán khác như KNN, SVM, RF... được dùng để dự đoán vị trí. Gần đây giải pháp sử dụng mô hình

học máy tổng hợp/kết hợp (Ensemble Learning model -ELM) cũng đã được áp dụng. Nhìn chung, kết quả các nghiên cứu cho thấy các thuật toán học máy đã giúp hệ thống định vị ước tính vị trí chính xác hơn và có thể áp dụng linh hoạt cho nhiều môi trường khác nhau. Mô hình ELM mặc dù đã kết hợp nhiều thuật toán và đã cho hiệu quả định vị tốt hơn các mô hình cơ sở, nhưng mô hình ELM vẫn còn tồn tại khả năng quá khớp và cách hoạt động của mô hình ELM cũng có thể bỏ qua các điểm mạnh của từng thuật toán. Bởi vậy, xây dựng mô hình học máy có thể tận dụng tối đa hiệu quả của các thuật toán, giảm nguy cơ quá khớp và tăng chất lượng định vị cho hệ thống định vị trong nhà vẫn là một thách thức.

2. Mục tiêu nghiên cứu của luận án

Với nhiệm vụ nghiên cứu để có được các giải pháp định vị trong nhà hiệu quả, luận án đặt ra mục tiêu nghiên cứu: làm thế nào để tăng khả năng xác định vị trí trong nhà hiệu quả và chính xác. Để đạt được mục tiêu này, căn cứ trên cơ sở phân tích các nghiên cứu liên quan, luận án đưa ra hai giải pháp:

1. Giải pháp thứ nhất: Cải thiện khả năng dự đoán chính xác vị trí của phương pháp fingerPrinting truyền thống bằng các biến đổi giá trị RSS thông qua phương pháp lựa chọn Access Point (AP) và phương pháp chọn cụm.

2. Giải pháp thứ hai: Tăng hiệu quả và độ chính xác của phương pháp fingerPrinting dựa trên học máy bằng mô hình học máy hai giai đoạn, trong đó kết quả huấn luyện của giai đoạn trước dùng để sinh dữ liệu huấn luyện cho giai đoạn thứ hai.

3. Nội dung nghiên cứu

a. Nghiên cứu các phương pháp lựa chọn AP, phương pháp phân cụm bằng vector RSS, phân cụm bằng vị trí và phương pháp chọn cụm.

b. Nghiên cứu các mô hình học máy, trong đó chú trọng vào nghiên cứu các mô hình học máy tích hợp nhiều mô hình học máy đồng thời.

c. Xây dựng, thực thi môi trường định vị trong nhà thực tế trên một mặt bằng. Cài đặt, thử nghiệm, đánh giá các phương pháp đề xuất của giải pháp thứ nhất trên môi trường tự xây xây dựng.

d. Cài đặt, thực nghiệm, đánh giá mô hình học máy được đề xuất trong giải pháp thứ hai trên bộ dữ liệu công cộng đa tòa, đa tầng và so sánh với các công bố khác trên cùng tập dữ liệu.

CHƯƠNG 1: TỔNG QUAN VỀ CÁC GIẢI PHÁP ĐỊNH VỊ TRONG NHÀ DỰA TRÊN DỮ LIỆU SÓNG KHÔNG DÂY

1.1. Các công nghệ không dây dùng định vị trong nhà

GPS là công cụ định vị ngoài trời phổ biến nhất và được sử dụng rộng rãi, GPS yêu cầu tầm nhìn thẳng (Line-Of-Sight - LOS) giữa các vệ tinh và thiết bị cầm tay. Tuy nhiên, vật cản (như trần nhà và tường) làm cho GPS bị suy giảm chất lượng do phản xạ tín hiệu và suy giảm tín hiệu. Điều này dẫn đến GPS không đạt hiệu quả cao và gần như không thích hợp cho việc định vị trong nhà. Có nhiều công nghệ không dây khác nhau được sử dụng thay thế GPS để định vị trong nhà. Trong đó, các công nghệ không dây được dùng phổ biến bao gồm: Nhận dạng tần số vô tuyến (Radio Frequency Identification-RFID), băng thông siêu rộng (Ultra Wide Band UWB), Bluetooth, ZigBee và WiFi.

Hệ thống nhận dạng tần số vô tuyến (RFID) có khả năng định vị và theo dõi trong nhà, nhưng triển khai RFID khó khăn vì không được hỗ trợ trên các thiết bị di động người dùng.

Công nghệ băng thông siêu rộng (Ultra-wideband - UWB) hấp dẫn vì không bị nhiễu, có khả năng xuyên qua vật liệu và độ nhạy thấp với hiệu ứng đa đường. Tuy nhiên, tiến trình tiêu chuẩn hóa UWB chậm và chi phí cao làm hạn chế việc sử dụng nó trong các sản phẩm tiêu dùng và thiết bị di động. Định vị bằng Bluetooth có ưu điểm là đơn giản, tiêu thụ năng lượng thấp, tốc độ kết nối nhanh, tốc độ truyền cao, tín hiệu ổn định và an toàn, nhưng vẫn có sai số định vị cao do hiện tượng đa đường trong môi trường trong nhà. Zigbee là giao thức truyền thông tầm ngắn có tiêu thụ điện năng thấp và giá thành rẻ, nhưng hạn chế trong phạm vi định vị, sai số lớn và khả năng chống nhiễu kém. So với các công nghệ không dây khác, hệ thống định vị dựa trên WiFi có nhiều ưu điểm như chi phí thấp, khả năng mở rộng cao, khả năng định vị với sai số hợp lý và khả năng cải thiện độ chính xác. Mạng WiFi phổ biến và hạ tầng liên tục phát triển, làm cho nó trở thành một ứng cử viên lý tưởng cho định vị trong nhà và là công nghệ được nghiên cứu rộng rãi nhất.

Do đó, trong luận án, WiFi là công nghệ không dây được lựa chọn cho bài toán định vị trong nhà, vì nó khả thi và có tiềm năng, không đòi hỏi thêm cơ sở hạ tầng.

1.2. Tổng quan các phương pháp định vị trong nhà bằng dữ liệu sóng WiFi

1.2.1. Các phương pháp

Các phương pháp định vị dựa trên WiFi có thể phân làm hai loại: phương pháp dựa trên thuộc tính về không gian và thời gian của tín hiệu nhận được (Time and Space Attributes of Received Signal-TSARS) hay còn gọi là phương pháp dựa trên phạm vi, và phương pháp định vị dựa trên cường độ tín hiệu nhận được (Received Signal Strength-RSS).

Phương pháp định vị trong nhà dựa trên phạm vi bao gồm các phương pháp Thời gian đến (Time of Arrival-ToA), Góc đến (Angle of Arrival-AoA) và Chênh lệch thời gian đến(Time Difference of Arrival- TDoA). Trong đó, ToA tính toán khoảng cách theo Thời gian đến, TDoA đo thời gian trễ, trong khi AoA đo góc của tín hiệu đến được gửi bởi các điểm truy cập khác nhau (Access Point-AP).

Công nghệ định vị dựa trên RSS sử dụng cường độ của tín hiệu nhận được để xác định vị trí của người dùng. RSS là cường độ công suất tín hiệu thực tế nhận được tại máy thu, thường được đo bằng decibel-milliwatts (dBm) hoặc milliWatts (mW). RSS có thể được sử dụng để ước tính khoảng cách giữa AP và thiết bị thu. Giá trị RSS càng cao thì khoảng cách giữa thiết bị thu và AP càng nhỏ. Có hai phương pháp chính dùng định vị trong nhà dựa trên RSS : tiệm cận (proximity), và dấu vân tay (Fingerprinting).

1.2.2. Đánh giá các phương pháp

Các ưu điểm và nhược điểm của các phương pháp dựa trên kết quả phân tích, đánh giá các khía cạnh độ phức tạp và tác động của môi trường được tổng hợp trong Bảng 1.1.

Từ các phân tích, thống kê ưu điểm, nhược điểm của từng phương pháp định vị, có thể thấy FingerPrinting là một trong các phương pháp định vị trong nhà đơn giản, có tính khả thi cao nhất và được sử dụng rộng rãi nhất trong rất nhiều nghiên cứu cũng như ứng dụng thực tế. FingerPrinting cũng là phương pháp NCS lựa chọn để nghiên cứu, phát triển các giải pháp nhằm tăng hiệu quả của hệ thống định vị trong nhà.

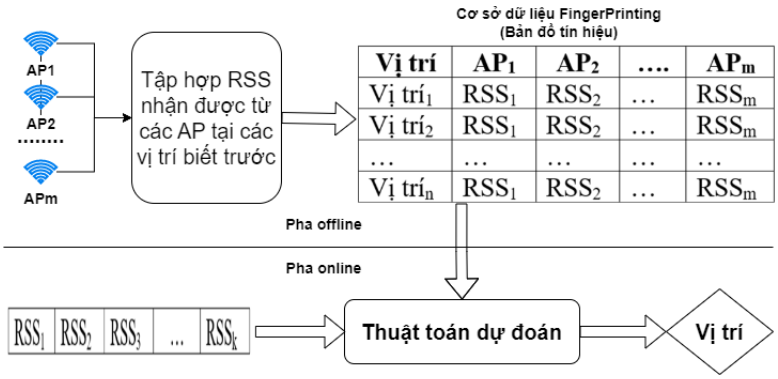
1.3. Định vị trong nhà bằng phương pháp fingerPrinting

1.3.1. Kiến trúc hệ thống định vị bằng phương pháp fingerPrinting.

Hệ thống định vị trong nhà bằng phương pháp fingerPrinting dựa trên RSS của WiFi được phân thành hai giai đoạn, giai đoạn thu thập dữ liệu ngoại tuyến (offline) và giai đoạn đối sánh trực tuyến (online) như trong Hình 1.1. Trong đó: Giai đoạn offline: Tại mỗi vị trí/điểm

Bảng 1.1: *Tổng hợp ưu điểm, nhược điểm của các phương pháp định vị trong nhà*

Phương pháp	**Ưu điểm**	**Nhược điểm**
ToA	Cung cấp độ chính xác cao trong môi trường LoS; Thuật toán khá đơn giản	Yêu cầu đồng bộ thời gian giữa AP và máy thu thường yêu cầu thêm phần cứng. Hiệu suất định vị giảm với môi trường trong nhà phức tạp không đảm bảo LoS
TDoA	Cung cấp độ chính xác cao trong môi trường LoS; Thuật toán khá đơn giản	Yêu cầu đồng bộ thời gian giữa các AP thường yêu cầu thêm phần cứng. Hiệu suất định vị giảm với môi trường trong nhà phức tạp không đảm bảo LoS
AoA	Cung cấp độ chính xác cao trong môi trường LoS	Có thể yêu cầu thêm phần cứng phức tạp như ăng-ten định hướng; yêu cầu các thuật toán tương đối phức tạp. Hiệu suất giảm trong môi trường phức tạp không đảm bảo LoS
Tiệm cận	Thuật toán đơn giản không yêu cầu bổ sung phần cứng	Độ chính xác thấp, hiệu suất định vị giảm với môi trường trong nhà phức tạp.
FingerPrinting	Không cần bổ sung phần cứng; ít chịu ảnh hưởng bởi tác động của môi trường; độ chính xác chấp nhận được; Không yêu cầu vị trí của AP	Có rất nhiều thuật toán dùng ước lượng vị trí. Quá trình chuẩn bị cơ sở dữ liệu tốn nhiều thời gian và công sức nhưng có thể phải thay đổi khi số lượng và vị trí AP thay đổi



Hình 1.1: *Kiến trúc hệ thống định vị trong nhà bằng phương pháp fingerPrinting*

tham chiếu (Reference Point-PR) đã xác định trước trên bản đồ định vị, cường độ của tín hiệu nhận được (RSS) của các AP lân cận được thu thập, chúng tạo thành vectơ RSS của vị trí với các thành phần của vectơ tuân theo cùng thứ tự của chuỗi AP. Các vectơ RSS, cùng với các vị trí được lưu trữ cùng nhau tạo thành cơ sở dữ liệu fingerPrinting (bản đồ tín hiệu); Giai đoạn online: Bằng cách so sánh và khớp vectơ RSS online thu được tại vị trí của thiết bị với các vector RSS trong cơ sở dữ liệu fingerPrinting (CSDL fingerPrinting) bằng thuật toán dự đoán, chúng ta có thể ước lượng được vị trí của thiết bị.

