JTE HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Predicting Marathon Finishing Times Using Ensemble Learning: An Empirical Study on Boston Marathon Data

Anh Khoa Mai, Ha Quynh Giao Nguyen, Cong Manh Hoang, Thi Ngoc Phuong Truong*
Ho Chi Minh City University of Technology and Education, Vietnam

*Corresponding author. *Email: phuongttn@hcmute.edu.vn*

ARTICLE INFO

Received: 10/06/2025
Revised: 01/07/2025
Accepted: 15/08/2025
Published: 28/11/2025

KEYWORDS

Ensemble Learning;
Marathon Prediction;
Boston Marathon;
Machine learning;
Performance forecasting.

ABSTRACT

This study proposes an ensemble machine learning model to predict marathon finishing times, using empirical data from the Boston Marathon spanning 2015–2017. After thorough preprocessing and feature engineering—including intermediate checkpoint times (5K, 10K, Half Marathon), age, gender, nationality, and year of participation—six models were implemented and evaluated: K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Case-Based Reasoning (CBR), a prior benchmark model (FA-PP-R-ML), Long Short-Term Memory (LSTM), and a novel ensemble model combining Linear Regression, Random Forest, and MLPRegressor via a meta-learning approach. Experimental results on the test set demonstrate that the proposed ensemble model achieved the highest predictive performance, with a Mean Absolute Error (MAE) of 7.32 minutes, Root Mean Squared Error (RMSE) of 11.06 minutes, and R^2 score of 0.928—outperforming all baseline models in both accuracy and robustness. Visualization techniques such as scatter plots and boxplots further confirmed the model's high agreement between predicted and actual values. Nevertheless, the study acknowledges several limitations, including a constrained dataset limited to three years of a single event, a narrow scope of model comparison, simplifications in algorithmic assumptions, and limited hyperparameter tuning. Future work should explore more diverse datasets, incorporate exogenous factors (e.g., weather, elevation), adopt advanced modeling techniques such as attention mechanisms, graph-based learning, or AutoML, and enhance model interpretability to support real-world applications in athlete coaching and performance forecasting.

Dự đoán thời gian hoàn thành Marathon bằng mô hình học máy tổng hợp: Một nghiên cứu thực nghiệm trên dữ liệu Boston Marathon

Mai Anh Khoa, Nguyễn Hà Quỳnh Giao, Hoàng Công Mạnh, Trương Thị Ngọc Phượng*
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh, Việt Nam

*Tác giả liên hệ. *Email: phuongttn@hcmute.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 10/06/2025
Ngày hoàn thiện: 01/07/2025
Ngày chấp nhận đăng: 15/08/2025
Ngày đăng: 28/11/2025

TỪ KHÓA

Học kết hợp;
Dự đoán Marathon;
Marathon ở Boston;

TÓM TẮT

Nghiên cứu này đề xuất một mô hình học máy tổng hợp (ensemble) nhằm dự đoán thời gian hoàn thành cuộc thi Boston Marathon dựa trên dữ liệu thực nghiệm từ các năm 2015–2017. Sau khi tiến hành tiền xử lý và xây dựng đặc trưng bao gồm thời gian tại các mốc (5K, 10K, Half), tuổi, giới tính, quốc tịch và năm thi đấu, nhóm tác giả đã triển khai sáu mô hình: KNN, ANN, CBR, FA-PP-R-ML, LSTM và một mô hình Ensemble mới kết hợp Linear Regression, Random Forest và MLPRegressor thông qua một meta-model. Kết quả thực nghiệm trên tập kiểm tra cho thấy mô hình Ensemble đạt hiệu năng vượt trội với MAE = 7,32 phút, RMSE = 11,06 phút và R^2 = 0,928, vượt trội so với các mô hình còn lại cả về độ chính xác và tính ổn định. Các biểu đồ trực quan như scatter plot và boxplot cũng cho

JTE
HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Học máy;

Dự đoán hiệu suất.

thấy sự phù hợp cao giữa giá trị dự đoán và thực tế. Tuy nhiên, nghiên cứu vẫn còn một số hạn chế về quy mô dữ liệu, phạm vi so sánh mô hình, giả định đơn giản hóa trong thiết kế thuật toán và mức độ tinh chỉnh tham số còn cơ bản. Từ đó, nghiên cứu đề xuất mở rộng đánh giá trên các bộ dữ liệu đa dạng hơn, tích hợp thêm dữ liệu ngoại sinh và áp dụng các kỹ thuật hiện đại như attention, học sâu đồ thị (graph-based learning) và AutoML, đồng thời tăng cường khả năng diễn giải nhằm hướng đến ứng dụng thực tiễn trong huấn luyện thể thao và dự báo thi đấu.

Doi: https://doi.org/10.54644/jte.2025.1924

1. Giới thiệu

Dự đoán thành tích trong thi đấu thể thao đang ngày càng trở nên quan trọng, đặc biệt ở các môn yêu cầu sức bền như marathon [1]. Việc ước lượng chính xác thời gian hoàn thành giúp vận động viên (VĐV) và huấn luyện viên (HLV) xây dựng được lộ trình luyện tập hợp lý, tối ưu hóa chiến lược thi đấu, đồng thời đánh giá tiến bộ trong quá trình huấn luyện.

Trong những năm gần đây, các nghiên cứu sử dụng các phương pháp truyền thống như hồi quy tuyến tính, K-Nearest Neighbors (KNN), cũng như các phương pháp học sâu như mạng neural nhân tạo (ANN) và mạng ghi nhớ dài hạn (LSTM) đã cho thấy những kết quả đáng khích lệ. Tuy nhiên, các nghiên cứu này vẫn còn tồn tại những hạn chế nhất định: phạm vi dữ liệu còn hạn chế về mặt không gian và thời gian, ít khai thác các đặc trưng trung gian như thời gian tại các mốc 5K, 10K, Half Marathon; đồng thời, hiệu suất của các mô hình đơn lẻ thường bị ảnh hưởng bởi nhiễu dữ liệu [2] hoặc cấu trúc phi tuyến phức tạp chưa được học đầy đủ. Ngoài ra, khả năng tổng quát hóa và độ ổn định của các mô hình hiện tại vẫn chưa được đánh giá toàn diện trên các tập kiểm tra độc lập.

Nhằm khắc phục các hạn chế nêu trên, nghiên cứu này đề xuất một mô hình học máy tổng hợp (ensemble learning) kết hợp ba thuật toán gồm Linear Regression, Random Forest và MLPRegressor thông qua một meta-model học tổng hợp. Dữ liệu được thu thập từ ba mùa giải Boston Marathon (2015–2017) [3], được chuẩn hóa và mã hóa theo hướng tối ưu cho các thuật toán học máy. Mục tiêu của nghiên cứu là xây dựng một mô hình có độ chính xác cao, khả năng khái quát hóa tốt và ổn định với dữ liệu thực tế, từ đó mở rộng khả năng ứng dụng trong các hệ thống tư vấn chiến lược huấn luyện và dự báo thành tích thi đấu cá nhân.

