

# Một kỹ thuật định vị trong nhà bằng WiFi hiệu quả sử dụng học máy kết hợp

Ngô Văn Bình<sup>1</sup>, Vũ Văn Hiệu<sup>2</sup>

<sup>1</sup> Trường Đại học FPT, Hà Nội

<sup>2</sup> Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Tác giả liên hệ: Vũ Văn Hiệu, vvhiu@ioit.ac.vn

Ngày nhận bài: 21/08/2022, ngày sửa chữa: 15/11/2022, ngày duyệt đăng: 25/11/2022

Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n2.1124

**Tóm tắt:** Trong nghiên cứu này, chúng tôi đề xuất mô hình học tích hợp nhiều mô hình học máy theo hai pha. Pha thứ nhất kết hợp tổ hợp các mô hình phân lớp k-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machines (SVM) được tổ hợp dữ liệu mới. Pha thứ hai sử dụng một bộ phân lớp trên tổ hợp dữ liệu mới. Mô hình này được áp dụng trong giai đoạn huấn luyện mô hình. Trong giai đoạn kiểm thử, dữ liệu thu từ thiết bị điện thoại được tiền xử lý, sau đó dữ liệu được đưa vào mô hình dự báo để ước tính một vị trí tòa-tầng cụ thể trong tòa nhà. Đề xuất được cài đặt và đánh giá trên tập dữ liệu UJIIndoorLoc đạt kết quả dự báo chính xác 98.73%. Kết quả này cao hơn kết quả của các mô hình sử dụng từng thuật toán thành phần và cũng cao hơn kết quả của các nghiên cứu khác dùng cùng thuật toán và trên cùng tập dữ liệu.

**Từ khóa:** dấu vết WiFi, cường độ tín hiệu thu-RSS, hệ thống định vị trong nhà, học máy.

---

**Title:** An Efficient WiFi Indoor Positioning Technique using Combined Machine Learning

**Abstract:** In this study, we propose a learning model that integrates many machine learning models in two phases. The first phase combines the combination of k-Nearest Neighbor (KNN), Logistic Regression (LR) and Support Vector Machines (SVM) classification models to combine new data. The second phase uses a classifier on the new dataset. This model is applied during the model training phase. During the testing phase, the data collected from the mobile device is preprocessed, then the data is fed into a predictive model to estimate a specific floor location in the building. The proposal is installed and evaluated on the UJIIndoorLoc dataset with accurate prediction results of 98.73%. This result is higher than the results of models using each component algorithm and also higher than the results of other studies using the same algorithm and on the same dataset.

**Keywords:** WiFi fingerprinting, Received Signal Strength-RSS, indoor positioning system, machine learning.

---

## I. ĐẶT VẤN ĐỀ

Hệ thống định vị trong nhà (IPS) rất hữu ích trong thực tế. Trong các tòa nhà, IPS ngoài việc trợ giúp điều hướng trong cứu hộ, khẩn cấp nó còn hữu ích trong các môi trường như bảo tàng, bệnh viện, mega mall. Do đó, IPS đã trở thành một lĩnh vực nghiên cứu rộng rãi trong hơn 20 năm qua.

Các phương pháp định vị trong nhà được giới thiệu trong tài liệu [1]. Trong đó, phương pháp dựa trên cường độ tín hiệu (Received Signal Strength - RSS) của sóng radio được phát ra từ các điểm truy cập (Access point - AP) và thu được từ các điểm tham chiếu (Reference Point-RP) được sử dụng nhiều nhất và phổ biến trong các hệ thống định vị trong nhà. Có rất nhiều kỹ thuật định vị trong nhà dựa trên giá trị RSS, trong số đó fingerprinting là kỹ thuật được

dùng hết sức phổ biến [2, 3].

Có hai giai đoạn trong fingerprinting: giai đoạn huấn luyện và giai đoạn kiểm thử. Giai đoạn huấn luyện: thực hiện đo các chỉ số RSS của các AP từ các RP đã biết để xây dựng cơ sở dữ liệu fingerprinting. Giai đoạn kiểm thử, giá trị RSS được lấy mẫu theo thời gian thực tại vị trí chưa biết, giá trị này được so sánh với cơ sở dữ liệu fingerprinting để tìm ra vị trí định vị phù hợp nhất.

IPS dùng fingerprinting đối mặt với hai vấn đề chính. Thứ nhất, các vector RSS thu được có giá trị biến đổi bất thường, không ổn định dẫn đến số lượng dữ liệu lớn nhưng không chất lượng [4]. Để khắc phục vấn đề này một số phương pháp lọc và giảm kích thước dữ liệu như phương pháp phân tích thành phần chính (Principal

Component Analysis-PCA) [5] và phân tích biệt thức tuyến tính (Linear Discriminant Analysis-LDA) [6] đã được áp dụng nhằm tăng chất lượng tập dữ liệu. Thách thức thứ hai là chất lượng định vị cần có độ chính xác và hiệu quả cao. Nhiều thuật toán học máy như KNN, SVM, LR, Decision Tree(DT), Random Forest (RF), Light Gradient Boosting Machine (LGBM), Convolutions Neural Network (CNN), Deep Neural Networks (DNN), Recurrent Neural Network (RNN), ... đã được thử nghiệm nhằm nâng cao chất lượng định vị và đã đạt được nhiều kết quả đáng kể, tuy nhiên cho đến hiện tại chưa có thuật toán nào được cho là có thể áp dụng cho nhiều môi trường [7, 8].