1.3.2. Cơ sở dữ liệu fingerPrinting

Sau quá trình xây dựng ta thu được CSDL fingerPrinting như trong Hình 1.1. Trong đó, CSDL fingerPrinting bao gồm nhiều fingerPrinting, mỗi một fingerPrinting của tín hiệu WiFi bao gồm ba yếu tố: vị trí, địa chỉ duy nhất hoặc địa chỉ MAC của AP (AP_{id}) và vector RSS với các thành phần tuân theo thứ tự của chuỗi AP nhận được ở vị trí tương ứng. Mỗi lần lấy mẫu, với tổng số AP là m thì fingerPrinting tại RP thứ i được định nghĩa trong Công thức (1.1):

$$f_i = [(ViTri_i), RSS_1, RSS_2, ..., RSS_m] \tag{1.1}$$

Trong đó, giá trị RSS của AP không phát hiện được tại RP sẽ được đặt giá trị mặc định (thông thường là 100). Cơ sở dữ liệu fingerPrinting thu được từ n vị trí có cấu trúc trong (1.2).

$$D_n(F_i) = \{f_{i_1}, f_{i_2}, ..., f_{i_k}\} \tag{1.2}$$

1.4. Các yếu tố ảnh hưởng đến chất lượng định vị của hệ thống định vị trong nhà bằng fingerPrinting

Các vật cản tĩnh, động cùng với các yếu tố thiết bị thu, phát có thể làm suy giảm tín hiệu. Bên cạnh đó, các vật cản tĩnh (như cửa sổ, cửa ra vào, tường, đồ vật...) tồn tại trong không gian trong nhà cùng với sự di chuyển của con người, việc đóng, mở các cửa làm cho tín hiệu được truyền qua các đường khác nhau, khiến tín hiệu đến được máy thu vào những thời điểm khác nhau, dẫn đến tín hiệu có thể bị chồng chéo. Hiện tượng này được gọi là hiệu ứng đa đường.

Do fingerPrinting dựa vào RSS để ước tính vị trí của người dùng nên hiệu ứng đa đường dẫn và suy giảm tín hiệu gây hậu quả đáng kể đối với định vị trong nhà, không chỉ chi phí lưu trữ đắt đỏ mà chi phí tính toán cũng tăng lên kéo theo tốc độ xử lý chậm, đặc biệt là suy giảm hiệu quả và độ chính xác của hệ thống định vị. Do đó, việc cải thiện chất lượng, tăng hiệu quả của RSS đồng thời tăng độ chính xác, hiệu suất của hệ thống định vị là rất có giá trị.

1.5. Các phương pháp tăng hiệu quả, độ chính xác định vị của phương pháp fingerPrinting

1.5.1. Phương pháp chọn AP.

Phương pháp FingerPrinting sử dụng tất cả các RSS từ các điểm truy cập để xác định vị trí, nhưng với quá nhiều RSS, hiệu ứng đa đường làm giảm độ chính xác và tăng gánh nặng hệ thống. Hầu hết các giải pháp chọn AP dựa trên độ lớn của RSS, vì AP có RSS mạnh nhất mang lại độ chính xác cao. Feng Chen và cộng sự chọn AP mạnh nhất trong giai đoạn online, và sử dụng tiêu chí Fisher trong giai đoạn offline. Thuật toán MaxMean sắp xếp các phép đo RSS trung bình từ nhiều AP và chọn AP mạnh nhất để định vị. Một nghiên cứu khác chia AP thành ngưỡng RSS khác nhau và chọn AP cùng ngưỡng cao nhất trong giai đoạn online. Thuật toán xếp hạng phần dư chọn AP ít nhạy cảm và loại bỏ AP ít xuất hiện trong FingerPrinting. Cách tiếp cận dựa trên phân biệt nhóm, lựa chọn nhóm tối ưu dựa trên thông tin chung giữa các AP. Phương pháp chọn AP dựa trên RSS đơn giản nhưng bỏ qua các AP còn lại, nhưng do hiệu ứng đa đường, cùng một AP tại các thời điểm khác nhau có thể có giá trị RSS khác nhau. Điều này có nghĩa, tại thời điểm lấy mẫu, AP có thể gần nhưng RSS lại thấp. Do đó, cần nghiên cứu giải pháp chọn AP mà không "lãng phí" AP.

1.5.2. Phương pháp phân cụm

Hai phương pháp phân cụm được sử dụng phổ biến là K-mean và phân cụm lan truyền độ tương đương (APC). Swangmuang sử dụng K-mean và tăng tốc độ định vị 50%. Seyed Alireza Razavi áp dụng K-mean và đã giảm thời gian tính toán. Abdullah sửa đổi K-mean bằng phân kỳ Bregman và kết quả giảm sai số trung bình. Torres-Sospedra và cộng sự cải tiến K-mean bằng cách kết hợp chọn AP mạnh nhất, tốt độ định vị đã tăng lên. Boyuan Wang kết hợp RSS và vị trí trong K-mean để cải thiện độ chính xác. Andrei Cramariuc và cộng sự sử dụng K-mean và APC, với APC có độ phức tạp tính toán thấp hơn, nhưng độ chính xác không bằng K-mean. Chen Feng và cộng sự áp dụng APC và đã giảm sai số trung bình. Zengshan Tian và cộng sự áp dụng phân cụm APC dựa trên vị trí và sai số trung bình cũng giảm. Pejman sử dụng phân cụm CSDL fingerPrinting dựa trên RSS và điểm tham chiếu, tăng hiệu suất dự đoán. Jingxue Bi và cộng sự áp dụng APC trong cả hai giai đoạn giúp độ chính xác tăng lên. Limin Wang và cộng sự cải thiện APC bằng đánh giá mật độ dữ liệu. Genming Ding và cộng sự sử dụng mạng thần kinh nhân tạo với mô hình được phân cụm bằng APC. Cả hai nghiên cứu đã giảm thời gian định vị và sai số. Các phương pháp phân cụm đã đóng góp vào tăng tốc và cải thiện định vị, nhưng hiệu ứng đa đường và suy giảm tín hiệu có thể làm cho giá trị RSS thay đổi tại cùng một vị trí ở các thời điểm khác nhau. Do đó, việc lựa chọn cụm theo cách so sánh giá trị RSS thu được ở giai đoạn online với tâm cụm có thể dẫn đến nhầm lẫn về tâm cụm, đặc biệt khi vị trí thực tế của đối tượng nằm ở giữa hai hoặc nhiều cụm. Trong trường hợp này, nếu giá trị RSS online bị thay đổi, khoảng cách giữa giá trị RSS online và tâm cụm cũng sẽ thay đổi, dẫn đến việc lựa chọn cụm sai. Bởi vậy, phương pháp lựa chọn cụm cần được cải thiện để đảm bảo độ chính xác và chất lượng định vị tốt hơn.

1.5.3. Phương pháp fingerPrinting dựa trên thuật toán học máy

CSDL fingerPrinting thường lớn với nhiều bản ghi và trường dữ liệu. Để tăng tốc xử lý và cải thiện định vị, nhiều thuật toán học máy (Machine Learning-ML) đã được áp dụng. Học máy có khả năng tìm hiểu và xác định mẫu trong dữ liệu, dựa trên quá trình học để đưa ra quyết định cho dữ liệu mới. Với fingerPrinting dựa trên học máy, mô hình học máy được huấn luyện để tìm mối quan hệ giữa vector RSS và vị trí. Khi áp dụng mô hình vào vector RSS ở giai đoạn online, độ chính xác và hiệu suất định vị tăng lên đáng kể.

1.5.3.1. Phương pháp fingerPrinting dựa trên mô hình học máy độc lập

Các thuật toán học máy đã đóng góp đáng kể trong việc giải quyết bài toán định vị trong nhà dựa trên phương pháp fingerPrinting. KNN đã được sử dụng từ rất sớm và đã cho thấy hiệu quả vượt trội so với fingerPrinting. SVM cũng được áp dụng và mang lại kết quả định vị chính xác gần như tương đương với KNN. RF được sử dụng trong không gian không có tường hoặc vật cản, và đã cải thiện đáng kể độ chính xác và thời gian thực hiện. LR và các biến thể của nó cũng đã cho kết quả tốt và cải thiện độ chính xác so với fingerPrinting. Ngoài ra, các thuật toán như DNN và LightGBM cũng đã được áp dụng và mang lại hiệu suất cao hơn trong việc định vị. Các thuật toán khác như LDA và NB (Naive Bayes) cũng đã được thử nghiệm và cho kết quả định vị khá tốt. Nhìn chung, áp dụng các thuật toán học máy đã nâng cao khả năng định vị chính xác và cải thiện hiệu suất của hệ thống so với phương pháp fingerPrinting truyền thống.

Mỗi thuật toán có ưu điểm và hạn chế riêng, và sự lựa chọn thuật toán phụ thuộc vào yêu cầu của bài toán và dữ liệu. Tuy nhiên, nếu chỉ sử dụng một thuật toán trong hệ thống định vị, có thể bỏ sót khả năng của các thuật toán khác. Do đó, nhiều nhóm nghiên cứu đã sử dụng mô hình kết hợp (Ensemble Learning model -ELM) nhằm tận dụng tốt hơn ưu điểm của các thuật toán và tăng hiệu quả định vị của hệ thống.