2. Khảo sát các nghiên cứu trước đó

Trong những năm gần đây, việc áp dụng trí tuệ nhân tạo (AI) và học máy vào lĩnh vực thể thao nói chung và chạy marathon nói riêng đã thu hút sự quan tâm đáng kể từ cộng đồng nghiên cứu. Nhiều công trình đã tập trung vào việc dự đoán hiệu suất vận động viên và đề xuất các kế hoạch tập luyện cá nhân hóa [4]. Tuy nhiên, phần lớn các nghiên cứu vẫn tồn tại một số hạn chế nhất định về mặt ứng dụng thực tế, khả năng mở rộng và tương tác người dùng. Dưới đây là tổng quan các hướng tiếp cận tiêu biểu và những khoảng trống còn lại trong lĩnh vực này.

Một nghiên cứu được công bố vào năm 2023 [4] đã so sánh hiệu quả giữa mạng neural nhân tạo (Artificial Neural Networks – ANN) và thuật toán k-nearest neighbors (KNN) trong việc dự đoán thời gian hoàn thành marathon. Dữ liệu từ 820 vận động viên bao gồm các đặc trưng như thời gian chạy 10 km, chỉ số BMI, tuổi và giới tính. Kết quả cho thấy KNN vượt trội hơn với sai số tuyệt đối trung bình (MAE) là 2,4%, so với 5,6% của ANN, đồng thời đạt hệ số tương quan cao ($r > 0,90$ và $p < 0,001$). Tuy vậy, nghiên cứu chưa đề cập đến việc tích hợp mô hình vào các ứng dụng thực tiễn, cũng như chưa xem xét đến yếu tố tương tác người dùng.

Một hướng tiếp cận khác là sử dụng Case-Based Reasoning (CBR) năm 2020 [5], được triển khai trong một nghiên cứu của Đại học Dublin (UCD Research Repository). Nghiên cứu này tận dụng dữ liệu từ hơn 21000 vận động viên và 1,5 triệu phiên tập luyện để xây dựng hệ thống dự đoán và đề xuất kế hoạch tập luyện phù hợp cho vận động viên nghiệp dư. Mặc dù có tiềm năng triển khai trên các nền

tảng như Strava, nghiên cứu chưa cung cấp thông tin chi tiết về độ chính xác của mô hình hoặc phản hồi từ người dùng thực tế.

Một nghiên cứu khác đã phát triển phương pháp FA-PP-R-ML (Factor Analysis – Probabilities Prediction – Ranking Machine Learning) năm 2024 [6] nhằm xác định các yếu tố ảnh hưởng đến hiệu suất marathon và dự đoán kết quả thi đấu. Mô hình cho kết quả chính xác với sai số ±0,1 giờ. Tuy nhiên, tương tự các nghiên cứu trước, công trình này không đề cập đến việc triển khai thành ứng dụng hoặc khả năng tương tác với người dùng cuối.

Nghiên cứu của Muijlwijk và cộng sự năm 2024 [7] nhấn mạnh tầm quan trọng của việc tích hợp yếu tố tương tác giữa huấn luyện viên và hệ thống AI. Nghiên cứu được thực hiện với 71 huấn luyện viên cho thấy rằng khi người dùng có thể điều chỉnh các yếu tố đầu vào, độ chính xác của mô hình và mức độ tin tưởng của người dùng đều được cải thiện. Tuy nhiên, quy mô nghiên cứu còn hạn chế, và chưa mở rộng đến nhóm vận động viên ở các trình độ khác nhau.

Một xu hướng mới là sử dụng mạng neural hồi tiếp Long Short-Term Memory (LSTM) năm 2024 [8] để xử lý dữ liệu chuỗi thời gian như tốc độ, nhịp tim và tần suất bước chân. Một nghiên cứu trên 50 vận động viên đã cho kết quả ấn tượng với hệ số xác định (R^2) đạt 0,92 cho tốc độ và 0,88 cho nhịp tim, cùng với MAE chỉ 0,12 m/s cho tốc độ. Mặc dù vậy, nghiên cứu vẫn còn hạn chế trong việc tích hợp kết quả dự đoán vào kế hoạch tập luyện dài hạn hoặc tương tác với huấn luyện viên.

Một nghiên cứu tiêu biểu khác của El-Kassabi et al. (2020) [9] đã đề xuất phương pháp học sâu để dự báo hiệu suất của các vận động viên trong các giải đấu thể thao. Nghiên cứu sử dụng mạng neural để xử lý các đặc trưng liên quan đến lịch sử thi đấu và điều kiện thi đấu nhằm đưa ra dự đoán cá nhân hóa. Mặc dù mô hình cho thấy khả năng học phi tuyến hiệu quả, nhưng chưa có đánh giá toàn diện về khả năng tổng quát hóa hoặc mở rộng mô hình sang các môn thể thao khác nhau.

Từ tổng quan các công trình kể trên, có thể nhận thấy một số khoảng trống nổi bật trong lĩnh vực nghiên cứu dự đoán hiệu suất chạy marathon bằng trí tuệ nhân tạo. Các thuật toán học máy như Linear Regression, KNN, ANN, LSTM hay Case-Based Reasoning để dự đoán thời gian hoàn thành cuộc thi marathon. Tuy đạt được những kết quả khả quan, phần lớn các công trình vẫn sử dụng mô hình đơn lẻ với các hạn chế nhất định: mô hình tuyến tính khó học được quan hệ phi tuyến phức tạp; mạng nơ-ron dễ overfitting; các phương pháp "lazy learning" như KNN hay CBR thiếu khả năng tổng quát hóa; trong khi các mô hình phân tích yếu tố và hồi quy tuyến tính dù đơn giản nhưng thiếu ổn định khi mở rộng quy mô dữ liệu.

Trước thực trạng đó, mô hình học máy tổng hợp (ensemble learning) nổi lên như một giải pháp hiệu quả nhờ khả năng kết hợp nhiều thuật toán bổ trợ nhau [10]. Cụ thể, việc phối hợp Linear Regression (học tuyến tính), Random Forest (xử lý phi tuyến và kháng nhiễu tốt) và MLPRegressor (khai thác các quan hệ phức tạp bằng mạng học sâu) trong một khung học tổng hợp giúp mô hình tận dụng ưu điểm của từng thành phần, giảm sai số và tăng độ ổn định. Hơn nữa, meta-model tổng hợp đầu ra của các mô hình con giúp tối ưu hóa kết quả cuối cùng và cải thiện khả năng khái quát hóa [11]. Đây chính là hướng tiếp cận mà nghiên cứu này đề xuất nhằm nâng cao hiệu quả dự đoán và khả năng ứng dụng thực tiễn trong lĩnh vực phân tích hiệu suất marathon.

