

NGHIÊN CỨU ỨNG DỤNG MỘT SỐ MÔ HÌNH HỌC MÁY TRONG DỰ BÁO LƯỢNG MƯA TẠI THÀNH PHỐ VINH

Nguyễn Hữu Hải^{1*}, Phạm Quang Dũng¹, Trần Đức Quỳnh², Nguyễn Thị Thảo¹

¹Khoa Công nghệ thông tin, Học viện Nông Nghiệp Việt Nam

²Trường Quốc tế, Đại học Quốc gia Hà Nội

*Tác giả liên hệ: nhhai@vnua.edu.vn

Ngày nhận bài: 07.11.2025

Ngày chấp nhận đăng: 10.06.2026

TÓM TẮT

Lượng mưa là yếu tố khí hậu thiết yếu, ảnh hưởng trực tiếp đến hệ sinh thái tự nhiên, con người và đặc biệt là hoạt động sản xuất nông nghiệp, quy hoạch đô thị. Trong bối cảnh biến đổi khí hậu ngày càng phức tạp, với sự gia tăng của các hiện tượng thời tiết cực đoan như mưa lớn, bão và lũ lụt, việc dự báo chính xác lượng mưa trở nên quan trọng nhưng cũng đầy thách thức. Nghiên cứu này tập trung đánh giá hiệu quả của các mô hình học máy trong dự báo lượng mưa tại khu vực cụ thể. Dữ liệu lượng mưa thường có tính phi tuyến, sự biến động, mức độ ngẫu nhiên cao, nhiều nhiễu. Vì vậy, nghiên cứu đề xuất một phương pháp cải tiến dữ liệu đầu vào đa bước, nhằm nâng cao độ chính xác của mô hình dự báo. Nhiều thuật toán học máy được triển khai và đánh giá thông qua các chỉ số như độ chính xác (Accuracy), F1 có trọng số (Weighted F1-score) và chỉ số AUC. Kết quả cho thấy hiệu suất của các mô hình dự báo được cải thiện đáng kể với dữ liệu mới. XGBoost là mô hình có cải thiện đáng kể nhất với mức tăng trung bình 15,5% cả về độ chính xác và điểm F1 có trọng số so với khi chỉ sử dụng dữ liệu gốc. Kết quả nghiên cứu có tiềm năng ứng dụng trong phân tích và dự báo lượng mưa ở cấp độ khu vực.

Từ khóa: Dự báo lượng mưa, học máy, lựa chọn đặc trưng, trích chọn đặc trưng, dữ liệu chuỗi thời gian, phương pháp trí tuệ nhân tạo.

Rainfall Forecasting in Vinh City Using Machine Learning Models

ABSTRACT

Rainfall is a vital climatic factor that directly influences natural ecosystems, human life, and particularly agricultural production. In the context of increasingly complex climate change, with the growing frequency of extreme weather events such as heavy rainfall, storms, and floods, accurate rainfall prediction has become both crucial and challenging. This study focuses on evaluating the performance of various machine learning models in forecasting rainfall in a specific region. Rainfall data are typically characterized by strong nonlinearity, high variability, considerable randomness, and substantial noise. Therefore, this research proposes a multi-step data enhancement approach to improve model accuracy. Several machine learning algorithms are implemented and evaluated using key performance metrics such as Accuracy, Weighted F1-score, and AUC. The results indicate that the forecasting models achieved substantial performance improvements when trained on new dataset. Among them, XGBoost exhibited the most notable gains, with an average increase of 15.5% in both accuracy and weighted F1-score compared with using only the original data. The findings demonstrate the potential applicability of the proposed approach in regional rainfall analysis and forecasting, contributing to more reliable climate prediction and decision support, especially in agricultural production and water resources management.

Keywords: Rainfall prediction, machine learning, feature extraction, feature selection, time series data, artificial intelligence methods.

1. ĐẶT VẤN ĐỀ

Biến đổi khí hậu tác động mạnh mẽ đến

hầu hết các lĩnh vực kinh tế - xã hội và sự sống trong tự nhiên. Trong đó, lượng mưa là yếu tố khí hậu chủ chốt, ảnh hưởng trực tiếp đến hệ

thống thủy văn, quản lý nguồn nước, tưới tiêu nông nghiệp và các hoạt động sản xuất khác. Sự thay đổi bất thường của lượng mưa có thể gây ra các hiện tượng thời tiết cực đoan như hạn hán, mưa lớn và lũ lụt, dẫn đến thiệt hại nghiêm trọng cho sản xuất nông nghiệp, cơ sở hạ tầng và đời sống dân sinh, ví dụ năm 2020, miền Trung chịu thiệt hại kinh tế hơn 36.000 tỷ đồng do thiên tai. Riêng đợt mưa lũ tháng 10 đã làm hư hại 38,5ha lúa và hoa màu, gây tổn thất trên 28.000 tỷ đồng (theo Cục Quản lý Đê điều và Phòng chống thiên tai).