1.5.3.2. Phương pháp fingerPrinting dựa trên các mô hình học máy kết hợp

Mô hình học máy kết hợp (Ensemble Learning Model-ELM) bao gồm một tập hợp các mô hình được kết hợp để tạo thành một mô hình mạnh hơn. Ý tưởng chính của Ensemble Learning là kết hợp các dự đoán của nhiều mô hình khác nhau để đưa ra một dự đoán cuối cùng có độ chính xác cao hơn. Cụ thể, việc kết hợp DNN và KNN trong một nghiên cứu đã đem lại kết quả tốt hơn với sai số từ 1,39m đến 1,5m. Sử dụng mô hình Ensemble Learning (ELM) cũng đã mang lại kết quả đáng chú ý, với sai lệch khoảng 4m trong 80% thử nghiệm và RMSE là 8,79m và 8,83m cho trục X và trục Y. Một nghiên cứu khác đã phát triển mô hình ELM dựa trên KNN, DNN, RF và SVM, và kết quả "voting" của các mô hình đã dự đoán vị trí với sai lệch 1,1 trong 60,38% thử nghiệm. Tuy nhiên, mặc dù các phương pháp này đã cải thiện độ chính xác và hiệu suất của mô hình, vẫn tồn tại một số thách thức. Một vấn đề phổ biến là khả năng quá khớp (overfitting) khi huấn luyện các mô hình trên cùng một tập dữ liệu. Ngoài ra, việc đánh trọng số hoặc sử dụng cơ chế bầu chọn ("voting") kết quả dự đoán của các mô hình cơ sở có thể làm giảm độ tin cậy của dự đoán cuối cùng. Để giải quyết những vấn đề này, cần xây dựng các mô hình mới có khả năng hạn chế quá khớp và nâng cao hiệu quả thông qua kết quả huấn luyện của các mô hình cơ sở.

Kết chương 1 Trong chương 1, đầu tiên luận án trình bày tổng quát bài toán định vị trong nhà dựa trên dữ liệu sóng không dây và các vấn đề của bài toán. Tiếp đó, các công nghệ không dây phổ biến được dùng trong bài toán định vị trong nhà được giới thiêu, sau khi đánh giá và so sánh các công nghệ thì WiFi là công nghệ phù hợp nhất. Hệ thống định vị trong nhà dựa trên dữ liệu sóng WiFi có thể thực thi bằng nhiều kỹ thuật, phương pháp khác nhau. Trong số đó, phương pháp fingerPrinting được được đánh giá cao nhất do có chi phí thấp, phù hợp với môi trường trong nhà, dễ triển khai và độ chính xác chấp nhận được. Tuy nhiên, phương pháp fingerPrinting phải đối mặt với hai thách thức làm giảm độ chính xác và hiệu quả định vị của hệ thống, đó là hiệu ứng đa đường và suy giảm tín hiệu sóng. Để tăng chất lượng, hiệu suất định vị của phương pháp fingerPrinting, nhiều giải pháp đã được đưa ra bởi nhiều nhóm nghiên cứu.

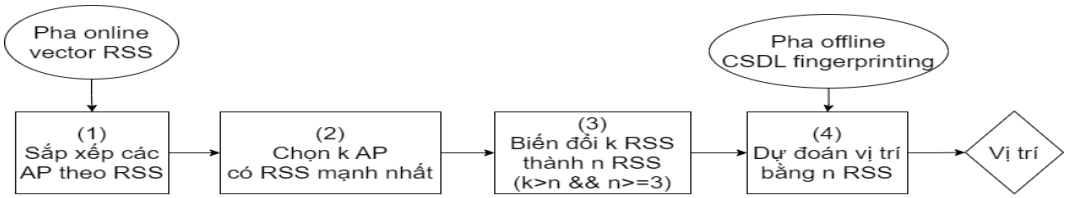
CHƯƠNG 2: PHƯƠNG PHÁP CHỌN AP VÀ PHÂN CỤM CƠ SỞ DỮ LIỆU FINGERPRINTING

2.1. Đặt vấn đề

Trong các tòa nhà và trung tâm thương mại, việc trang bị nhiều AP phát WiFi để đảm bảo chất lượng truy cập Internet đã trở nên phổ biến. Tuy nhiên, việc tăng số lượng và mật độ AP cũng đặt ra những thách thức cho quá trình định vị trong nhà bằng trên phương pháp fingerPrinting dựa trên RSS của WiFi. Vấn đề đầu tiên là hiện tượng đa đường gây ảnh hưởng đáng kể đến chất lượng định vị. Nhiều nghiên cứu đã tìm hiểu số lượng AP cần thiết và đề xuất cách chọn AP dựa trên giá trị của RSS để tăng chất lượng định vị. Tuy nhiên, tác động của hiệu ứng đa đường và suy giảm tín hiệu có thể làm thay đổi giá trị RSS của cùng một AP ở cùng một vị trí dẫn đến việc một số AP có thể bị loại nhầm. Do đó, luận án đề xuất phương pháp lựa chọn AP mới để giảm khả năng loại nhầm AP và tác động của hiệu ứng đa đường, từ đó tăng độ chính xác. Vấn đề thứ hai là độ lớn của cơ sở dữ liệu fingerPrinting tăng theo số lượng AP, làm tăng chi phí tính toán và giảm tốc độ định vị. Phương pháp phân cụm đã được áp dụng để giải quyết vấn đề này. Tuy nhiên, vẫn còn vấn đề chọn cụm trong giai đoạn trực tuyến và trong kết quả thực nghiệm của đề xuất chọn AP của luận án, hiện tượng mốt số kết quả dự đoán vị trí bị "nhảy" đi quá xa. Do đó, luận án đề xuất một phương pháp chọn cụm mới nhằm khắc phục sai lệch vị trí và cải thiện chất lượng định vị cũng như khắc phục vấn đề của đề xuất chọn AP.

2.2. Đề xuất phương pháp chọn AP

Phương pháp chọn AP được đề xuất dựa trên hai yếu tố: (1) Lựa chọn AP với giá trị RSS khả thi nhất cho quá trình định vị. (2) Sử dụng các AP có giá trị RSS mạnh nhất để đạt độ chính xác cao hơn. Tuy nhiên, hiệu ứng đa đường và suy giảm tín hiệu làm khó phân biệt giá trị RSS và có thể dẫn đến việc lựa chọn sai. Đồng thời, phương pháp chọn AP chỉ tập trung vào N AP có giá trị RSS cao nhất, bỏ qua các giá trị RSS khác và có thể gây mất mát thông tin quan trọng. Do đó, luận án đề xuất phương pháp chọn AP mới ở giai đoạn online. Hình 2.1 thể hiện lưu đồ thực hiện phương pháp chọn AP được đề xuất. Các bước thực hiện được thể hiện trong Thuật toán 2.1. Độ phức tạp thuật toán của phương pháp sẽ tăng nhanh theo giá trị k bởi số tam giác tạo ra là C(k, 3) = k! / (3! * (k - 3)!). Do đó, NCS đề nghị sử dụng số RSS tối thiểu là 3 và cao nhất là 5.



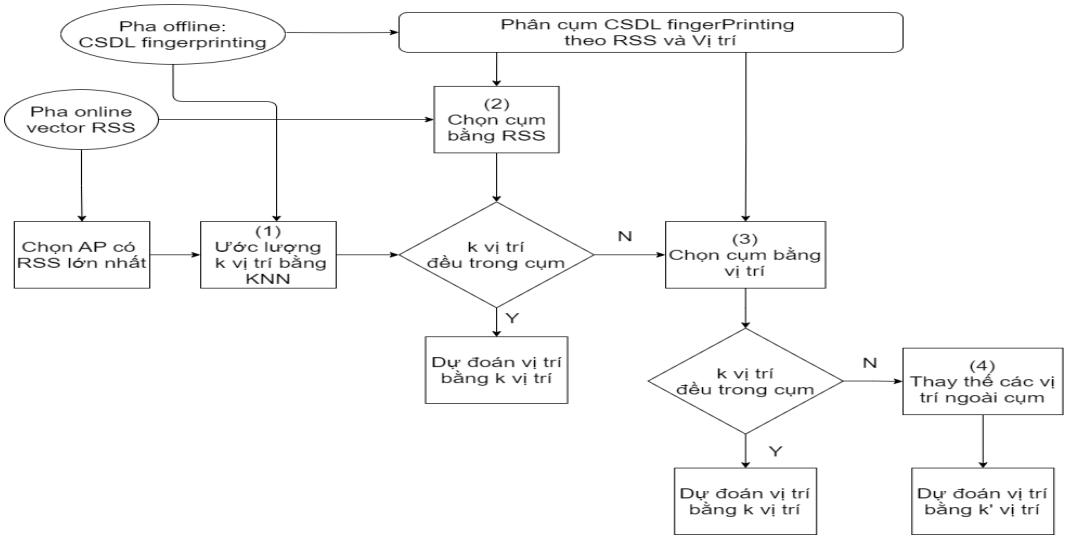
Hình 2.1: *Lưu đồ phương pháp chọn AP được đề xuất*

Thuật toán 2.1: Thuật toán định vị bằng các AP có RSS mạnh nhất.

Dữ liệu vào: $\mathscr{R} \leftarrow \{RSS_1, RSS_2, ..., RSS_m\}$ *(m giá trị RSS thu được từ m AP tại vị trí chưa xác định)*
Dữ liệu ra: $\mathscr{V}$: Vị trí được dự đoán.
begin
    Bước 1: Chọn các RSS có giá trị mạnh nhất
    Sắp xếp $\mathscr{R}$ theo chiều giảm dần;
    $\mathscr{R}_k \leftarrow \{RSS_1, RSS_2, ..., RSS_m\}$; *(k giá trị RSS lớn nhất từ $\mathscr{R}$)*
    Bước 2: Biến đổi tập $\mathscr{R}_k$ thành tập $\mathscr{R}_n$ chứa RSS mới
    Khởi tạo n là số lượng RSS cần dùng để dự đoán vị trí.
    while $k >= n$ **do**
        $\mathscr{S}_t$=tập gồm t các tam giác tạo ra từ k RSS trong $\mathscr{R}_k$;
        $\mathscr{P} \leftarrow \emptyset$; *(tập các trọng tâm tam giác)*
        for $i = 1$ *to* t **do**
            $\mathscr{P}$=$\mathscr{P} \cup$ Trọng tâm tam giác thứ i trong $\mathscr{S}_t$
        end
        Sắp xếp giá trị $\mathscr{P}_t$ theo chiều giảm dần
        k'= k-1
        $\mathscr{R}_{k'} \leftarrow \mathscr{P}_t$; *(k' phần tử đầu tiên trong $\mathscr{P}_t$)*
        $\mathscr{R}_k \leftarrow \mathscr{R}_{k'}$
    end
    Bước 4: Tính vị trí cần định vị.
    Xác định vị trí cần định vị bằng tập RSS mới trong $\mathscr{R}_k$; *(k=n)*
    $\mathscr{V} \leftarrow$ Vị trí dự đoán;
    Return $\mathscr{V}$;
end

2.3. Đề xuất phương pháp chọn cụm

Trong phần này, luận án đề xuất một phương pháp chọn cụm, trong đó kết hợp phương pháp chọn cụm bằng các RSS online với thuật toán KNN. Lưu đồ hoạt động của phương pháp được thể hiện trong Hình 2.2. Trong các bước thực hiện phương pháp chọn cụm, phần



Hình 2.2: *Lưu đồ phương pháp chọn cụm*

thay thế các vị trí ngoài cụm bằng các vị trí lân cận cụm nhằm mục tiêu kéo k vị trí lại gần nhau hơn, khi đó khả năng dự đoán vị trí có thể chính xác hơn do các vị trí ở xa có thể làm cho vị trí được dự đoán dịch chuyển ra xa. Bên cạnh đó, việc thay thế vị trí về bản chất cũng là thay đổi giá trị RSS, việc này cũng có thể làm hạn chế tác động của hiệu ứng đa đường và suy giảm tín hiệu. Quá trình thực thi của phương pháp đề xuất được thể hiện trong Thuật toán 2.2.