Trong nghiên cứu này chúng tôi đề xuất phương pháp giảm kích thước dữ liệu và phương pháp định vị bằng mô hình mới, trong đó chúng tôi tích hợp các thuật toán học máy KNN [9, 10], SVM [11], LR [12] để dự báo tòa-tầng bằng bài toán phân lớp, các mô hình được thực nghiệm trên tập dữ liệu UJIIndoorLoc<sup>1</sup>[13]. Cụ thể những đóng góp của bài báo như sau:

- Tiền xử lý dữ liệu trong giai đoạn kiểm thử
- Đề xuất phương pháp giảm kích thước dữ liệu cho bài toán phân lớp ước lượng vị trí tòa nhà và tầng
- Đề xuất mô hình kết hợp các thuật toán học máy phân lớp áp dụng vào bài toán ước lượng vị trí tầng bằng tổ hợp các thuật toán phân lớp LR, KNN, SVM

Bài báo có cấu trúc như sau: Phần I đặt vấn đề, Phần II về các nghiên cứu liên quan, Phần III mô hình đề xuất. Phần IV thực nghiệm và kết quả mới, Phần V kết luận.

## II. CÁC NGHIÊN CỨU LIÊN QUAN

Các phương pháp học máy truyền thống được áp dụng trong bài toán định vị trong nhà dùng RSS của sóng WiFi được sử dụng rất nhiều. Có thể coi KNN là thuật toán học máy được sử dụng đầu tiên và mở ra thời kỳ áp dụng học máy vào bài toán định vị trong nhà. Năm 2000 nhóm nghiên cứu của Microsoft Research đã sử dụng KNN để phát triển hệ thống định vị có tên RADAR [9], kết quả KNN tốt hơn rất nhiều thuật toán fingerprinting. Trong [14] các tác giả đã sử dụng KNN kết hợp với thông tin vị trí người dùng trong quá khứ. Kết quả cho thấy phương pháp mới đạt hiệu suất định vị cao hơn 45% so với KNN. Các tác giả của nghiên cứu [15] kết hợp DNN và KNN. Kết quả độ sai số trung bình từ 1,39 đến 1,5m tùy vào giá trị K và so sánh với các phương pháp DT, KNN, DNN, SVM, và RF thì mô hình kết hợp có kết quả tốt hơn.

Giải pháp dựa trên thuật toán phân lớp SVM đã được đề xuất trong [16, 17]. Kết quả của các phương pháp đều cho thấy SVM cho độ chính xác cao hơn thuật toán

fingerprinting. Độ chính xác là 2m trong 77% và 98,75% trong 93,75% các trường hợp kiểm thử.

LR được các tác giả sử dụng trong [12] sau khi tối ưu dữ liệu, kết quả độ chính xác đạt 95.83%. Chenlu Xiang và các cộng sự dùng LR kết hợp với tối ưu dữ liệu ở môi trường phòng thí nghiệm tiêu chuẩn trong [18, 19] đều cho kết quả sai số định vị đạt 92cm.

### Các nghiên cứu sử dụng tập dữ liệu UJIIndoorLoc.

Trong [20] các tác giả dùng KNN phân cụm dự báo vị trí tòa nhà và KNN hồi quy để dự báo tọa độ và RF để dự báo tòa-tầng. Sau khi loại bỏ các AP chỉ xuất hiện dưới 2 lần, kết quả cho thấy RF dự đoán vị trí tòa nhà đúng 100%, dự đoán tầng bằng RF đạt kết quả lần lượt 97.95%, 90.87% and 95.86% cho ba tòa nhà, kết quả dự đoán kinh độ vĩ độ ở ba tòa nhà có giá trị (RMSE) trong khoảng 7.1 đến 12.8, Mean Absolute Error (MAE) trong khoảng 5 đến 9.

Trong [21] sử dụng cách kết hợp hệ thống COSELM (Constraint Online Sequential Extreme Learning Machine) với KNN thành hệ thống có tên AFARLS. Kết quả AFARLS có hiệu suất dự đoán tầng 95.41%, sai số vị trí 6.4m, trong khi KNN dự đoán tầng đạt 89.92%. Trong [22], các tác giả đã biến đổi tập dữ liệu huấn luyện theo cách riêng của mình và tạo ra tập dữ liệu thử nghiệm mới từ tập dữ liệu đã biến đổi, trong đó 520 đặc trưng chỉ còn lại 60 đặc trưng. Thử nghiệm được thực thi bởi 6 thuật toán, trong đó thuật toán KNN nhận diện tầng đạt 97%.

Áp dụng lượng tử hóa trước khi áp dụng KNN, SVM, RF trong [23], kết quả nghiên cứu có độ chính xác vị trí tương ứng với các thuật toán KNN, SVM, RF là 67.49%, 62,71% và 68.5%, thuật toán RF dùng để ước lượng vị trí tầng đạt độ chính xác 97%.

Như vậy các thuật toán KNN, SVM, LR đã được nhiều nhóm nghiên cứu sử dụng bằng nhiều cách khác nhau. Theo kiến thức tốt nhất mà chúng tôi có, hiện tại trên tập dữ liệu UJIIndoorLoc chưa có nghiên cứu nào kết hợp giữa các thuật toán như mô hình chúng tôi đề xuất. Do đó trong bài báo này chúng tôi tiến hành xây dựng mô hình kết hợp tổ hợp các thuật toán KNN, SVM, LR cho phần huấn luyện mô hình ở giai đoạn huấn luyện và thực nghiệm trên bộ dữ liệu UJIIndoorLoc nhằm nâng cao hiệu suất định vị.