3. Phương pháp nghiên cứu

3.1. Cách tiếp cận nghiên cứu

Nghiên cứu này sử dụng phương pháp thực nghiệm kết hợp với mô phỏng dữ liệu để so sánh hiệu quả của các thuật toán dự đoán và tối ưu hóa tốc độ trong chạy marathon. Cụ thể, các thuật toán bao gồm KNN, ANN, CBR, LSTM và mô hình tổ hợp Ensemble Learning được triển khai và đánh giá trên cùng một tập dữ liệu được chuẩn hóa (bao gồm các mốc thời gian chia đoạn - split times, tuổi và giới tính). Việc triển khai đa thuật toán cho phép đánh giá hiệu năng tương đối giữa các mô hình học máy truyền thống và các mô hình học sâu (deep learning), từ đó làm cơ sở để xây dựng một mô hình tối ưu hóa phù hợp hơn cho mục tiêu nghiên cứu.

Việc lựa chọn kết hợp các thuật toán truyền thống và hiện đại giúp khai thác được thế mạnh riêng của từng phương pháp: KNN [4] đơn giản và dễ hiểu, phù hợp với dữ liệu có tính gần nhau; CBR [5]

tận dụng kiến thức từ các trường hợp tương tự trong quá khứ; ANN [4] và LSTM [8] có khả năng học các quan hệ phi tuyến phức tạp và mô hình hóa chuỗi thời gian; trong khi đó, mô hình Ensemble kết hợp các đặc điểm mạnh của nhiều thuật toán để tăng độ chính xác và độ ổn định của dự đoán.

Trong các nghiên cứu trước, phần lớn chỉ lựa chọn một hoặc hai thuật toán để triển khai, thường là Linear Regression hoặc Random Forest do tính đơn giản và hiệu quả ban đầu. Một số nghiên cứu gần đây bắt đầu ứng dụng LSTM [8] nhưng chưa có sự so sánh có hệ thống giữa các mô hình. Ngoài ra, nhiều nghiên cứu chỉ dừng lại ở việc dự đoán thành tích, mà chưa tiếp cận khía cạnh tối ưu hóa chiến lược phân bố tốc độ cho từng vận động viên cụ thể.

Khác với các hướng tiếp cận trên, nghiên cứu này tập trung vào việc đánh giá định lượng độ chính xác dự đoán và khả năng thích nghi cá nhân của từng mô hình. Đồng thời, nó tích hợp khả năng dự đoán với một thuật toán tổng hợp có khả năng học từ nhiều mô hình khác nhau, từ đó đưa ra gợi ý chiến lược mang tính cá nhân hóa cao hơn, mở rộng hướng ứng dụng từ chỉ "dự đoán" sang "gợi ý chiến lược".

3.2. Dữ liệu nghiên cứu

3.2.1. Nguồn dữ liệu (datasets)

Dữ liệu được thu thập từ nền tảng **Kaggle**, cụ thể là bộ dữ liệu có tên *Finishers Boston Marathon 2015, 2016 & 2017* do tác giả *Rojour* đăng tải [3]. Bộ dữ liệu này được trích xuất từ trang web chính thức của giải Boston Marathon thông qua kỹ thuật web scraping với mã nguồn trích xuất và các tệp liên quan cũng được công khai trên GitHub của tác giả tại [12] cho phép kiểm chứng quy trình và tính xác thực của dữ liệu.

Giải Boston Marathon là một trong những giải chạy bộ lâu đời và có tính cạnh tranh cao tại Hoa Kỳ, với phần lớn vận động viên phải đạt chuẩn thành tích nhất định mới đủ điều kiện tham dự. Vì vậy, dữ liệu thu được có độ tin cậy cao và mang tính đại diện cho các nghiên cứu liên quan đến hiệu suất thi đấu marathon.

Sau khi tổng hợp, tập dữ liệu bao gồm hàng chục nghìn bản ghi của ba năm liên tiếp (2015–2017), với các trường thông tin chính như: thời gian tại các mốc quãng đường (5K, 10K, 15K, 20K, Half, 25K, 30K, 35K, 40K), nhịp độ trung bình (pace, tính bằng phút/km), thời gian hoàn thành chính thức (official time – được chọn làm biến mục tiêu), tuổi và giới tính của vận động viên (được mã hóa nhị phân: 1 = nam, 2 = nữ). Các trường phụ như tên, số báo danh (Bib), quốc tịch,… sẽ được xử lý và lọc bỏ trong bước tiền xử lý dữ liệu để đảm bảo tính nhất quán và phù hợp với mục tiêu nghiên cứu.