2.4. Xây dựng môi trường thực nghiệm thực tế

Sau khi thiết kế và thực thi, NCS có được môi trường thực nghiệm bài toán định vị trong nhà như sau:Diện tích thực nghiệm trên một mặt sàn có diện tích 250m^2 với sơ đồ thực tế các phòng, hành lang; Số lượng AP là 39, trong đó có 6 AP được đặt cố định bởi nhóm nghiên cứu; Tổng số có 154 vị trí được gắn tọa độ (x,y).Tất cả các thử nghiệm được mhóm thực hiện trên thiết bị Samsung Galaxy S4. Do tính chất của điện thoại có màn hình độ phân giải 16:9, nên nhóm thiết kế ảnh bản đồ khớp với màn hình.Từ đó phát sinh vấn đề, tỉ lệ ảnh bản đồ trong điện thoại và thực tế không khớp nhau. Sau khi đo đạc và chia tỉ lệ bản đồ theo hệ trục tọa độ (X, Y), giá trị dùng để quy đổi theo Công thức (2.1).

$$[X : 1m = 4.175; Y : 1m = 5.9] \tag{2.1}$$

Dựa trên tỉ lệ quy đổi, sai lệch giữa vị trí dự đoán và vị trí thực tế sẽ được tính bằng đơn vị mét (m). Cụ thể, gọi (X send), (Y send) là tọa độ vị trí thực tế, (X receive) và (Y receive) là tọa độ vị trí được định vị bởi hệ thống. Sai lệch vị trí định vị được tính bằng m từ tọa độ vị trí [(X send), (Y send)] đến [(X receive),(Y receive)] theo Công thức (2.2).

$$Error(m) = \sqrt{((X_{send} - X_{receive})/4.175)^2 + ((Y_{send} - Y_{receive})/5.9)^2} \tag{2.2}$$

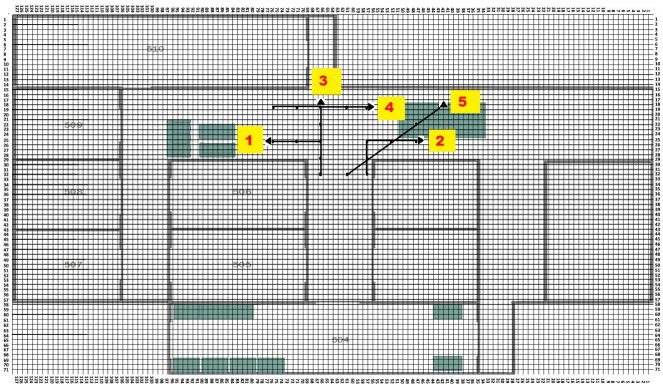
Thuật toán 2.2: Thuật toán chọn cụm.

Dữ liệu vào: $\mathscr{C}_n$=$(C_1, C_2, ..., C_n)$; *n cụm đã được tạo trước ở pha offline*
$\mathscr{R}_m \leftarrow \{RSS_1, RSS_2, ..., RSS_m\}$ *m giá trị RSS thu được từ vị trí chưa biết*
Dữ liệu ra: $\mathscr{V}$: Vị trí định vị.
begin
  Bước 1: Tính k vị trí và chọn cụm
  $\mathscr{P}_k \leftarrow \{P_1, P_2, ..., P_k\}$ *k vị trí "láng giềng" từ KNN bằng m' RSS chọn từ m RSS*; Chọn cụm bằng các RSS trong $\mathscr{R}_m$
  Bước 2: Kiểm tra k vị trí có trong cụm
  if *(k vị trí nằm trong cụm)* **then**
    $\mathscr{V} \leftarrow$ Vị trí dự đoán bằng danh sách các vị trí của $\mathscr{P}_k$
    Return $\mathscr{V}$;
  end
  Bước 3: Chọn cụm theo vị trí
  Chọn cụm theo vị trí bằng k vị trí của $\mathscr{P}_k$
  if *(k vị trí nằm trong cụm)* **then**
    $\mathscr{V} \leftarrow$ Vị trí dự đoán bằng danh sách các vị trí của $\mathscr{P}_k$
    Return $\mathscr{V}$;
  end
  Bước 4: Tìm cụm có chứa nhiều vị trí trong $\mathscr{P}_k$ nhất và thay thế vị trí
  max=0; $\mathscr{C}_{max} \leftarrow \emptyset$
  for $i = 1$ *to n* **do**
    temp=số các các vị trí của $\mathscr{P}_k$ có trong $\mathscr{C}_i$;
    if *max<temp* **then**
      max=temp;
      $\mathscr{C}_{max} \leftarrow \mathscr{C}_i$;
    end
  end
  Thay thế các vị trí không có trong $\mathscr{C}_{max}$ bằng các vị trí lân cận các vị trí của $\mathscr{P}_k$ có trong $\mathscr{C}_{max}$
  $\mathscr{P}_{k'}$: tập vị trí mới
  Bước 5: Định vị bằng danh sách các vị trí của $\mathscr{P}_{k'}$
  $\mathscr{V} \leftarrow$ Vị trí dự đoán;
  Return $\mathscr{V}$;
end

2.5. Kết quả và đánh giá phương pháp chọn AP

2.5.1. Nội dung và kịch bản thực nghiệm.

Luận án tiến hành thực nghiệm và so sánh hai phương pháp chọn AP: Phương pháp chọn AP dựa trên giá trị RSS lớn nhất và phương pháp chọn AP được đề xuất trong luận án. Phương pháp chọn AP dựa trên giá trị RSS lớn nhất sẽ chọn ra n giá trị RSS lớn nhất, trong khi phương pháp chọn AP đề xuất sẽ chọn ra m giá trị RSS mạnh nhất (trong đó m > n) và chuyển đổi thành n giá trị RSS mới. Giá trị n trong thử nghiệm là 3. NCS và nhóm đã tiến hành các kịch bản thực nghiệm dựa trên di chuyển hàng ngày của người dùng, có 5 kịch bản di chuyển thể hiện trong Hình 2.3, bao gồm: đi thẳng ngang, đi thẳng dọc, đi cua gấp khúc 90 độ sang phải, đi cua gấp khúc 90 độ sang trái, đi chéo. Tổng số 250 mẫu đã được ghi nhận cho cả 5 kịch bản di chuyển.



Hình 2.3: *Kịch bản thử nghiệm đề xuất chọn AP*

2.5.2. Kết quả thực nghiệm và đánh giá

Kết quả thực nghiệm các phương pháp được tiến hành theo từng kịch bản di chuyển. Tổng số có 250 lần thực hiện thực nghiệm. Sau đây là kết quả tổng hợp và đánh giá.

Kết quả thực nghiệm của hai phương pháp được đánh giá dựa trên sai lệch vị trí trung bình trên các kịch bản. Bảng 2.1 hiển thị sai lệch vị trí trung bình của phương pháp chọn AP dựa trên giá trị RSS mạnh nhất, trong khi Bảng 2.2 thể hiện sai lệch vị trí trung bình của phương pháp chọn AP được đề xuất. Kết quả cho thấy, sai lệch vị trí trung bình của hai phương pháp trên tất cả các kịch bản lần lượt là 3.23m và 2.46m. Điều này cho thấy, phương pháp chọn AP đề xuất giảm sai lệch trung bình khoảng 24% so với phương pháp chọn AP dựa trên giá trị RSS mạnh nhất.

Bảng 2.1: *Sai lệch vị trí trung bình của phương pháp chọn AP có RSS mạnh nhất*

Số kịch bản	**Sai lệch (X)**	**Sai lệch (Y)**	**Sai lệch trung bình (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
		Trung bình sai lệch	**3.23**

Bảng 2.2: *Sai lệch vị trí trung bình phương pháp chọn AP được đề xuất*

Số kịch bản	**Sai lệch (X)**	**Sai lệch (Y)**	**Sai lệch trung bình (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
		Trung bình sai lệch	**2.46**

Các kết quả thực nghiệm cùng với đánh giá kết quả giữa hai phương pháp chọn AP dựa trên giá trị RSS mạnh nhất và phương pháp chọn AP dựa trên các biến đổi giá trị RSS đã chứng minh tính khả thi của phương pháp được đề xuất trong luận án, và khả năng cải thiện chất lượng định vị vị trí của phương pháp fingerPrinting. Tuy nhiên, trong quá trình thực

Bảng 2.3: *Thống kê số lượng sai lệch vị trí của phương pháp chọn AP đề xuất*

Kich bản	Sai lệch			
	>=4m	>=5m	>=6m	>=7m
1	2	2	0	0
2	0	0	0	0
3	3	1	0	0
4	3	1	0	0
5	3	0	0	0

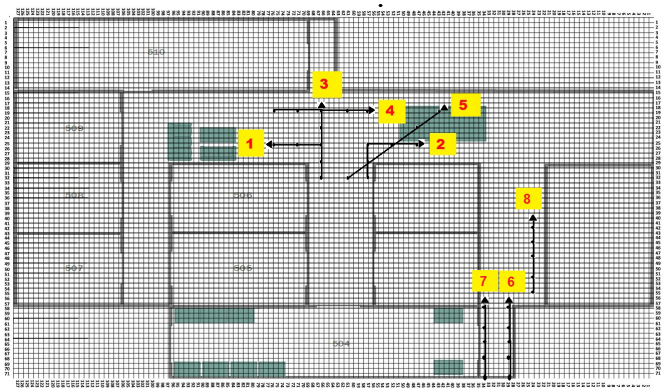
nghiệm, phương pháp đề xuất vẫn còn một số trường hợp vị trí dự đoán có sai lệch lớn hơn 4m so với vị trí thực như trong Bảng 2.3. Vì vậy, để giải quyết vấn đề này và nâng cao độ chính xác của quá trình định vị, luận án đã nghiên cứu phương pháp phân cụm và đề xuất một phương pháp chọn cụm tương ứng. Hy vọng rằng, phương pháp này sẽ giải quyết được vấn đề sai lệch lớn trong kết quả thực nghiệm và cải thiện độ chính xác của quá trình định vị.

2.6. Kết quả và đánh giá phương pháp chọn cụm.

Trong giai đoạn đầu tiên, luận án tiến hành thử nghiệm cả hai phương pháp phân cụm k-means và APC (phân cụm độ lan truyền tương đương) để lựa chọn phương pháp phân cụm phù hợp với môi trường đã xây dựng. Dựa trên những kết quả thử nghiệm, luận án chọn phương pháp APC làm phương pháp phân cụm cho các thử nghiệm tiếp theo.

2.6.1. Nội dung và kịch bản thực nghiệm

Phương pháp được thực nghiệm tại hai khu vực khác nhau trên bản đồ. Các khu vực và hướng di chuyển thể hiện trong hình 2.4. Sở dĩ có việc chia làm hai khu vực bởi bản đồ định vị không đồng đều và phân bố AP cũng không đồng đều, điều này dẫn đến chất lượng RSS tại các khu vực là khác nhau. Đầu vào của thuật toán KNN vẫn là phương pháp chọn AP đã đề xuất với số lượng RSS được chọn là 4RSS.