## III. ĐỀ XUẤT MÔ HÌNH ĐỊNH VỊ

### 1. Mô tả tập dữ liệu

Tập dữ liệu UJIIndoorLoc[13] được nhiều nghiên cứu sử dụng [8], UJIIndoorLoc có số lượng RP là 933, các RP thuộc 3 tòa nhà cao 4 hoặc 5 tầng với tổng diện tích 108703m<sup>2</sup>. UJIIndoorLoc có 21049 mẫu, trong đó 19938 mẫu dùng cho huấn luyện và 1111 mẫu dùng cho kiểm thử. Mỗi mẫu gồm 529 đặc trưng, 520 đặc trưng đầu tiên là giá trị RSS thu được từ 520 AP, giá trị RSS nằm trong

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/ujiindoorloc>

khoảng từ -104dB đến 0dB, giá trị RSS=100 được đặt mặc định với các AP không phát hiện được khi lấy mẫu. Chín đặc trưng còn lại là các nhân gồm kinh độ (Longitude), vĩ độ (Latitude), số hiệu tòa nhà (BuildingID), tầng trong tòa nhà (Floor), số hiệu vùng lấy mẫu (SpaceID), vị trí tương quan xác định vùng lấy mẫu là trong hay ngoài các phòng (Relative Position), mã người thu thập dữ liệu (UserId), mã điện thoại thu thập dữ liệu (PhoneId) và thời gian thu thập (Timestamp).

## 2. Đề xuất kỹ thuật giảm kích thước dữ liệu cho bài toán phân lớp

Tập dữ liệu UJIIndoorLoc như mô tả bên trên là tập dữ liệu lớn, do đó cần tiến hành xử lý dữ liệu nhằm giảm thời gian tính toán và tăng chất lượng định vị.

Trong UJIIndoorLoc, mỗi tòa nhà là một lớp bao gồm nhiều dòng dữ liệu, một dòng dữ liệu gốc có dạng  $[RSS_1, RSS_2, \dots, RSS_{520}, \text{buildingID}]$ , trong đó 520 giá trị đầu tiên là các giá trị RSS đo được từ 520 AP, giá trị cuối cùng là số hiệu ba tòa nhà, các tòa nhà được đánh số từ 0 đến 2. Với mỗi tòa nhà chúng tôi lấy giá trị RSS trung bình của mỗi AP, kết quả thu được 520 giá trị trung bình. Kết quả thu được tập dữ liệu mới được thể hiện trong Phương trình (1) với 520 dữ liệu đầu là giá trị mới của các đặc trưng và cuối cùng là số hiệu tòa nhà.

$$\begin{pmatrix} \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 0 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 1 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 2 \end{pmatrix} \quad (1)$$

Bài toán phân lớp giá trị giữa các điểm càng chênh lệch lớn thì càng dễ phân lớp, chẳng hạn như tòa nhà được phân lớp tốt hơn nếu các đặc trưng có nhiều thông tin của tòa nhà. Do đó, chúng tôi tìm sự khác biệt tuyệt đối giữa các cặp tòa nhà 0-1, 0-2 và 1-2 bằng cách ghép chéo các cặp tương ứng trong tập dữ liệu (1), chúng tôi thu được ba mảng tương ứng với các cặp tòa nhà. Sắp xếp ba mảng thu được theo thứ tự giảm dần các đặc trưng và loại bỏ các đặc trưng có sự khác biệt thấp, trùng nhau. Kết quả mỗi mảng còn lại N đặc trưng có sự khác biệt cao nhất. Hợp nhất ba mảng lại thành một tập hợp đặc trưng có  $N*3$  giá trị cho mô hình phân lớp tòa nhà. Phương pháp giảm kích thước dữ liệu tòa nhà thể hiện trong Thuật toán 1.

Tập dữ liệu dùng phân lớp tầng được chúng tôi xử lý tương tự như tòa nhà. Một dòng dữ liệu gốc có dạng  $[RSS_1, RSS_2, \dots, RSS_{520}, \text{floor}]$ , trong đó 520 giá trị đầu tiên là các giá trị RSS đo được từ 520 AP, giá trị cuối cùng là số hiệu các tầng, các tầng được đánh số từ 0 đến 4, kết