3.2.2. Tiền xử lý dữ liệu

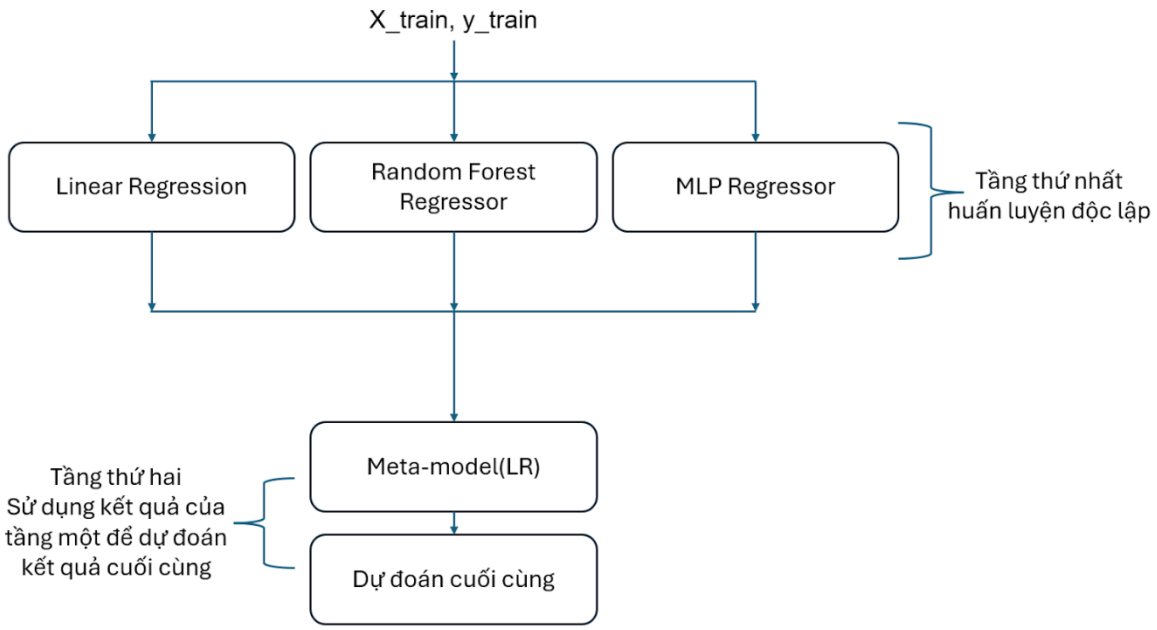
Tập dữ liệu sử dụng gồm kết quả Boston Marathon các năm 2015–2017 [3], được hợp nhất thành một tập duy nhất và bổ sung cột năm để phân biệt. Các trường dữ liệu không rõ nghĩa hoặc thiếu thông tin được loại bỏ, đồng thời chỉ giữ lại các mốc 5 km, 10 km và nửa chặng nhằm tránh mô hình đoán trước kết quả từ mốc 40 km. Các bản ghi thiếu dữ liệu quan trọng (tuổi, giới tính, thời gian hoàn thành, các mốc 5K, 10K, Half) bị loại bỏ; giới tính được mã hóa nhị phân, quốc tịch được one-hot encoding, và dữ liệu thời gian chuyển đổi sang số phút. Các đặc trưng đầu vào gồm: Age, Gender, Year, 5K_Min, 10K_Min, Half_Min và Country. Tập dữ liệu được chia theo tỷ lệ 80:20 cho huấn luyện và kiểm thử, sau đó chuẩn hóa bằng z-score (StandardScaler).

3.3. Xây dựng mô hình đề xuất

Trong bài nghiên cứu này, nhóm đề xuất một thuật toán tổ hợp theo hướng stacking ensemble learning [13] nhằm kết hợp ưu điểm của nhiều thuật toán để nâng cao hiệu suất của bài toán. Mô hình sẽ là sự kết hợp của ba thuật toán học máy bao gồm *LinearRegression*, *RandomForestRegressor* và *MLPRegressor*, kết hợp với meta-model (mô hình tầng thứ 2) đơn giản *LinearRegression* để đưa ra dự đoán cuối cùng. Việc áp dụng mô hình tổng hợp nhằm mục đích tận dụng được ưu điểm riêng của từng thuật toán, tạo nên một hệ thống dự đoán cân bằng, chính xác và ổn định hơn so với việc sử dụng từng mô hình đơn lẻ.

Khác với các mô hình trước đây như *ANN, KNN, CBR,* hay *LSTM* (tham khảo các bài công bố mô hình gốc [14], [15], [16], [17]) mô hình mà nhóm đề xuất không chỉ phụ thuộc vào một hướng tiếp cận duy nhất mà còn kết hợp các đặc tính khác nhau của từng mô hình như khả năng nắm bắt mối quan hệ tuyến tính của *LinearRegression,* khả năng kháng nhiễu và học tốt dữ liệu phi tuyến của *RandomForestRegressor,* và khả năng học được các biểu diễn phức tạp nhờ vào cấu trúc mạng neural của *MLPRegressor (hình 1).* Sự kết hợp này được kỳ vọng sẽ khắc phục được hạn chế của từng thuật toán đơn lẻ, giảm overfitting và tăng tính tổng quát của mô hình. Ngoài ra, nhóm kỳ vọng mô hình sẽ cải thiện độ chính xác toàn diện, đặc biệt là khi làm việc với dữ liệu vừa có đặc trưng tuần tự, vừa có đặc trưng tĩnh.

Kiến trúc mô hình bao gồm hai tầng. Tầng thứ nhất gồm ba thuật toán học máy *LinearRegression, RandomForestRegressor* và *MLPRegressor* được huấn luyện độc lập để tạo ra các dự đoán sơ cấp. Tầng thứ hai (meta-model) đảm nhiệm vai trò nhận dự đoán từ tầng một để làm đầu vào sau đó huấn luyện mô hình *LinearRegression* khác và cho ra kết quả dự đoán cuối cùng. Tham khảo *hình 1.*



Hình 1. *Mô hình Học kết hợp 2 tầng*

3.4. Phương pháp tính hệ số hồi quy trong mô hình Linear Regression

Trong kiến trúc mô hình hai tầng được đề xuất, thuật toán *LinearRegression* được sử dụng ở cả tầng cơ sở (base learner) dùng để huấn luyện độc lập và tầng tổng hợp (meta-model) dùng để cho ra dự đoán cuối cùng. Ở cả hai tầng này, quá trình huấn luyện đều nhằm mục tiêu tìm ra các hệ số hồi quy tối ưu để mô hình hóa mối quan hệ giữa biến đầu vào và biến mục tiêu. Các hệ số hồi quy (trọng số) này được tính toán theo phương pháp bình phương tối thiểu thông thường (*Ordinary Least Squares – OLS)* [18].

3.4.1. Ở tầng cơ sở

Tại tầng này, mô hình Linear Regression được huấn luyện trên tập dữ liệu ban đầu với các đặc trưng đầu vào gốc. Mục tiêu là tìm các hệ số hồi quy $w_1, w_2, \dots, w\_n$ và hệ số chệch b sao cho hàm dự đoán:

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \quad (1)$$

tiệm cận tốt nhất theo giá trị thực tế y, theo tiêu chí tối thiểu hóa tổng sai số bình phương giữa $\hat{y}$ và y. Các hệ số được tính theo công thức ma trận:

$$w = (X^T X)^{-1} X^T y \quad (2)$$

Trong đó, X là ma trận đặc trưng đầu vào, y là vector biến mục tiêu, và w là vector hệ số hồi quy.

3.4.2. Ở tầng tổng hợp

Tầng tổng hợp đóng vai trò như một bộ tổng hợp dự đoán (meta-model), trong đó đầu vào không còn là đặc trưng ban đầu mà là các dự đoán sơ cấp từ các mô hình ở tầng cơ sơ (bao gồm *LinearRegression, RandomForestRegressor* và *MLPRegressor).* Ba đầu vào này được ký hiệu lần lượt là y_{lr}, y_{rf}, y_{nn}. Mô hình *LinearRegression* trong tầng này học một hàm tuyến tính có dạng:

$$\hat{y} = a_1 y_{lr} + a_2 y_{rf} + a_3 y_{nn} + b \quad (3)$$

Trong đó:

- a_1, a_2, a_3*: là các trọng số phản ánh mức độ đóng góp của từng mô hình cơ sở.*
- b*: là hệ số chệch.*

Tương tự tầng đầu tiên, các trọng số này được tính bằng phương pháp OLS với công thức:

$$\alpha = (Z^T Z)^{-1} Z^T y \quad (4)$$

Với Z là ma trận đầu vào mới gồm các cột tương ứng với dự đoán của mô hình cơ sở.