Hình 2.4: *Kịch bản thử nghiệm đề xuất chọn cụm*

2.6.2. Kết quả thực nghiệm và đánh giá.

Bảng 2.4 thể hiện kết quả định vị vùng 1 có các kịch bản từ 1 đến 5. Bảng 2.5 thể hiện kết quả vùng 2 của các kịch bản 6 đến 8. Kết quả thực nghiệm trên hai vùng cho kết quả rất khác nhau, tại vùng 1 với các kịch bản từ 1 đến 5, sai lệch trung bình giữa vị trí dự đoán và vị trí thực là 4,08m, nhưng với vùng 2 từ kịch bản 6 đến 8 sai lệch trung bình giảm gần

Bảng 2.4: *Kết quả vùng 1, các kịch bản từ 1 đến 5*

Số kịch bản	**Sai lệch (X)**	**Sai lệch (Y)**	**Sai lệch trung bình (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
Trung bình sai lệch			**4.08**

Bảng 2.5: *Kết quả vùng 2, các kịch bản từ 6 đến 8*

Số kịch bản	**Sai lệch (X)**	**Sai lệch (Y)**	**Sai lệch trung bình (m)**
6	1.73	0.51	1.93
7	1.59	0.44	1.68
8	1.84	1.68	2.92
Trung bình sai lệch			2.18

2m còn 2,18m. Với bài toán định vị trong nhà, con số chênh lệch 2m không phải là nhỏ. Sự chêch lệch này được giải thích là do sự phân bố không đồng đều trên bản đồ cả về mặt sơ đồ lẫn AP (chú ý rằng, phân vùng 2 được nhóm đặt thêm 6 AP cố định). So sánh kết quả với đề xuất chọn AP thì chất lượng định vị khi dùng phân cụm tại vùng một với các kịch bản từ 1 đến 5 bị giảm, sai lệch trung bình khi chưa áp dụng phân cụm là 2.46m, sau khi áp dụng phân cụm tăng lên 4.08m. Phân vùng 2, với các kịch bản từ 6 đến 8 có vẻ tốt hơn với sai số trung bình 2.18m. Tuy nhiên do mô hình định vị bằng các AP có RSS mạnh nhất không thử nghiệm trên phân vùng này nên không có cơ sở để so sánh.

Có nhiều nguyên nhân dẫn đến phương pháp đề xuất không đạt kỳ vọng, trong đó có bản đồ không đủ lớn, các vị trí thu thập dữ liệu chỉ tập trung vào các hành lang dẫn đến phân bố không đồng đều, số lượng AP cũng có thể gây ra phân cụm, chọn cụm không được như mong muốn.

Kết chương 2

Trong Chương 2, luận án trình hai phương pháp xử lý dữ liệu ở giai đoạn đoạn online nhằm khắc phục tác động của hiệu ứng đa đường, suy giảm tín hiệu lên RSS để tăng độ chính xác định vị. Các phương pháp đã được thực nghiệm trên môi trường thực tế được NCS cùng nhóm nghiên cứu xây dựng công phu. Trong số hai phương pháp đề xuất, kết quả của phương pháp chọn AP cho thấy sự khả thi của phương pháp. Phương pháp chọn cụm tuy chưa đạt được kết quả mong đợi nhưng giúp khẳng định thêm sự thiếu hụt về dữ liệu, phân bố không đồng đều các RP, AP là nguyên nhân gây ra giảm chất lượng định vị và gây bất lợi cho phương pháp phân cụm.

CHƯƠNG 3: MÔ HÌNH HỌC MÁY HAI GIAI ĐOẠN

3.1. Đặt vấn đề

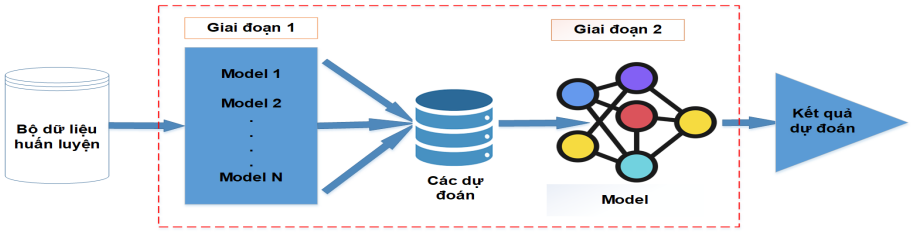
Mỗi thuật toán học máy mang những lợi thế riêng so với các thuật toán khác. Do đó, việc kết hợp các thuật toán học máy khác nhau có thể tạo ra một giải pháp toàn diện cho một ứng dụng cụ thể. Bằng cách hợp nhất thông tin từ các thuật toán học máy khác nhau, Mô hình học máy kết hợp (ELM) có thể cải thiện độ chính xác và hiệu suất của hệ thống tổng thể so

với các mô hình của các thuật toán riêng lẻ. Mô hình ELM tập trung vào việc kết hợp các dự đoán của các mô hình riêng lẻ để tạo ra dự đoán cuối cùng. Trong khi mỗi mô hình con trong ELM có thể có xu hướng riêng để có thể xẩy ra hiện tượng quá khớp dữ liệu. Khi các mô hình con có xu hướng này, mô hình kết hợp có thể bị ảnh hưởng và kế thừa những đặc điểm không mong muốn này. Điều này dẫn đến việc mô hình kết hợp cũng bị quá khớp dữ liệu huấn luyện và khó có thể thể dự đoán tốt trên dữ liệu mới.

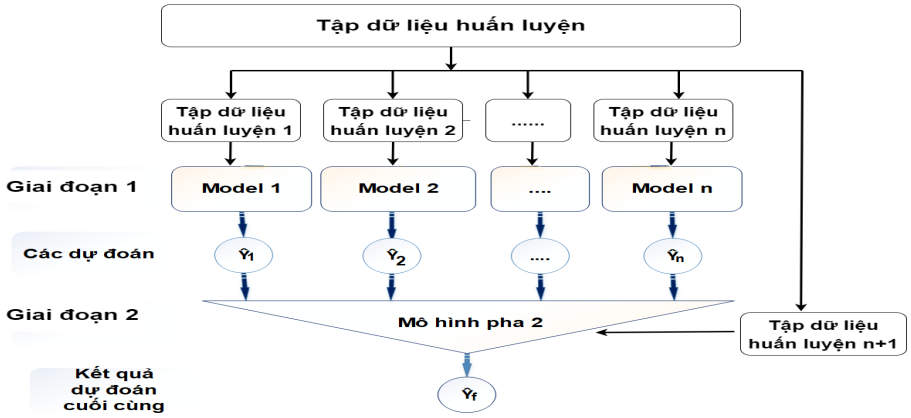
Trong chương này, luận án đề xuất một mô hình học máy hai giai đoạn. Thay vì tổng hợp các dự đoán của các mô hình riêng lẻ để tạo ra dự đoán cuối cùng như ELM, mô hình học máy hai giai đoạn hợp nhất các kết quả huấn luyện từ các mô hình riêng lẻ trong giai đoạn đầu tiên, tận dụng sự đa dạng và khác biệt giữa các mô hình để sinh ra dữ liệu huấn luyện cho giai đoạn tiếp theo. Mô hình hai giai đoạn có khả năng cung cấp quá trình huấn luyện liên tục và tăng cường hiệu quả cũng như độ chính xác trong dự đoán vị trí. Ngoài ra, việc sử dụng dữ liệu huấn luyện phát sinh từ nhiều mô hình khác nhau trong giai đoạn một giúp giảm khả năng bị quá khớp của mô hình tổng thể.

3.2. Mô hình đề xuất

Trong phần này, luận án đề xuất mô hình huấn luyện hai giai đoạn có mục tiêu tăng tính đa dạng và độ chính xác của dữ liệu huấn luyện cho mô hình giai đoạn hai. Phương pháp huấn luyện mô hình hai giai đoạn tận dụng tính đa dạng của các mô hình trong giai đoạn một và kết hợp kết quả của chúng để sinh ra dữ liệu huấn luyện đa dạng và và cung cấp khả năng dự đoán chính xác hơn cho giai đoạn hai. Điều này giúp giảm khả năng quá khớp và cung cấp một mô hình có khả năng dự đoán và tổng quát hóa tốt hơn trên dữ liệu mới. Mô hình đề



Hình 3.1: *Mô hình huấn luyện hai giai đoạn*



Hình 3.2: *Quá trình huấn luyện hai giai đoạn của mô hình*

xuất của luận án được hiển thị trong Hình 3.1. Quá trình huấn luyện mô hình hai giai đoạn đã được hiển thị trong Hình 3.2, trong đó $\hat{Y}_1$ $\hat{Y}_2$, ... và $\hat{Y}_n$ là kết quả dự đoán của n mô hình trong giai đoạn đầu tiên, các kết quả này sẽ được dùng cùng với bộ dữ liệu testing để để sinh

bộ dữ liệu huấn luyện cho thuật toán ở giai đoạn tiếp theo. $\hat{Y}_f$ là kết quả cuối cùng của giai đoạn thứ hai. Quá trình huấn luyện chi tiết của mô hình được trình bày trong Thuật toán 3.1 với độ phức tạp tính toán $\mathscr{O}\left(Max\left(\|\mathscr{D}_i\|\right) * m * n\right)$.

Thuật toán 3.1: Thuật toán huấn luyện mô hình hai giai đoạn

Dữ liệu vào: $\mathscr{D} \leftarrow \{x_i, y_i\}_1^m, x_i \subset \mathbb{X}$, $y_i \subset \mathbf{y}$. Với $\mathbb{X}$ là tập các đặc trưng, **y** là tập các nhãn, m là số các dòng trong tập dữ liệu.
Dữ liệu ra: $\hat{Y}_f$
begin
  Step 1:
  Khởi tạo $\{M_1, M_2, ..., M_n\}$; *n thuật toán học máy cho pha đầu tiên*
  Chia $\mathscr{D}$ thành các tập con $\{\mathscr{D}_1, \mathscr{D}_2, ..., \mathscr{D}_n, \mathscr{D}_{n+1}\}$; *n+1 tập con của* $\mathscr{D}$
  $\mathscr{D}^{'} \leftarrow \emptyset$; *Tập dữ liệu huấn luyện của pha thứ hai*
  Step 2: Huấn luyện bằng các thuật toán của pha đầu tiên
  for $i = 1$ *to* n **do**
    $\left(X_i^{train}, y_i^{train}, X_i^{test}, y_i^{test}\right) \leftarrow \mathscr{D}_i$; *Chia* $\mathscr{D}_i$ *thành các tập huấn luyện và kiểm thử*
    $Model_i^0 \leftarrow train\left(M_i, \left(X_i^{train}, y_i^{train}\right)\right)$; *Mô hình của* M_i
    $\hat{Y}_i \leftarrow Model_i^0(X_i^{test})$; *Kết quả dự đoán của* $Model_i^0$
    $\mathscr{D}_i^{'} \leftarrow \left(X_i^{test}, \hat{Y}_i\right)$; *Dữ liệu kết hợp cho giai đoạn hai*
    $\mathscr{D}^{'} \leftarrow \mathscr{D}^{'} \cup \mathscr{D}_i^{'}$;
  end
  Step 3: Huấn luyện bằng thuật toán của giai đoạn hai Khởi tạo: $M_{Combine}$;
  $Model^1 \leftarrow train\left(M_{Combine}, \mathscr{D}'\right)$; *Huấn luyện mô hình ở pha thứ hai*
  $\hat{Y}_f \leftarrow Model^1\left(\mathscr{D}_{n+1}\right)$; *Kết quả dự đoán của* $Model^1$
end

3.3. Môi trường thực nghiệm và bài toán định vị

3.3.1. Bộ dữ liệu thực nghiệm

Mô hình học máy hai giai đoạn được thực nghiệm trên tập dữ liệu UJIIndoorLoc, đây là tập dữ liệu đa tòa nhà, đa tầng có nhiều nhóm nghiên cứu sử dụng và phù hợp với bài toán ở chương 3 của luận án. Bộ dữ liệu UJIIndoorLoc được thực hiện bởi nhóm nghiên cứu thuộc Đại học Jaume I Tây Ban Nha. Hệ thống định vị trong nhà của Trường Đại học này được xây dựng trên 3 tòa nhà, mỗi tòa nhà có 4 hoặc 5 tầng, tổng diện tích 108.703m^2. UJIIndoorLoc có tổng cộng 21.049 mẫu, trong đó 19.938 mẫu cho training dataset và 1.111 mẫu cho validation Dataset.