### Thuật toán 1: Thuật toán giảm kích thước đặc trưng

```

1 Dữ liệu vào:
2  $\mathcal{B}_0 \leftarrow$ : các dòng dữ liệu của buildingID=0;
3  $\mathcal{B}_1 \leftarrow$ : các dòng dữ liệu của buildingID=1;
4  $\mathcal{B}_2 \leftarrow$ : các dòng dữ liệu của buildingID=2;
5 Dữ liệu ra:  $\mathcal{B}$ : Tập hợp các đặc trưng đã rút gọn.
6 begin
7   Bước 1:
8     1.1.  $\mathcal{BA}_0 \leftarrow$  Tập giá trị RSS trung bình của  $\mathcal{B}_0$ ;
9     1.2.  $\mathcal{BA}_1 \leftarrow$  Tập giá trị RSS trung bình của  $\mathcal{B}_1$ ;
10    1.3.  $\mathcal{BA}_2 \leftarrow$  Tập giá trị RSS trung bình của  $\mathcal{B}_2$ ;
11   Bước 2:
12     2.1.  $\mathcal{B}_{01} \leftarrow \mathcal{BA}_0 \cup \mathcal{BA}_1$ ;
13     2.2.  $\mathcal{B}_{02} \leftarrow \mathcal{BA}_0 \cup \mathcal{BA}_2$ ;
14     2.3.  $\mathcal{B}_{12} \leftarrow \mathcal{BA}_1 \cup \mathcal{BA}_2$ ;
15   Bước 3: Sắp xếp  $\mathcal{B}_{01}$ ,  $\mathcal{B}_{02}$  và  $\mathcal{B}_{12}$  theo thứ tự giảm dần;
16   Bước 4: Loại bỏ các đặc trưng có sự khác biệt thấp, trùng nhau;
17   Bước 5:  $\mathcal{B} \leftarrow \mathcal{B}_{01} \cup \mathcal{B}_{02} \cup \mathcal{B}_{12}$ ;
18 end

```

quả thể hiện theo Phương trình (2).

$$\begin{pmatrix} \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 0 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 1 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 2 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 3 \\ \overline{RSS_1} & \overline{RSS_2} & \dots & \overline{RSS_{520}} & 4 \end{pmatrix} \quad (2)$$

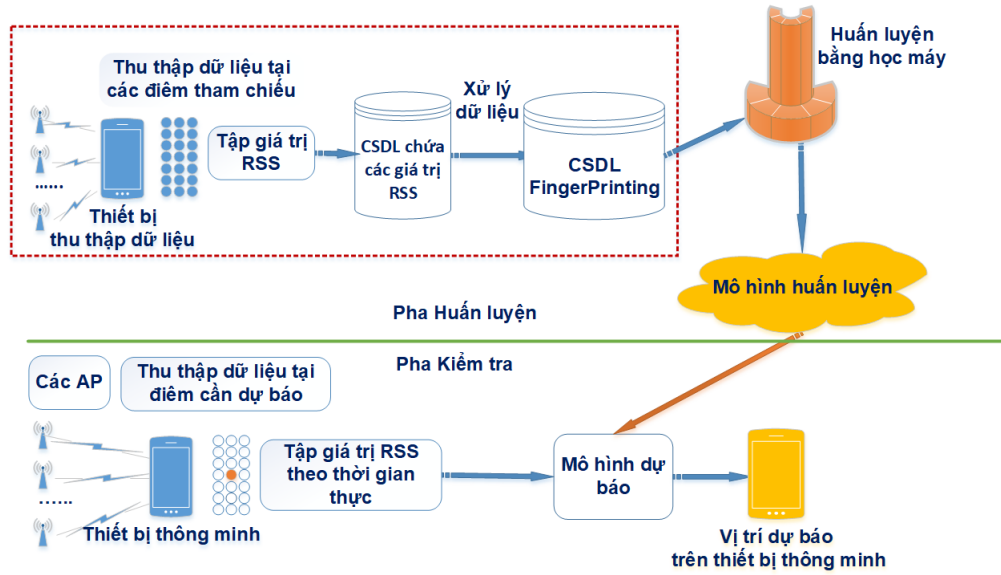
Bài toán phân lớp tầng tìm các đặc trưng nhiều thông tin tương tự, dữ liệu phân lớp tầng được xử lý tương tự như dữ liệu tòa nhà. Chúng tôi kết hợp các cặp tầng 0-1, 0-2, 0-3, 0-4, 1-2, 1-3, 1-4, 2-3, 2-4, 3-4 tạo thành 10 mảng. Sau khi loại bỏ các đặc trưng có độ chênh lệch thấp bằng phương pháp đã áp dụng cho tòa nhà, chúng tôi gộp các mảng lại để tạo thành tập dữ liệu đầu vào có  $10*N$  đặc trưng cho mô hình phân lớp tầng.

Cuối cùng, gộp các tập dữ liệu lại với nhau, chúng tôi thu được 13 lớp cho các tầng của các tòa nhà. Giá trị N được tính theo phương pháp heuristically nó ảnh hưởng trực tiếp đến hiệu suất của thuật toán.

### 3. Mô hình đề xuất

Mô hình cơ bản của bài toán định vị trong nhà bằng fingerprinting dùng sóng WiFi dựa trên học máy được thể hiện trong Hình 1. Trong mô hình, thuật toán học máy sẽ được áp dụng trên tập dữ liệu fingerprinting (tập dữ liệu huấn luyện) để đưa ra mô hình huấn luyện tại giai đoạn huấn luyện. Ở giai đoạn kiểm thử, tập RSS thời gian thực được đưa vào mô hình dự báo để tìm ra vị trí ước lượng theo mô hình huấn luyện.