Việc sử dụng *LinearRegression* ở tầng meta cho phép mô hình học được sự kết hợp tuyến tính tối ưu giữa các mô hình cơ sở, từ đó cải thiện độ chính xác của dự đoán cuối cùng. Các trọng số thu được còn mang ý nghĩa định lượng trong việc đánh giá tầm quan trọng tương đối của từng mô hình thành phần.

3.5. Huấn luyện và đánh giá

Mô hình sử dụng chiến lược học stacking gồm ba thuật toán cơ sở *LinearRegression, RandomForestRegressor* và *MLPRegressor*. Dữ liệu được chuẩn hóa bằng kỹ thuật chuẩn hóa z-score để đảm bảo các đặc trưng có cùng thang đo, sau đó được huấn luyện trên tập dữ liệu huấn luyện. Tham số huấn luyện cụ thể là:

- *LinearRegression* sử dụng cấu hình mặc định của thư viện *sklearn.*
- *RandomForestRegressor* với số lượng cây quyết định là (*n_estimators=100*) và *random_state=42* để đảm bảo khả năng tái kiểm tra.
- MLPRegressor được cấu hình với một lớp ẩn gồm 100 nút (*hidden_layer_sizes=(100,)*), số lần lặp tối đa là 1000 (*max_iter=1000*) nhằm đảm bảo mô hình có thể hội tụ.

Sau khi các thuật trên hoàn thành huấn luyện, ta sẽ thu được các dự đoán sơ cấp trên tập kiểm tra. Sau đó, các dự đoán này được kết hợp thành một đặc trưng mới (*X_meta)* để huấn luyện mô hình meta là *LinearRegression* nhằm tạo ra đầu ra dự đoán cuối cùng.

Hiệu năng của mô hình được đánh giá qua các tiêu chí:

- *MAE (Mean Absolute Error)* – sai số tuyệt đối trung bình, thể hiện độ lệch trung bình giữa dự đoán và thực tế.
- *RMSE (Root Mean Squared Error)* – sai số căn bậc hai trung bình, nhấn mạnh hơn vào các sai số lớn.
- *R^2 (R-squared)* – hệ số xác định, phản ánh mức độ mô hình giải thích được phương sai trong dữ liệu.

Các chỉ số MAE, RMSE và R^2 được sử dụng để đánh giá độ chính xác, tính ổn định và khả năng khái quát hóa của mô hình. Tất cả các thuật toán đều được huấn luyện trên cùng một tập dữ liệu, áp dụng cùng phương pháp chuẩn hóa và đánh giá theo cùng bộ tiêu chí, nhằm loại bỏ nhiễu và đảm bảo sự khác biệt hiệu năng phản ánh đúng bản chất mô hình.

4. Kết quả và thảo luận

4.1. Kết quả thực nghiệm

4.1.1. Kết quả dự đoán

Bảng 1. *Bảng đánh giá hiệu năng của các mô hình*

Tên mô hình	MAE (phút)	RMSE (phút)	R^2
KNN	8,77	12,99	0,901
ANN	7,65	12,14	0,913
CBR	9	13,19	0,898
FA-PP-R-ML	8,81	12,72	0,905
LSTM	7,53	11,78	0,918
Ensemble (đề xuất)	7,32	11,06	0,928

Theo kết quả ở *bảng 1*, các mô hình học sâu như ANN và LSTM, cùng với mô hình tổng hợp Ensemble, cho kết quả dự đoán vượt trội hơn so với các mô hình đơn giản như KNN và CBR. Mô hình Ensemble đặc biệt nổi bật với độ chính xác cao nhất, thể hiện qua MAE thấp nhất (7,32 phút) và R^2 cao nhất (0,928), cho thấy tính ổn định và hiệu quả trong việc dự đoán thời gian hoàn thành marathon.

4.1.2. So sánh hiệu năng đối với kết quả nghiên cứu trước đó

Bảng 2. *Bảng so sánh*

Tên mô hình	Ưu điểm	Nhược điểm
ANN	Học phi tuyến tốt, hiệu quả cao	Cần tinh chỉnh kỹ lưỡng, dễ overfitting
CBR	Dễ hiểu, logic gần gũi với con người	Không cần tổng quát hóa, phụ thuộc vào dữ liệu gần
FA-PP-R-ML	Kết hợp giảm chiều và hồi quy, dễ giải thích	Giảm chiều có thể mất thông tin quan trọng
LSTM	Tốt với dữ liệu tuần tự, linh hoạt	Phức tạp, yêu cầu nhiều dữ liệu và tài nguyên.
Ensemble (đề xuất)	Tận dụng ưu điểm của nhiều mô hình, chính xác cao nhất	Cần huấn luyện nhiều mô hình con, thời gian huấn luyện cao.

Theo so sánh ở *bảng 2*, Ensemble đạt được kết quả định lượng tốt nhất với MAE và RMSE thấp nhất, cùng với R^2 cao nhất. Về mặt định tính, phương pháp này kết hợp hiệu quả thông tin từ các mô hình tuyến tính (Linear), phi tuyến (MLP) và mô hình cây (RF), giúp giảm thiểu rủi ro overfitting và tăng cường tính tổng quát của mô hình.