3.3.2. Bài toán định vị

Bộ dữ liệu UJIIndoorLoc đại diện cho môi trường định vị trong nhà đa tòa, đa tầng. Do đó, bài toán định vị trong nhà được giải quyết bằng mô hình luận án đề xuất được phát biểu như sau: Cho hệ thống định vị trong nhà gồm có B tòa nhà, mỗi tòa nhà gồm có F tầng. Trong mỗi tầng được lắp đặt nhiều AP. Gọi ap_i là giá trị RSSI nhận được từ AP_i tại một điểm lấy mẫu trong tòa B_i và ở tầng F_j. Nếu tổng số AP có trong tất cả các tòa nhà là N thì mỗi lần lấy mẫu ta nhận được một véc tơ đặc trưng như Phương trình (3.1).

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

trong đó $ap_i = \overline{-104,0}$ và $ap_i = 100$ nếu không có tín hiệu. Vector đặc trưng f_i có một nhãn tương ứng là kinh độ và vĩ độ (ký hiệu là x_i và y_i), tòa nhà xác định b_t và tầng f_t xác định. Sau khi lấy mẫu ở tất cả các điểm tham chiếu chúng ta có một cơ sở dữ liệu $\mathscr{D}$ chứa các vector đặc trưng cùng với nhãn tương ứng của chúng như Phương trình (3.2).

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

Để huấn luyện, chúng ta biết giá trị cường độ của N RSS và nhãn tương ứng, ví dụ như $(a_1, x_1, y_1, b_{t1}, f_{t1})$. Để dự báo, chúng ta biết các giá trị RSS cho (a_2), và ước lượng nhãn tương ứng là $(x_2, y_2, b_{t2}, f_{t2})$

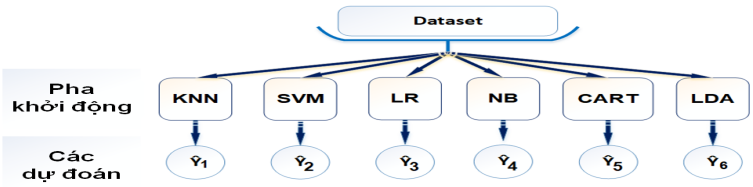
Như vậy chúng ta có tập dữ liệu $\mathscr{D} = \{\mathbb{X}, \mathscr{Y}\}$, trong đó tập $\mathbb{X} = [(f_i, f_2, ..., f_N)]$ là tập các đặc trưng và $\mathscr{Y} = [(x_1, y_1, b_{t1}, f_{t1}), ..., (x_N, y_N, b_{tN}, f_{tN})]$ là tập các nhãn tương ứng.

Trong đó, bài toán cần xác định vị trí người dùng/thiết đang ở tòa nhà nào, tầng nào (tòa-tầng nào) dựa trên các nhãn tòa B_i và tầng F_j và đang ở vị trí nào dựa trên các nhãn kinh độ và vĩ độ. Trong bộ dữ liệu UJIIndoorLoc, các tòa B_i và tầng F_j chứa các giá trị rời rạc và kinh độ, vĩ độ (x_i,y_i) chứa các giá trị liên tục. Do đó, dựa trên tính chất dữ liệu của các nhãn, luận án xây dựng hai mô hình: mô hình phân lớp thực thi bài toán dự đoán tòa-tầng và mô hình hồi quy thực thi bài toán ước lượng vị trí.

3.4. Mô hình phân lớp hai giai đoạn dự đoán tòa tầng

3.4.1. Xây dựng và đề xuất mô hình phân lớp hai giai đoạn dự đoán tòa tầng

3.4.1.1. Xây dựng mô hình



Hình 3.3: *Quy trình thực thi các mô hình phân lớp độc lập dự đoán tòa-tầng*

Dựa trên kết quả nghiên cứu các thuật toán học máy ở chương 1, NCS đã chọn một số thuật toán phân lớp để chọn ra các thuật toán tốt nhất cho giai đoạn một của mô hình. Các thuật toán bao gồm LR, LDA, KNN, CART, GB và SVM và qui trình hoạt động được thể hiện trong Hình 3.3.

Hiệu suất của các mô hình độc lập được tổng hợp thể hiện rõ nét hơn thông qua chỉ số macro averages. Bảng 3.1 thể hiện các chỉ số macro averages. Các chỉ số của các mô hình SVM, KNN và LR đều cao hơn các mô hình còn lại. Chỉ số của LR chỉ nhỉnh hơn của CART một chút, nhưng theo các khảo cứu đã có thì LR có nhiều ưu điểm hơn CART và để giảm tải cho hệ thống, luận án chỉ chọn thuật toán LR.

Khả năng dự đoán đúng tòa-tầng của các mô hình được thể hiện trong Bảng 3.2 Một lần nữa, các mô hình SVM,KNN và LR lại có khả năng dự đoán đúng tầng tốt hơn các mô hình CART, LDA và NBR. Tổng hợp các kết quả so sánh hiệu suất và kết quả dự đoán. Trong giai đoạn thứ hai, NCS chọn thuật toán Logistic Regression (LR). Dựa trên các kết quả này, mô hình phân lớp hai giai đoạn dự đoán tòa-tầng được luận án đề xuất trong phần tiếp theo.

Bảng 3.1: *Tổng hợp hiệu suất của các mô hình độc lập dự đoán tòa-tầng bằng chỉ số Macro averages*

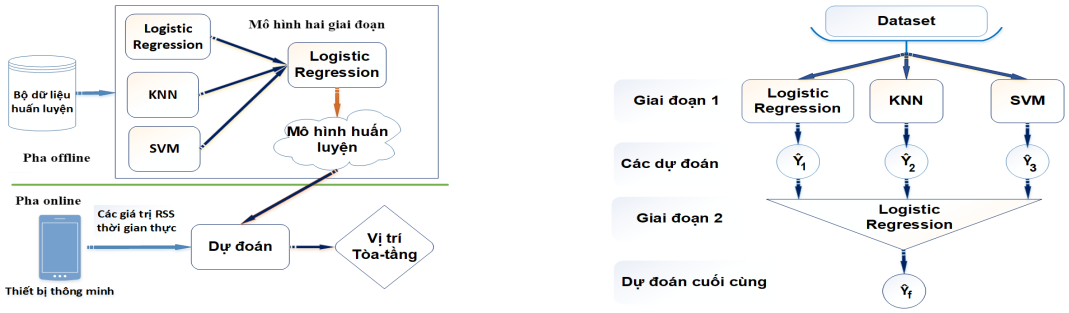
Macro averages	**SVM**	**KNN**	**LR**	**CART**	**LDA**	**NB**
Precision	**98.43**	**97.71**	**96.62**	96.50	94.42	63.70
Recall	**98.47**	**97.98**	**96.69**	96.71	94.26	55.37
F1 score	**98.45**	**97.83**	**96.65**	96.60	94.33	47.42

Bảng 3.2: *Kết quả dự đoán đúng tòa-tầng và thời gian thực thi của các mô hình độc lập*

	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. Đề xuất mô hình phân lớp hai giai đoạn dự đoán tòa-tầng

Mô hình phân lớp hai giai đoạn dự đoán tòa-tầng cùng với quá trình hoạt động của nó được thể hiện trong Hình 3.4. Trong đó hình 3.4a hiển thị mô hình hai giai đoạn. Hình 3.4b hiển thị quá trình thực thi giữa hai giai đoạn của mô hình, trong đó $\hat{Y}_1$ $\hat{Y}_2$ và $\hat{Y}_3$ là kết quả dự đoán của giai đoạn thứ nhất, bộ kết quả này kết hợp với bộ dữ liệu testing để sinh dữ liệu huấn luyện cho thuật toán LR để tạo ra kết quả cuối cùng $\hat{Y}_f$.



(a) *Cấu trúc mô hình* ***(b)*** *Quá trình huấn luyện mô hình*

Hình 3.4: *Mô hình phân lớp hai giai đoạn dự đoán tòa-tầng*

3.4.2. Kết quả thực nghiệm và đánh giá mô hình phân lớp hai giai đoạn dự đoán tòa tầng

Hiệu suất và kết quả dự đoán đúng của mô hình đề xuất thể hiện rõ ở Bảng 3.3. Các thông

Bảng 3.3: *Hiệu suất và kết quả dự đoán đúng của mô hình đề xuất dự đoán tòa-tầng*

	Macro avg Precision	**Macro avg Recall**	**Macro avg F1-Score**	**Accuracy**	**Time(s)**
Mô hình đề xuất	98.71	98.61	98.66	98.73	99.31

số trong Bảng 3.3 thể hiện kết quả hiệu suất và độ chính xác. Các chỉ số đánh giá này chỉ ra rằng mô hình đề xuất dự đoán vị trí theo tầng có hiệu suất cao và có thể dự đoán đúng tầng với tỉ lệ 98,73%.

Mô hình phân lớp hai giai đoạn dự đoán tòa-tầng có hiệu suất và tỉ lệ dự đoán đúng tầng cao. Tuy nhiên, để đánh giá sự cải thiện thực sự, cần so sánh kết quả với các mô hình độc lập.