Trong bài báo này chúng tôi đề xuất mô hình học máy mới, mô hình này dùng các thuật toán phân lớp nhằm ước



Hình 1. Mô hình định vị trong nhà cơ bản dựa trên học máy

lượng vị trí theo tòa-tầng. Khác với mô hình cơ bản, mô hình của chúng tôi tích hợp nhiều thuật toán và chia làm hai pha như Hình 2. Pha đầu tiên, tập dữ liệu fingerprinting được huấn luyện bằng tổ hợp các thuật toán phân lớp LR, KNN và SV. Pha một tạo ra tổ hợp dữ liệu mới, tổ hợp dữ liệu này là đầu vào cho thuật toán huấn luyện LR ở pha thứ hai. Như vậy, mô hình của chúng tôi huấn luyện tập dữ liệu theo hai pha liên tiếp thay vì một pha như mô hình cơ bản. Hình 3 mô tả chi tiết hoạt động của tổ hợp thuật toán theo hai pha. Trong đó  $\hat{Y}_1$ ,  $\hat{Y}_2$  và  $\hat{Y}_3$  là kết quả huấn luyện của ba thuật toán trong pha một,  $\hat{Y}_f$  là kết quả của pha hai và cũng là mô hình huấn luyện cuối.

## IV. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

### 1. Tiền xử lý dữ liệu

Chúng tôi phân tích dữ liệu UJIIndoorLoc và chúng tôi nhận thấy cần tiền xử lý dữ liệu như sau: thứ nhất số lượng các AP nhận được tại một RP nhỏ nhất là 0, lớn nhất là 51 và trung bình là xấp xỉ 18 AP, trong khi bài toán định vị trong nhà chỉ cần tối thiểu 3 AP là có thể xác định được vị trí và theo thống kê thì số RP có AP nhỏ hơn 6 (bao gồm cả các RP không có AP nào) là 389, do đó chúng tôi quyết định loại bỏ các RP có AP nhỏ hơn 6, điều này hầu như không ảnh hưởng đến chất lượng định vị nhưng lại tăng tốc độ tính toán. Thứ hai tại một RP có thể nhận được giá trị RSS từ AP của tòa nhà khác, các AP này không có tác dụng cho định vị thậm chí làm giảm chất lượng, do đó chúng tôi cũng loại bỏ các AP này. Tiền xử lý dữ liệu được áp dụng cho pha kiểm thử của mô hình phân lớp.

### 2. Các chỉ số đánh giá

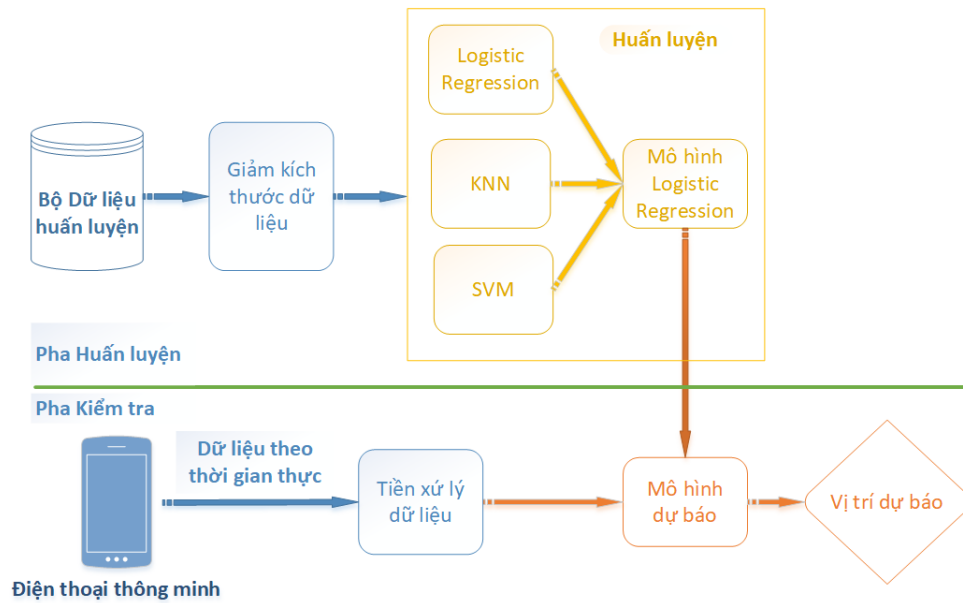
Bài toán phân lớp có các chỉ số đánh giá dựa trên ma trận nhầm lẫn (confusion matrix) để đếm được số lượng các giá trị gồm True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive; True Negative (TN): số lượng điểm của lớp negative được phân loại đúng là negative; False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive và False Negative (FN): số lượng điểm của lớp positive bị phân loại nhầm thành negative. Gọi T là tổng số các điểm của cả hai lớp khi đó các chỉ số đánh giá phân lớp bao gồm:

- Accuracy: Độ chính xác  $Accuracy = \frac{TP+TN}{T}$
- Precision: tỉ lệ số điểm true positive trong số những điểm được phân loại là positive.  $Precision = \frac{TP}{TP+FP}$
- Recall: tỉ lệ số điểm true positive trong số những điểm thực sự là positive.  $Recall = \frac{TP}{TP+FN}$
- F1-Score: trung bình điều hòa (harmonic mean) của các tiêu chí Precision và Recall, F1-Score đánh giá khách quan khả năng thực thi của mô hình học máy.  $F_1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$