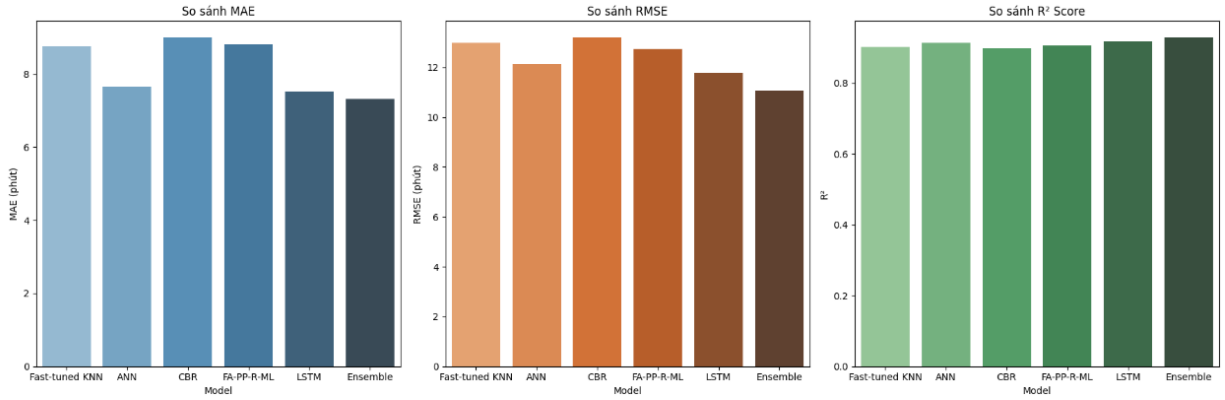
4.1.3. Trực quan hóa kết quả

Để so sánh hiệu năng giữa các mô hình, nhóm nghiên cứu sử dụng ba chỉ số MAE, RMSE và R^2 (Hình 2). Kết quả cho thấy mô hình Ensemble đạt MAE thấp nhất (7,32 phút), RMSE nhỏ nhất (11,06 phút) và R^2 cao nhất (0,928), khẳng định khả năng dự đoán vượt trội và ổn định hơn so với các mô hình khác.

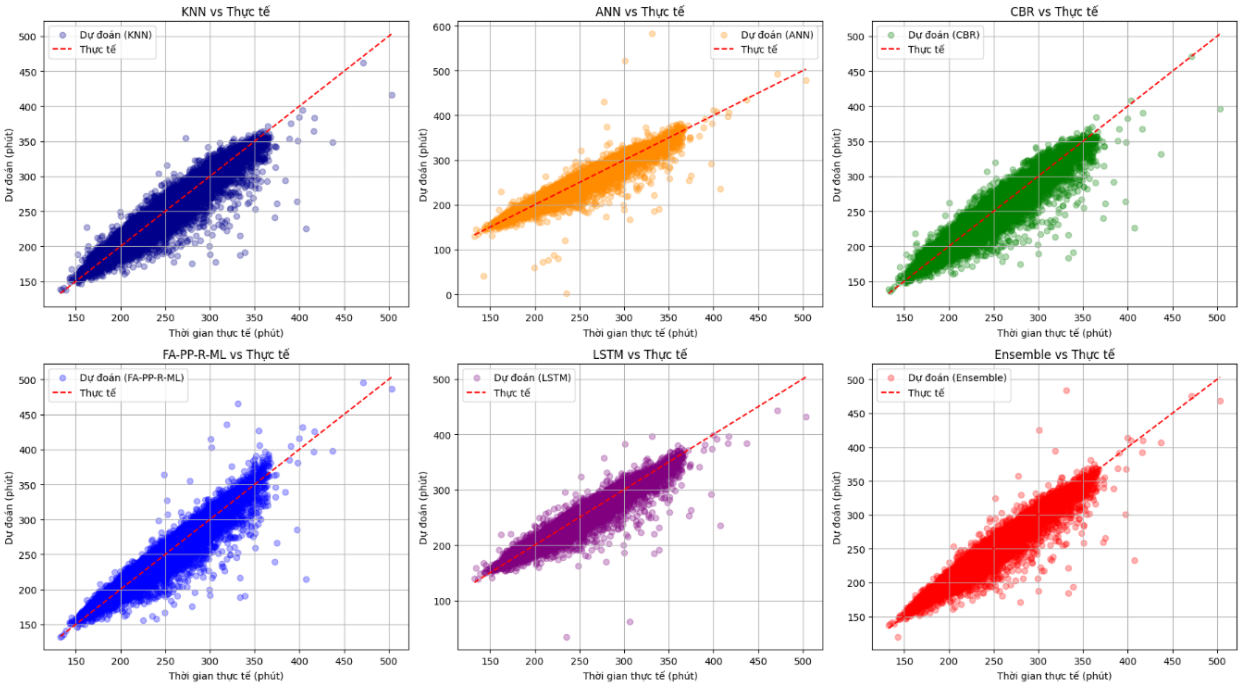
Kết quả (*hình 3*) cho thấy ANN và LSTM mô phỏng tốt mối quan hệ đầu vào–đầu ra, trong khi mô hình Ensemble thể hiện phân bố điểm gần sát đường lý tưởng nhất, khẳng định độ chính xác và ổn định vượt trội. Ngược lại, KNN và CBR có sai lệch lớn ở các điểm biên, phản ánh hạn chế trong khả năng tổng quát hóa.

JTE
HCMUTE
ISSN: 1859-1272

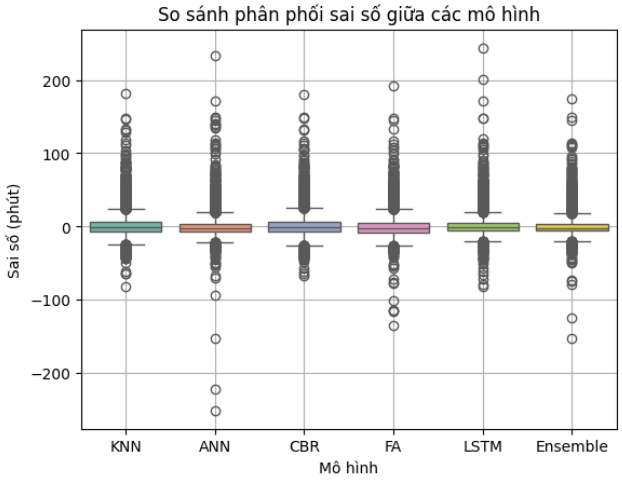
TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn



Hình 2. *Kết quả so sáng các thông số MAE, RMSE, và R^2 của mô hình học tập hợp với những mô hình khác*



Hình 3. *Phân bố của các giá trị dự đoán theo từng mô hình*



Hình 4. *So sánh phân bố sai số giữa các mô hình*

Kết quả từ *hình 4* cho thấy mô hình Ensemble có sai số trung bình thấp và mức dao động nhỏ nhất, thể hiện qua boxplot hẹp và gần trục 0, đồng thời hạn chế các điểm ngoại lai. Ngược lại, KNN và CBR có sai số phân tán rộng và nhiều ngoại lai, phản ánh tính không ổn định. ANN và LSTM cũng đạt kết quả tốt nhưng vẫn xuất hiện một số ngoại lệ. Nhìn chung, Ensemble chứng minh được độ chính xác và độ tin cậy vượt trội khi áp dụng thực tế.

4.2. Thảo luận

4.2.1. Phân tích kết quả

Kết quả thực nghiệm cho thấy mô hình Ensemble (kết hợp Linear Regression, Random Forest và MLPRegressor) đạt hiệu năng cao nhất với MAE = 7,32 phút, RMSE = 11,06 phút và R^2 = 0,928. Điều này khẳng định khả năng dự đoán chính xác, ổn định và giải thích được gần 93% phương sai của dữ liệu.