Bảng 3.4: *So sánh hiệu suất và kết quả dự đoán của mô hình đề xuất và các mô hình độc lập dự đoán tòa-tầng*

	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%
Mô hình đề xuất	**98.71%**	**98.61%**	**98.66%**	**98.73%**

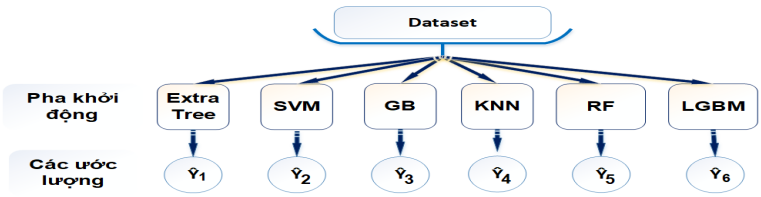
Bảng 3.4 hiển thị so sánh hiệu suất và kết quả dự đoán đúng tòa-tầng của mô hình dự đoán tòa-tầng với các mô hình độc lập. Kết quả cho thấy, về mặt hiệu suất, tất cả các chỉ số Precision, Recall, F1-Score của mô hình đề xuất đều nhỉnh hơn các mô hình độc lập. Điều này chỉ ra rằng phương pháp tiếp cận huấn luyện liên tục của các mô hình học máy, trong đó mô hình trước đó cung cấp dữ liệu cho mô hình sau đã thành công và hoàn toàn khả thi khi thực thi bài toán dự đoán tầng.

3.5. Mô hình hồi quy hai giai đoạn ước lượng vị trí

3.5.1. Xây dựng và đề xuất mô hình hồi quy hai giai đoạn ước lượng vị trí

3.5.1.1. Xây dựng và đề xuất mô hình hồi quy ước lượng kinh độ

Các thuật toán dùng để chọn ra các thuật toán tốt nhất cho giai đoạn một của mô hình hồi quy ước lượng kinh độ bao gồm các thuật toán hồi quy SVM, ExtraTree, GB, KNN, RF và LightGBM như trong Hình 3.5.



Hình 3.5: *Quy trình thực thi các mô hình hồi quy độc lập ước lượng kinh độ*

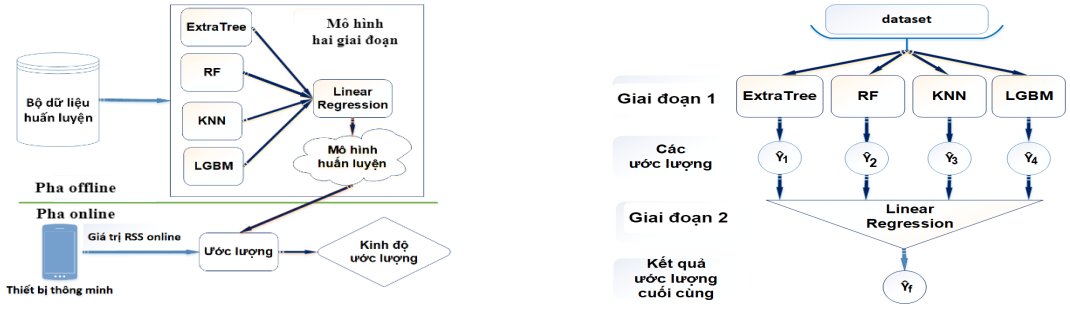
Bảng 3.5: *Hiệu suất và sai lệch của các mô hình hồi quy độc lập ước lượng kinh độ*

	SVM	**ExtraTree**	GB	**KNN**	**RF**	**LightGBM**
R2-Score(%)	96.94	**99.30**	96.7	**99.49**	**99.606**	**99.2**
MSE(m)	477.36	**109.4**	509.3	**79.39**	**61.5**	**112.4**
MAE(m)	13.85	**3.62**	16.02	**3.25**	**2.72**	**5.99**
Time(s)	59.11	**0.35**	9.63	**0.027**	**34.3**	**0.32**

Kết quả thực thi của các mô hình được hiển thị trong Bảng 3.5. Với kết quả này, trong mô hình hồi quy hai giai đoạn ước tính kinh độ, NCS chọn thuật toán hồi quy ExtraTree, KNN, RF và LightGBM cho giai đoạn đầu tiên và thuật toán Linear Regression cho giai đoạn thứ hai.

Mô hình hồi quy hai giai đoạn ước lượng kinh độ được luận án đề xuất thể hiện trong Hình 3.6. Trong đó 3.6a thể hiện mô hình và 3.6b thể hiện quá trình huấn luyện của mô hình. Trong giai đoạn đầu tiên, mô hình được huấn luyện bởi các thuật toán hồi quy ExtraTree, KNN, RF và LightGBM. Mô hình tiếp tục được huấn luyện bởi thuật toán hồi quy Linear

Regression giai đoạn 2 theo quy trình như trong thuật toán 3.1. Hình 3.6b hiển thị chi tiết quá trình huấn luyện hai giai đoạn, trong đó $\hat{Y}_1$, $\hat{Y}_2$,$\hat{Y}_3$ và $\hat{Y}_4$ là kết quả ước tính của bốn mô hình trong giai đoạn đầu tiên và $\hat{Y}_f$ là kết quả ước tính cuối cùng.



(a) Cấu trúc mô hình *(b) Quá trình huấn luyện mô hình*

Hình 3.6: *Mô hình hồi qui hai giai đoạn ước lượng kinh độ*

3.5.1.2. Xây dựng và đề xuất mô hình hồi quy ước lượng vĩ độ

Tương tự như khi xây dựng mô hình hồi quy ước lượng kinh độ, luận án cũng tiến hành thử nghiệm các mô hình độc lập bằng các thuật toán hồi quy như phần xây dựng mô hình ước lượng kinh độ và các thuật toán vẫn như trong Hình 3.5. Các kết quả thử nghiệm các mô hình hồi quy độc lập được thể hiện trong bảng 3.6.

Bảng 3.6: *Hiệu suất và sai lệch của các mô hình hồi quy độc lập ước lượng vĩ độ*

	SVM	**ExtraTree**	GB	**KNN**	**RF**	**LightGBM**
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**

Theo kết quả này,luận án chọn các thuật toán ExtraTree, KNN, RF và LightGBM cho giai đoạn thứ nhất và thuật toán Linear Regression cho giai đoạn hai. Dễ dàng có thể nhận thấy mô hình hồi quy ước lượng vĩ độ có cấu trúc giống như mô hình hồi quy ước lượng kinh độ được thể hiện trong Hình 3.6

3.5.2. Kết quả và đánh giá mô hình hồi quy hai giai đoạn ước lượng vị trí

3.5.2.1. Kết quả và đánh giá mô hình hồi quy hai giai đoạn ước lượng kinh độ

Hiệu suất và kết quả ước lượng của mô hình ước lượng kinh độ được hiển thị trong bảng 3.7. Với kết quả này, giá trị R2-score là 99,621% cho biết mô hình đã nắm bắt thành công 99,621% độ biến thiên trong biến mục tiêu (kinh độ) bằng cách sử dụng các đặc trưng (vector RSS). Điều này cho thấy rằng mô hình phù hợp tốt với dữ liệu và có thể đưa ra dự đoán chính xác về dữ liệu mới. Giá trị chỉ số MAE là 2,7m cho thấy rằng, trung bình, các dự đoán của mô hình sai lệch khoảng 2,7m so với giá trị kinh độ thực.

Bảng 3.7: *Hiệu suất và kết quả ước lượng của mô hình hồi quy ước lượng kinh độ*

	R2-Score(%)	**MSE(m)**	**MAE(m)**	**Time(s)**
Mô hình đề xuất	99.621	59.32	2.70	165.00

Bảng 3.8 so sánh hiệu suất và sai lệch giữa kinh độ ước lượng và kinh độ thực của mô hình đề xuất ước lượng kinh độ với các mô hình độc lập. Hiệu suất thể hiện trong Bảng 3.8 cho thấy mô hình đề xuất phù hợp với dữ liệu tốt hơn, có nghĩa hiệu suất cao hơn. Ước lượng kinh độ của mô hình đề xuất cũng sai lệnh ít hơn so với các mô hình độc lập, 2,7m so với giá trị gần nhất là 2,73m và xa nhất là 6m.

Bảng 3.8: *So sánh hiệu suất và kết quả ước lượng của mô hình đề xuất và các mô hình độc lập ước lượng kinh độ*

	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
Mô hình đề xuất	**99.62%**	**59.32**	**2.7**

3.5.2.2. Kết quả và đánh giá mô hình hồi quy hai giai đoạn ước lượng vĩ độ

Bảng 3.9 hiển thị hiệu suất và kết quả ước lượng vĩ độ của mô hình. Kết quả này có phần tốt hơn mô hình ước lượng kinh độ. Trong đó, con số 99,52% của R2-score thể hiện mô hình phù hợp tốt với dữ liệu và có thể đưa ra dự đoán chính xác về dữ liệu mới. Với chỉ số MAE là 1,95m cho thấy rằng vĩ độ ước lượng lệch với vĩ độ thực 1,95m, giá trị này nhỏ hơn khi ước lượng kinh độ.

Bảng 3.9: *Hiệu suất và kết quả ước tính của mô hình hồi quy ước tính vĩ độ*

	R2-Score(%)	**MSE(m)**	**MAE(m)**	**Time(s)**
Mô hình đề xuất	99.52	21.66	1.95	170.82

Hiệu suất và kết quả ước tính vĩ độ của mô hình đề xuất so với các mô hình độc lập thể hiện trong Bảng 3.10. Tương tự như khi ước tính kinh độ, mô hình đề xuất ước tính vĩ độ cũng có các chỉ số hiệu suất cao hơn và sai lệch giữa vĩ độ ước tính và vĩ độ thực thấp hơn các mô hình độc lập.

Bảng 3.10: *So sánh hiệu suất và kết quả ước tính của mô hình đề xuất và mô hình độc lập ước tính vĩ độ*

	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
Mô hình đề xuất	**99.52%**	**21.66**	**1.95**

3.5.2.3. Tổng hợp kết quả dự đoán vị trí

Với chỉ số MAE của kinh độ là 2.7m và vĩ độ 1.95 thì sai lệch trung bình của vị trí ước lượng với vị trí thực tế tính theo Công thức Euclid (3.3) là 3.3m.

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

3.6. Kết quả và đánh giá mô hình đề xuất với dữ liệu thực tế

Trong phần này mô hình đề xuất được đánh giá bằng bộ dữ liệu validation. Luận án đã thử nghiệm trên tất cả các điện thoại. Tuy nhiên, do số lượng điện thoại rất nhiều nên NCS chỉ chọn kết quả đại diện là phoneID=14. Hình 3.7 hiển thị sự sai lệch của vị trí ước lượng với vị trí thực tế được thực hiện bằng điện thoại có PhoneID=14. Trong đó, hình 3.7a hiển thị sai lệch vị trí theo mặt cắt tọa độ không gian hai chiều của kinh độ và vĩ độ, trên một mặt phẳng có nhiều điểm chồng lên nhau. Hình 3.7b hiển thị kết quả sai lệch vị trí theo kinh độ và vĩ độ theo di chuyển của người dùng (điện thoại). Màu xanh lá cây đại diện cho vị trí ước tính. Màu cam đại diện cho vị trí thực tế. Các vị trí màu biểu thị sự trùng khớp giữa vị trí ước tính và vị trí thực tế hầu như trùng nhau. Những hình ảnh này một lần nữa xác nhận sự chính xác của mô hình đã được luận án đề xuất.

(a) Mặt cắt phoneID=14 *(b) Di chuyển của phoneID=14*

Hình 3.7: *Kiểm thử độ chính xác*

3.7. So sánh kết quả mô hình đề xuất với mô hình của các nghiên cứu khác

Trong phần này luận án so sánh kết quả thực nghiệm mô hình đề xuất với kết quả của các nghiên cứu khác trên cùng bộ dữ liệu UJIIndoorLoc.