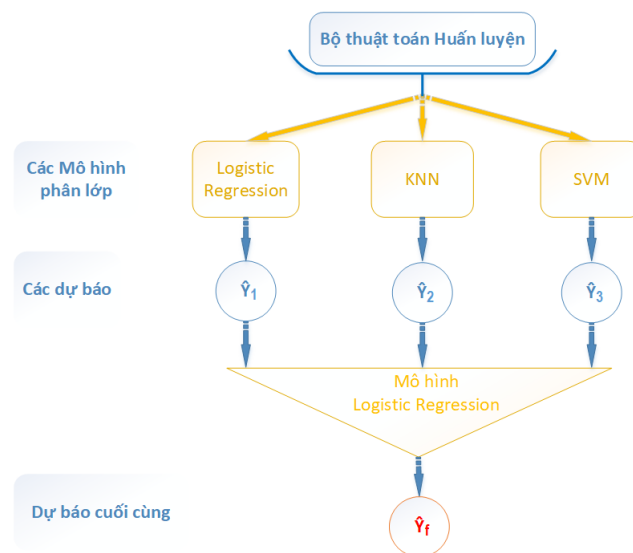
### 3. Kết quả thực nghiệm

Trong phần này chúng tôi cung cấp các kết quả thực nghiệm thể hiện hiệu quả của các thuật toán độc lập và phương pháp kết hợp mà chúng tôi đề xuất, kết quả cho thấy khi tích hợp các thuật toán kết quả tốt hơn khi sử dụng từng thuật toán độc lập

Mô hình kết hợp dự đoán tòa-tầng được chúng tôi thực hiện theo hai bước, đầu tiên chúng tôi thử nghiệm các mô hình dùng thuật toán phân lớp độc lập nhằm lựa chọn các



Hình 2. mô hình kết hợp phân loại tòa-tầng



Hình 3. Mô hình hoạt động tổ hợp thuật toán phân loại tòa-tầng

thuật toán tốt nhất. Các thuật toán chúng tôi thử nghiệm bao gồm KNN, SVM, LR, Naïve Bayes (NB), cây phân loại và hồi quy (Classification and Regression Tree-CART) và Linear Discriminant Analysis (LDA).

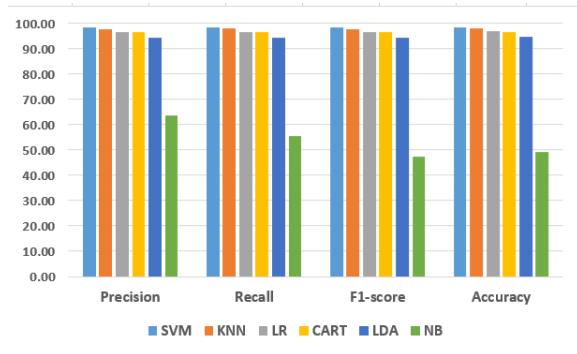
Các kết quả của các thuật toán độc lập được tổng hợp trong Bảng I, giá trị các chỉ số được so sánh ở biểu đồ trong Hình 4 và hình ảnh của Hình 5. Bảng I và các hình cho thấy tất cả các chỉ số của các thuật toán LR, KNN, SVM đều cho kết quả tốt hơn các thuật toán còn lại. Lưu ý, chỉ số của LR và CART rất gần nhau, với bài toán phân lớp chúng tôi chỉ chọn một, chỉ số của Recall của thuật

toán LR thấp hơn của thuật toán CART, tuy nhiên chỉ số F1-Score của LR lại tốt hơn CART, cho thấy mô hình của LR tốt hơn nên chúng tôi chọn LR. Dùng các kết quả này, chúng tôi chọn 3 thuật toán có kết quả tốt nhất LR, KNN và SVM vào trong mô hình kết hợp.

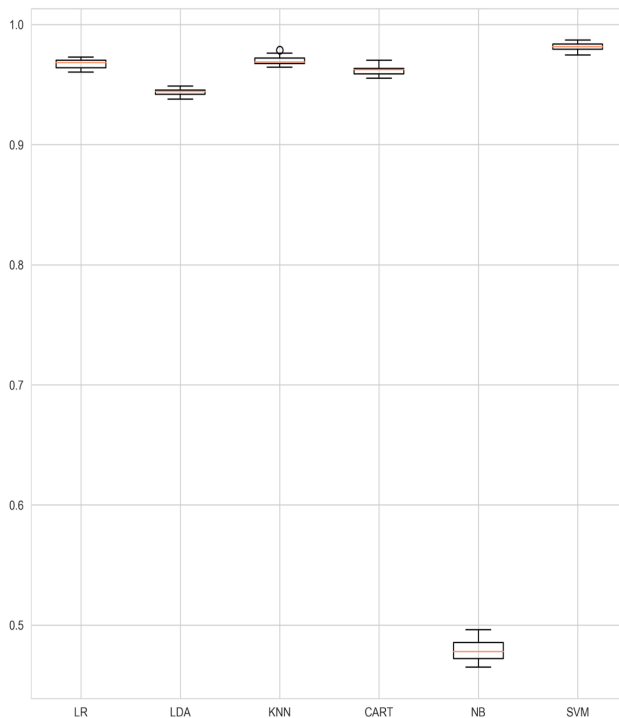
Bước hai, chúng tôi thử nghiệm mô hình kết hợp các thuật toán đã được chọn. Trong Bảng II là kết quả tổng hợp của các chỉ số Precision, Recall, F1-score và Accuracy của mô hình kết hợp. Kết quả của mô hình kết hợp so với các mô hình sử dụng các thuật toán độc lập được thể hiện ở biểu đồ trong Hình 6 cho thấy tất cả các chỉ số của mô

**Bảng I**  
**TỔNG HỢP KẾT QUẢ CÁC CHỈ SỐ**  
**CỦA CÁC THUẬT TOÁN ĐỘC LẬP**

|           | 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 |
| 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  |