Sở dĩ mô hình Ensemble hoạt động tốt hơn là nhờ vào khả năng kết hợp linh hoạt nhiều thuật toán học máy có thế mạnh khác nhau [10]. Linear Regression giúp nắm bắt mối quan hệ tuyến tính giữa các đặc trưng [19], Random Forest khai thác được cấu trúc phi tuyến và kháng nhiễu tốt [20], trong khi MLPRegressor (một dạng mạng neural) có khả năng học sâu các mẫu phức tạp [21]. Việc sử dụng mô hình tổng hợp (meta-model) để kết hợp đầu ra của các mô hình con cho phép tận dụng tối đa các tín hiệu mạnh mẽ trong dữ liệu, đồng thời giảm thiểu ảnh hưởng của sai số từ bất kỳ mô hình đơn lẻ nào.

Hiệu quả của mô hình đề xuất còn gắn liền với đặc điểm của tập dữ liệu và quy trình tiền xử lý hợp lý. Tập dữ liệu sử dụng gồm các thông tin chi tiết về vận động viên qua ba năm liên tiếp (2015–2017) [3], với các đặc trưng được chuẩn hóa và mã hóa rõ ràng: tuổi, giới tính, thời gian hoàn thành 5K, 10K, Half, quốc tịch và năm thi đấu. Việc chuyển đổi dữ liệu thời gian sang dạng số phút, mã hóa quốc gia bằng one-hot encoding, và chuẩn hóa toàn bộ đặc trưng bằng StandardScaler đã giúp mô hình học hiệu quả hơn và giảm thiểu độ lệch giữa các thuộc tính.

Tổng thể, thành công của mô hình Ensemble không chỉ đến từ kiến trúc mô hình mà còn nằm ở chiến lược xử lý và khai thác dữ liệu hợp lý, cho phép mô hình học được cả xu hướng toàn cục lẫn các mối quan hệ chi tiết, từ đó mang lại kết quả dự đoán vượt trội so với các mô hình trong các nghiên cứu trước.

4.2.2. Những hạn chế của nghiên cứu

Mặc dù đạt kết quả khả quan, nghiên cứu vẫn tồn tại một số hạn chế. Thứ nhất, dữ liệu chỉ giới hạn trong ba mùa Boston Marathon (2015–2017) [3], chủ yếu gồm vận động viên hoàn thành, nên thiếu tính đại diện và có nguy cơ thiên lệch khi áp dụng rộng hơn. Thứ hai, phạm vi so sánh mô hình còn hạn chế, mới dừng ở KNN, ANN, CBR, FA-PP-R-ML và LSTM, chưa đối chiếu với các mô hình hiện đại như XGBoost, LightGBM, CatBoost hay các mô hình học sâu tiên tiến. Thứ ba, một số giả định đơn giản hóa (quan hệ tuyến tính và ổn định giữa đặc trưng và kết quả) chưa phản ánh đầy đủ các yếu tố ngoại sinh như thời tiết, địa hình hay tình trạng sức khỏe. Cuối cùng, kỹ thuật huấn luyện chưa được tối ưu (chưa dùng Grid Search, Bayesian Optimization hay AutoML), và mức độ khái quát hóa chưa được kiểm chứng trên tập dữ liệu độc lập hay trong triển khai thực tế.

4.2.3. Đề xuất nghiên cứu tương lai

Về hướng phát triển, nhóm sẽ mở rộng dữ liệu (bổ sung BMI, thời tiết,...) và hợp tác với chuyên gia thể thao để nâng cao độ chính xác của mô hình. Các kỹ thuật hiện đại như Transformer, Attention, graph-based hay kiến trúc lai (CNN–LSTM–Transformer) sẽ được thử nghiệm nhằm xử lý dữ liệu phức tạp và đa dạng hơn. Bên cạnh đó, nhóm sẽ phát triển các phương pháp diễn giải (SHAP, LIME, Visual Attention) để tăng tính minh bạch, đồng thời xây dựng ứng dụng hỗ trợ luyện tập và dự đoán kết quả marathon. Ngoài thể thao, mô hình Ensemble còn có tiềm năng ứng dụng trong các lĩnh vực như môi trường, y tế, năng lượng; ví dụ trong dự báo chất lượng không khí [22]. Các nghiên cứu tiếp theo sẽ kiểm chứng khả năng tổng quát hóa trên nhiều tập dữ liệu khác lĩnh vực.

5. Kết luận

5.1. Kết quả đạt được

Nghiên cứu đã đề xuất mô hình học máy tổng hợp (Ensemble Learning) kết hợp Linear Regression, Random Forest và MLPRegressor thông qua cơ chế meta-learning, nhằm dự đoán thời gian hoàn thành marathon. Trên dữ liệu Boston Marathon 2015–2017, mô hình đạt MAE = 7,32 phút, RMSE = 11,06 phút và R^2 = 0,928, vượt trội so với các mô hình đối chứng (KNN, ANN, CBR, FA-PP-R-ML, LSTM). Các trực quan hóa như scatter plot và boxplot cũng cho thấy tính nhất quán cao giữa dự đoán và thực tế, khẳng định hiệu quả của hướng tiếp cận này trong phân tích và dự đoán hiệu suất thể thao.

5.2. Ý nghĩa ứng dụng thực tiễn

Kết quả đạt được cho thấy mô hình đề xuất có tiềm năng ứng dụng thực tiễn cao trong các hệ thống hỗ trợ huấn luyện viên, vận động viên và nền tảng tư vấn thể thao cá nhân hóa. Việc dự đoán chính xác thời gian hoàn thành có thể hỗ trợ xây dựng giáo án tập luyện, lập kế hoạch thi đấu và theo dõi hồi phục chấn thương, đồng thời tích hợp vào ứng dụng di động hoặc thiết bị đeo thông minh. Với khả năng mở rộng linh hoạt, mô hình có thể bổ sung dữ liệu ngoại sinh (thời tiết, BMI, địa hình) và áp dụng các kỹ thuật hiện đại (attention, graph-based, hybrid architectures) để nâng cao độ chính xác và tính diễn giải. Trong tương lai, nhóm nghiên cứu hướng tới phát triển ứng dụng hỗ trợ chạy bộ, tích hợp dự báo cá nhân hóa và khả năng diễn giải kết quả, nhằm gia tăng giá trị ứng dụng bền vững.

Xung đột lợi ích

Các tác giả tuyên bố không có xung đột lợi ích trong bài báo này.