Bảng 3.11 thể hiện kết quả các nghiên cứu và kết quả mô hình của luận án. Trong đó, kết quả của các nghiên cứu khác. Theo kết quả này, về dự đoán tầng, mô hình đề xuất đứng thứ

Bảng 3.11: *So sánh kết quả mô hình đề xuất với các nghiên cứu khác*

Nghiên cứu	**Dự đoán tầng**	**Sai lệch ước lượng vị trí (MAE) (m)**
Beenish Ayesha Akram ;	-	6,46
Shivam Wadhwa ;	97,95%	7,93
Gan và cộng sự ;	95,41%	6,4
Lu Yin và cộng sự ;	99,32%	96.73%
Charoenruengkit và cộng sự ;	97%	5,65
Mô hình đề xuất	**98,73%**	**3,3**

2, kém nghiên cứu của Lu Yin và cộng sự 0,59%; Lu Yin và cộng sự không sử dụng mô hình EML. Trong nghiên cứu của nhóm, bộ mã hóa tự động khử nhiễu chính là tác nhân chính để nâng cao chất lượng định vị. Bộ mã hóa này có tác dụng trích xuất các tính năng chính từ dữ liệu RSS thưa thớt và giảm ảnh hưởng của nhiễu và dữ liệu ngoại lệ trước khi đưa dữ liệu vào thuật toán LightGBM.

Kết chương 3

Mô hình hai giai đoạn mở ra cơ hội để kết hợp các kết quả huấn luyện từ nhiều mô hình riêng lẻ, tận dụng sự đa dạng và khác biệt của chúng. Điều này mang lại lợi ích trong việc nâng cao khả năng dự đoán và độ chính xác của mô hình tổng thể. Qua đó, mô hình cung cấp một phương pháp huấn luyện liên tục và tăng cường, giúp cải thiện hiệu quả và độ chính xác trong việc ước tính vị trí. Điều này đã được thể hiện qua các mô hình dự đoán tòa-tầng và ước lượng vị trí bằng kinh độ và vĩ độ. Các kết quả thực nghiệm cho thấy rằng mô hình hai giai đoạn được đề xuất là một phương pháp học máy hiệu quả.

KẾT LUẬN

Nghiên cứu **"Nghiên cứu các giải pháp định vị trong nhà hiệu quả bằng sóng không dây"** là một hướng tiếp cận bài toán định vị trong nhà bằng phương pháp fingerPrinting dùng cường độ sóng WiFi có tính thực tiễn cao, bởi các dịch vụ dựa trên vị trí không chỉ phát triển trên toàn cầu mà còn đang dần phát triển ở Việt nam. Tuy đã có nhiều ngiên cứu, giải pháp được công bố trong thời gian gần đây, nhưng vẫn còn nhiều thách thức chưa được giải quyết hoặc có thể cải tiến thêm bởi các môi trường trong nhà khác nhau thì có sự khác biệt và phức tạp khác nhau, thậm trí trong cùng môi trường, ở các thời điểm khác nhau có thể độ phức tạp là khác nhau, do sự thay đổi của các vật cản. Bài toán định vị trong nhà bằng fingerPrinting dùng RSS của sóng WiFi vẫn luôn đối mặt với hai thách thức chính: hiệu ứng đa đường và suy giảm tín hiệu sóng.

Để giải quyết vấn đề này, luận án tiến hành nghiên cứu tổng quan về các công nghệ, kỹ thuật, mô hình xây dựng và giải quyết các vấn đề của bài toán định vị trong nhà bằng fingerPrinting dựa trên RSS của WiFi. Từ các nghiên cứu về mặt lý thuyết cũng như thực nghiệm, luận án đã đề xuất 02 cải tiến cho phương pháp fingerPrinting truyền thống bao gồm: Biến đổi giá trị vector RSS online với mục tiêu giảm tác động của môi trường đến giá trị RSS bằng phương pháp chọn AP. Thay đổi cách chọn cụm và xử lý các vị trí ngoài cụm của phương pháp phân cụm APC, các thay đổi nhằm mục đích chọn đúng cụm khả thi nhất và đảm bảo sự hội tụ của các vị trí trong cụm, từ đó nâng cao độ chính xác định vị. Hai đề xuất này được thực nghiệm trên môi trường do NCS cùng nhóm nghiên cứu tự xây dựng đảm bảo các yêu cầu của môi trường định vị trong nhà. Kết quả, cải tiến đầu tiên giúp độ chính xác tăng 24%, cải tiến thứ 2 tuy chưa tăng được độ chính xác định vị bởi phân bố vị trí và AP không đều nhau cũng như số lượng mẫu trong CSDL fingerPrinting ít.

Trong phần tiếp theo, luận án áp dụng học máy vào phương pháp fingerPrinting và đã đề xuất một mô hình học máy hai giai đoạn nhằm tăng chất lượng và hiệu suất định vị. Mô hình đề xuất được thực nghiệm trên bộ dữ liệu đa tòa, đa tầng có diện tích và số lượng mẫu lớn. Kết quả, mô hình dự đoán vị trí theo tầng trung bình dự đoán đúng 98,73%. Mô hình ước tính vị trí có sai lệch trung bình theo kinh độ là 2,7m và 1,95m theo vĩ độ, độ lệch trung bình tính bằng định lý Pythagore là 3,3m. Các kết quả này cao hơn kết quả của các mô hình cơ sở và so với các nghiên cứu khác thì kết quả của luận án cũng được xếp ở vị trí cao. Tuy nhiên, mô hình đề xuất vẫn còn gặp một số vấn đề dựa trên kết quả thu được từ quá trình thực nghiệm. Đầu tiên, sử dụng nhiều thuật toán khác nhau trong giai đoạn đầu tiên để tạo ra một loạt các dự đoán và ước lượng đa dạng để cải thiện độ chính xác tổng thể của mô hình có thể gây ra khó khăn trong việc lựa chọn siêu tham số cho từng thuật toán. Điều này đặc biệt quan trọng bởi các siêu tham số này có thể ảnh hưởng đến hiệu suất của mô hình. Thứ hai, kết quả của giai đoạn một được sử dụng để tạo dữ liệu huấn luyện cho giai đoạn hai giúp mô hình hiểu được các mối quan hệ phức tạp hơn giữa các đặc trưng và nhãn, cũng như cải thiện khả năng

dự đoán, nhưng việc kết hợp các dự đoán của nhiều mô hình có thể dẫn đến tăng độ phức tạp, thời gian tính toán và có nguy cơ overfitting nếu không thực hiện cẩn thận.

Những đóng góp chính của luận án bao gồm:

- Đề xuất cải tiến phương pháp định vị bằng AP có RSS mạnh nhất để tăng độ chính xác định vị. Kết quả, sai lệch trung bình giữa vị trí dự đoán và vị trí thực giảm 24%.

- Đề xuất thay đổi phương pháp chọn cụm, tuy chưa đạt được kết quả như kỳ vọng, nhưng luận án rút ra được bài học, trong môi trường trong nhà có quy mô nhỏ, số lượng vị trí, AP ít, phân bố không đồng đều, phương pháp phân cụm, chọn cụm có thể không đạt được mục tiêu đề ra và cần tiếp tục cải tiến.

- Đề xuất Mô hình học máy huấn luyện hai giai đoạn với nhiệm vụ tăng độ chính xác và hiệu suất định vị. Mô hình này đã thể hiện sự thành công thông qua việc giải quyết hai bài toán dự đoán tòa-tầng và ước lượng vị trí trong tòa nhà. Trong đó, bài toán dự đoán tòa-tầng được thực thi bằng mô hình phân lớp, bài toán ước lượng vị trí được giải quyết bằng hai mô hình hồi quy ước lượng kinh độ và hồi quy ước lượng vĩ độ. Cả ba mô hình đã cho kết quả tốt hơn các mô hình độc lập về cả hiệu suất mô hình và độ chính xác, thể hiện tính khả thi của mô hình huấn luyện theo hai giai đoạn. So sánh với các mô hình khác trên cùng tập dữ liệu, kết quả của mô hình cũng được đánh giá cao.

Kết quả của luận án góp phần vào việc đưa ra các giải pháp hiệu quả tăng hiệu suất, chất lượng định vị trong nhà bằng fingerPrinting dùng RSS của WiFi, góp phần phát triển các dịch vụ dựa trên vị trí. Trong tương lai, luận án tiếp tục mở rộng các nghiên cứu các mô hình nâng cao hiệu suất, độ chính xác định vị và có thể áp dụng cho nhiều môi trường trong nhà khác nhau.

Các vấn đề có thể mở rộng bao gồm:

-Đề xuất cải tiến phương pháp định vị bằng AP có RSS mạnh nhất đã tăng độ chính xác định vị. Tuy nhiên, độ phức tạp thuật toán của phương pháp còn rất cao lên tới O(N4), điều này dẫn đến thời gian định vị tăng cao. Do đó, một trong các hướng nghiên cứu mà NCS sẽ tiếp tục là cải tiến thuật toán chọn AP sao cho giảm được độ phức tạp thuật toán, từ đó có thể giúp hệ thống xác định được vị trí nhanh hơn mà vẫn đảm bảo độ chính xác.

- Tiếp tục phát triển bài toán giảm kích thước, thuộc tính bằng kỹ thuật rút gọn thuộc tính bằng thuật toán tìm tập rút gọn sử dụng khoảng cách mờ, phần thử nghiệm ban đầu của hướng này đã cho kết quả khả quan và được công bố ở hội nghị gần đây.

- Nghiên cứu và áp dụng thuật toán học máy bán giám sát và không giám sát và bài toán phân cụm

- Thử nghiệm mô hình học máy kết hợp theo hai pha trên các tập cơ sở dữ liệu khác để kiểm nghiệm thêm nữa hiệu suất, chất lượng cũng như khả năng mở rộng của mô hình.

- Nghiên cứu thử nghiệm các phương pháp tiền xử lý dữ liệu cho tập dữ liệu huấn luyện.

- Nâng cấp mô hình học máy kết hợp theo hai pha bằng các thuật toán học sâu.

- Xây dựng mô hình định vị trong nhà thực tế ở trong các tòa nhà có diện tích lớn, áp dụng các công nghệ hiện đại như dùng robot để thu thập mẫu và kiểm thử.

CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

[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] Ngô Văn Bình, Vũ Văn Hiệu. "Một kỹ thuật định vị trong nhà bằng WiFi hiệu quả sử dụng học máy kết hợp", *Các công trình nghiên cứu, phát triển và ứng dụng CNTT và truyền thông - Tạp chí Thông tin và Truyền thông*, Số 2, tháng 12/2022.

[CT3] Binh Ngo Van, Vương Quang Phương, Hoang Do Thanh Tung. "Improve the Fingerprinting 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] Ngô Văn Bình, Vương Quang Phương, Hoàng Đỗ Thanh Tùng. "Thiết kế, Xây dựng và phân cụm bộ dữ liệu mẫu cho hệ thống định vị trong nhà". *Kỷ yếu Hội nghị quốc gia lần XX Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Quy Nhơn, tháng 11/2017.