Hình 4. Biểu đồ so sánh giá trị tổng hợp các chỉ số các mô hình phân lớp độc lập



Hình 5. Biểu đồ so sánh độ chính xác của các mô hình phân lớp độc lập

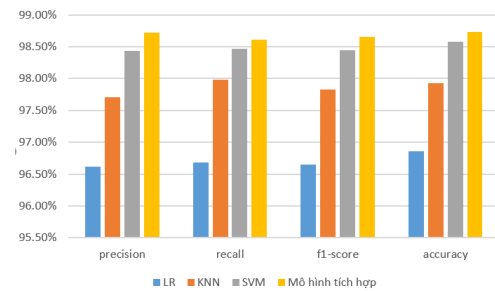
hình kết hợp có kết quả tốt hơn kết quả của tất cả các thuật

**Bảng II**  
**KẾT QUẢ ƯỚC LƯỢNG TÒA-TẦNG CỦA MÔ HÌNH KẾT HỢP**

|                    | Pre-<br>cision | Re-<br>call | F1-<br>Score | Accu-<br>racy | Time<br>(s) |
|--------------------|----------------|-------------|--------------|---------------|-------------|
| Mô hình<br>kết hợp | 98.71          | 98.61       | 98.66        | 98.73         | 99.31       |

toán LR, KNN và SVM.

Như vậy, mô hình kết hợp các thuật toán của chúng tôi đề xuất đạt hiệu suất và kết quả tốt hơn khi dùng các thuật toán độc lập.



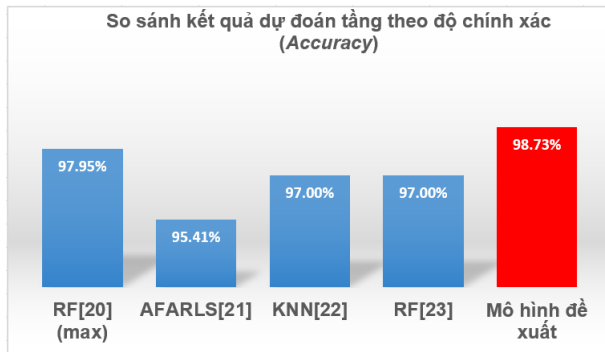
Hình 6. So sánh các chỉ số phân lớp giữa mô hình kết hợp và các mô hình độc lập

#### 4. So sánh kết quả với các nghiên cứu khác trên cùng tập dữ liệu UJIIndoorLoc

Hình 7 thể hiện kết quả dự đoán tòa-tầng của mô hình kết hợp mà chúng tôi đề xuất với các nghiên cứu khác, trong đó các nghiên cứu khác được thể hiện bằng tên thuật toán hoặc tên do nhóm nghiên cứu đặt cùng với số tham chiếu của nghiên cứu. Với nghiên cứu của chúng tôi, nhiều đơn vị đo được sử dụng để đánh giá kết hiệu suất mô hình cũng như độ chính xác, tuy nhiên, các nghiên cứu khác thông thường chỉ công bố độ chính xác, do đó, trong so sánh chúng tôi dùng chỉ số độ chính xác (Accuracy) để so sánh, ngoài ra với các công bố không sử dụng kết quả trung bình, chúng tôi sử dụng kết quả cao nhất mà nhóm công bố (ghi chú max). Kết quả cho thấy mô hình của chúng tôi dự đoán tốt hơn các mô hình khác sử dụng cùng các thuật toán.

#### V. KẾT LUẬN

Trong bài báo này, phương pháp giảm kích thước dữ liệu được đề xuất bằng cách phân tích theo cặp giữa các tòa nhà, giữa các tầng đã mang lại kết quả rất tốt cho các mô hình phân lớp độc lập và kết hợp. Kết quả mô hình phân lớp kết hợp có độ chính xác cao 98.73% và cao hơn so với kết quả của các mô hình dùng các thuật toán độc lập. Giá trị của các chỉ số Precision, Recall rất cao thể hiện hiệu



Hình 7. So sánh kết quả dự đoán tòa-tầng theo độ chính xác (Accuracy) của mô hình kết hợp với các nghiên cứu khác

suất cao của mô hình, giá trị F1-Score cao khẳng định độ ổn định của hệ thống, các kết quả này cho thấy mô hình của chúng tôi đã thành công và cho kết quả tốt. So sánh kết quả với các nghiên cứu khác trên cùng tập dữ liệu, mô hình cũng cho kết quả tốt hơn. Tuy nhiên, mô hình vẫn có nhược điểm là thời gian chạy cao do tích hợp nhiều thuật toán, tuy nhiên mô hình được áp dụng ở pha huấn luyện mô hình nên thời gian chạy lâu không phải là vấn đề lớn.

Mô hình kết hợp có rất nhiều hứa hẹn áp dụng cho các bài toán khác và chúng tôi có thể nâng cấp được mô hình khi thay thế bằng các thuật toán học sâu.