Tuyên bố dữ liệu sẵn có

Dữ liệu hỗ trợ cho các khám phá của nghiên cứu này khi độc giả yêu cầu một cách hợp lý sẽ được tác giả liên hệ cung cấp.

TÀI LIỆU THAM KHẢO

[1] A. Keogh, O. Sheridan, O. McCaffrey, S. Dunne, A. Lally, and C. Doherty, “The determinants of marathon performance: An observational analysis of anthropometric, pre-race and in-race variables,” *Int. J. Exerc. Sci.*, vol. 13, no. 6, pp. 1132–1142, 2020.
[2] W. Yong, P. Lingyun, and W. Jia, “Statistical analysis and ARMA modeling for the big data of marathon score,” *Sci. Sports*, vol. 35, no. 6, pp. 375–385, 2020.
[3] Rojour, “Finishers Boston Marathon 2015, 2016 & 2017,” *Kaggle*, 2017. [Online]. Available: https://www.kaggle.com/datasets/rojour/boston-results. Accessed: 2025.
[4] L. Lerebourg, D. Saboul, M. Clémençon, and J. B. Coquart, “Prediction of marathon performance using artificial intelligence,” *Int. J. Sports Med.*, vol. 44, no. 5, pp. 352–360, 2023.
[5] C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, “Using case-based reasoning to predict marathon performance and recommend tailored training plans,” in *Proc. 28th Int. Conf. Case-Based Reasoning (ICCBR 2020)*, 2020.
[6] J. Chen, “Factor and correlation analysis for predicting marathon race performance using machine learning algorithms,” *J. Electr. Syst.*, pp. 1948–1958, 2024.
[7] H. Muijlwijk, B. Smyth, M. C. Willemsen, and W. A. IJsselsteijn, “Benefits of human-AI interaction for expert users interacting with prediction models: A study on marathon running,” in *Proc. 29th Int. Conf. Intell. User Interfaces (IUI ’24)*, Greenville, SC, USA, 2024.
[8] Y. Ding, “Analyzing athletes’ physical performance and trends in athletics competitions using time series data mining algorithms,” *J. Electr. Syst.*, pp. 736–746, 2024.
[9] K. K. El-Kassabi and M. A. S. H. Taha, “Deep learning approach for forecasting athletes’ performance in sports tournaments,” unpublished.
[10] R. Huang, Z. Qian, H. Ma, Z. Han, and Y. Xie, “Sports performance prediction for college students through ensemble learning algorithm,” *IEICE Trans. Inf. Syst.*, vol. E108.D, no. 7, pp. 776–783, 2025.
[11] T. Anande, S. Alsaadi, and M. Leeson, “Enhanced modelling performance with boosting ensemble meta learning and Optuna optimization,” *SN Comput. Sci.*, vol. 6, Art. no. 12, 2024.
[12] Rojour, “boston_results: Scrapping and visualizing Boston Marathon results,” *GitHub*, 2017. [Online]. Available: https://github.com/rojour/boston_results. Accessed: 2025.
[13] D. H. Wolpert, “Stacked generalization,” *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.
[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, 1986.
[15] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[16] J. L. Kolodner, “An introduction to case-based reasoning,” *Artif. Intell. Rev.*, vol. 6, pp. 3–34, 1992.
[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[18] A. K. Kuchibhotla and L. D. Brown, “Model-free study of ordinary least squares linear regression,” arXiv preprint arXiv:1809.05296, Sep. 2018.

[19] S. Lee, "7 surprising stats where linear regression shapes sports data analysis," *Number Analytics, LLC*, Mar. 19, 2025. [Online]. Available: https://www.numberanalytics.com/blog/surprising-stats-linear-regression-sports-data-analysis. Accessed: Apr. 29, 2025.
[20] TechGoGreen, "Random forest algorithm," *TechGoGreen*, Jun. 20, 2023. [Online]. Available: https://techgogreen.com/random-forest-algorithm/?utm_source=chatgpt.com. Accessed: Apr. 29, 2025.
[21] A. Kumar, "Sklearn neural network example – MLPRegressor," *Analytics Yogi*, May 2, 2023. [Online]. Available: https://vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/. Accessed: Apr. 29, 2025.
[22] V. Hua, N. T. Dang, M. S. Nguyen, H. N. Bui, and A. B. Arun, "The impact of data imputation on air quality prediction problem," *PLoS One*, vol. 19, no. 9, Art. no. e0306303, 2024.
[23] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS 2017)*, Long Beach, CA, USA, 2017.
[24] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.
[25] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state of the art," *Knowl.-Based Syst.*, vol. 212, Art. no. 106622, 2021.

Anh Khoa Mai is a fourth-year student in Information Technology, majoring in Artificial Intelligence at Ho Chi Minh City University of Technology and Education. Currently, he is working as an intern at FPT Software Co., Ltd., Ho Chi Minh City. This paper is his first publication, developed from the idea of his undergraduate thesis. It provides him with an opportunity to further study Artificial Intelligence and Deep Learning, while practicing scientific research skills in an academic environment. Research areas: machine learning, deep learning, reinforcement learning, chatbot.

Email: anhkhoamai11040307@gmail.com. ORCID: https://orcid.org/0009-0007-2204-2040

Giao Quynh Ha Nguyen is currently a fourth-year student in Information Technology, majoring in Software Engineering at Ho Chi Minh City University of Technology and Education. Currently, she is working as an intern at FPT Software Co., Ltd., Ho Chi Minh City.
This paper is her first publication during her studies at HCMUTE, serving as an opportunity for her to practice research skills and synthesize specialized knowledge.
Research areas: Mobile Programming, Deep Learning.

Email: nguyenhaquynhgiao9569@gmail.com. ORCID: https://orcid.org/0009-0004-5643-207X

Cong Manh Hoang is currently a fourth-year student in Information Technology, majoring in Software Engineering at Ho Chi Minh City University of Technology and Education. This report is his first academic work during his studies, serving as an opportunity to practice research skills and synthesize specialized knowledge. Research areas: Mobile Programming, Web Programming.

Email: hoangmanh6889@gmail.com . ORCID: https://orcid.org/0009-0005-6456-2613

Phuong Thi Ngoc Truong is currently a lecturer at Ho Chi Minh City University of Technology and Education. She graduated from the University of Science, Ho Chi Minh City, in 2005 and pursued a Master's degree in Information Technology at Kookmin University, South Korea. She is now a Ph.D. candidate at the Computer Science Laboratory, University of Information Technology.
Her research interests include Computer Vision, Deep Learning, and Mobile Programming.

Email: phuongttn@hcmute.edu.vn. ORCID: https://orcid.org/0009-0003-9963-9874. Phone: +84 – 942920912.