## TÀI LIỆU THAM KHẢO

- [1] S. Chan and G. Sohn, "Indoor localization using wi-fi based fingerprinting and trilateration techniques for lbs applications," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. 4, p. C26, 2012.
- [2] R. Harle, "A survey of indoor inertial positioning systems for pedestrians," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1281–1293, 2013.
- [3] C. Basri and A. El Khadimi, "Survey on indoor localization system and recent advances of wifi fingerprinting technique," in *2016 5th international conference on multimedia computing and systems (ICMCS)*. IEEE, 2016, pp. 253–259.
- [4] E. Laitinen, E. S. Lohan, J. Talvitie, and S. Shrestha, "Access point significance measures in wlan-based location," in *2012 9th Workshop on Positioning, Navigation and Communication*, 2012, pp. 24–29.
- [5] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, 2016.
- [6] D. Chu, L.-Z. Liao, M. K. Ng, and X. Wang, "Incremental linear discriminant analysis: A fast algorithm and comparisons," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 2716–2735, 2015.
- [7] A. Nessa, B. Adhikari, F. Hussain, and X. N. Fernando, "A survey of machine learning for indoor positioning," *IEEE Access*, vol. 8, pp. 214 945–214 965, 2020.
- [8] N. Singh, S. Choe, and R. Punmiya, "Machine learning based indoor localization using wi-fi rssi fingerprints: an overview," *IEEE Access*, 2021.
- [9] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proceedings IEEE INFOCOM 2000. Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064)*, vol. 2. Ieee, 2000, pp. 775–784.
- [10] M. Y. Umair, K. V. Ramana, and D. Yang, "An enhanced k-nearest neighbor algorithm for indoor positioning systems in a wlan," *2014 IEEE Computers, Communications and IT Applications Conference*, pp. 19–23, 2014.
- [11] C. Sujin and G. Kousalya, *Fingerprint-Based Support Vector Machine for Indoor Positioning System*, 01 2019, pp. 289–298.
- [12] Z. Peng, Y. Xie, D. Wang, and Z. Dong, "One-to-all regularized logistic regression-based classification for wifi indoor localization," *2016 IEEE 37th Sarnoff Symposium*, pp. 154–159, 2016.
- [13] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014, pp. 261–270.
- [14] M. T. Hoang, Y. Zhu, B. Yuen, T. Reese, X. Dong, T. Lu, R. Westendorp, and M. Xie, "A soft range limited k-nearest neighbors algorithm for indoor localization enhancement," *IEEE Sensors Journal*, vol. 18, no. 24, pp. 10 208–10 216, 2018.
- [15] P. Dai, Y. Yang, M. Wang, and R. Yan, "Combination of dnn and improved knn for indoor location fingerprinting," *Wireless Communications and Mobile Computing*, vol. 2019, 2019.
- [16] L. Zhang, Y. Li, Y. Gu, and W. Yang, "An efficient machine learning approach for indoor localization," *China Communications*, vol. 14, no. 11, pp. 141–150, 2017.
- [17] Y. Rezagui, L. Pei, X. Chen, F. Wen, and C. Han, "An efficient normalized rank based svm for room level indoor wifi localization with diverse devices," *Mobile Information Systems*, vol. 2017, pp. 1–19, 07 2017.
- [18] C. Xiang, Z. Zhang, S. Zhang, S. Xu, S. Cao, and V. K. N. Lau, "Robust sub-meter level indoor localization - a logistic regression approach," *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2019.
- [19] C. Xiang, S. Zhang, S. Xu, X. Chen, S. Cao, G. C. Alexandropoulos, and V. K. N. Lau, "Robust sub-meter level indoor localization with a single wifi access point—regression versus classification," *IEEE Access*, vol. 7, pp. 146 309–146 321, 2019.
- [20] P. R. Shivam Wadhwa and R. Kaushik, "Machine learning based indoor localization using wi-fi fingerprinting," *International Journal of Recent Technology and Engineering*, 2019.
- [21] H. Gan, M. H. B. M. Khir, G. Witjaksono Bin Djaswadi, and N. Ramli, "A hybrid model based on constraint oselm, adaptive weighted src and knn for large-scale indoor localization," *IEEE Access*, vol. 7, pp. 6971–6989, 2019.
- [22] G. H. Apostolo, I. G. B. Sampaio, and J. Viterbo, "Feature selection on database optimization for wi-fi fingerprint indoor positioning," *Procedia Computer Science*, vol. 159, pp. 251–260, 2019, knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.
- [23] W. Charoenruangkit, S. Saejun, R. Jongfungfeuang, and K. Multhonggad, "Position quantization approach with multi-class classification for wi-fi indoor positioning system," in *2018 International Conference on Information Technology (InCIT)*, 2018, pp. 1–5.

## SƠ LƯỢC VỀ TÁC GIẢ



**Ngô Văn Bình** Nghiên cứu sinh tại Học Viện Khoa học Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Nhận bằng Cử nhân khoa học ngành Tin học năm 1999 tại Trường Đại học Khoa học Tự nhiên- Đại học Quốc gia Hà Nội. Nhận bằng Thạc sĩ năm 2004 tại Khoa Công nghệ - Đại học Quốc gia Hà Nội.

Hiện nay đang giảng dạy tại Đại học FPT.

Lĩnh vực nghiên cứu: Định vị trong nhà, học máy.

Email: binhnv11@fe.edu.vn



**Vũ Văn Hiệu** Nhận bằng tiến sĩ Cơ sở toán học cho tin học tại Học Viện Khoa học Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam năm 2017. Nhận bằng thạc sĩ Khoa học máy tính vào năm 2007 và Kỹ sư Công nghệ thông tin vào năm 2003.

Hiện nay đang nghiên cứu tại Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Lĩnh vực nghiên cứu: Thị giác máy tính, khoa học dữ liệu.

Email: vvhiệu@ioit.ac.vn