Tại Việt Nam những năm gần đây, tần suất và cường độ mưa lớn, bão và lũ lụt gia tăng, gây thiệt hại đáng kể về kinh tế và môi trường đặc biệt tại miền Trung nói chung và thành phố Vinh nói riêng. Với đặc điểm đô thị lớn ven biển, Vinh thường chịu ảnh hưởng của thời tiết cực đoan, gây thiệt hại kinh tế và ảnh hưởng đời sống dân cư và làm gia tăng ngập úng đô thị. Trong bối cảnh biến đổi khí hậu phức tạp, việc nghiên cứu và dự báo lượng mưa trở nên cấp thiết để giảm thiểu rủi ro, hỗ trợ quy hoạch nông nghiệp, đô thị và quản lý tài nguyên nước. Tuy nhiên, do tính biến động cao và khó dự đoán của khí hậu, ước lượng và dự báo mưa tiếp tục là thách thức lớn đối với các nhà khoa học.

Các nghiên cứu về ước tính và dự báo lượng mưa chủ yếu dựa trên dữ liệu radar và vệ tinh, sử dụng các phương pháp truyền thống như mô phỏng và phân tích ảnh viễn thám. Chẳng hạn, Công Thanh & cs. (2018) ước tính lượng mưa từ độ phản hồi radar bằng nhiều công thức khác nhau; Nguyễn Hoàng Minh & cs. (2022) nghiên cứu hiệu chỉnh lượng mưa ước tính từ radar nhằm nâng cao chất lượng mô hình dòng chảy.

Những năm gần đây, trí tuệ nhân tạo và học máy được ứng dụng nhiều trong dự báo mưa. Nguyễn Hồng Giang & cs. (2020) thử nghiệm MLP và kNN cho dự báo lượng mưa; Nguyễn Vinh Thư & cs. (2023) dùng Random Forest với dữ liệu vệ tinh, cho thấy độ tương quan cao dù mô hình hơi đánh giá thấp thực tế. Dương Thị Hà & cs. (2022) triển khai mô hình LSTM đa biến để dự báo đồng thời lượng mưa và nhiệt độ.

Dù việc ứng dụng học máy và học sâu trong dự báo mưa đang tăng, các nghiên cứu chuyên sâu tại Việt Nam vẫn còn ít. Vì vậy, cần tiếp tục mở rộng nghiên cứu ước tính và dự báo lượng mưa để nâng cao độ chính xác và hỗ trợ hiệu quả cho công tác khí tượng - thủy văn.

Lượng mưa là yếu tố khí hậu thiết yếu, ảnh hưởng trực tiếp đến hệ sinh thái tự nhiên và các hoạt động của con người, đặc biệt trong nông nghiệp, công nghiệp, thủy điện và quản lý nguồn nước. Các hiện tượng thời tiết cực đoan như hạn hán, lũ lụt, mưa bão đều chịu tác động mạnh từ sự biến động lượng mưa, gây thiệt hại lớn cho sản xuất và đời sống. Do đó, việc ước lượng và dự báo chính xác lượng mưa trở thành vấn đề quan trọng nhưng đầy thách thức trong nghiên cứu khí tượng, thu hút sự quan tâm của các tổ chức và nhà khoa học trên toàn thế giới.

Các phương pháp truyền thống trong dự báo lượng mưa chủ yếu dựa trên ảnh vệ tinh, radar khí tượng, mô phỏng vật lý và hồi quy tuyến tính. Tuy nhiên, chúng tồn tại hạn chế do quá trình xử lý phức tạp, tốn thời gian và không thể mô hình hóa các mối quan hệ phi tuyến trong dữ liệu mưa. Ngoài ra, các phương pháp này dễ bị ảnh hưởng bởi nhiễu và điểm bất thường trong dữ liệu khí tượng. Hiện nay, các kỹ thuật học máy và học sâu mang lại cách tiếp cận dựa trên dữ liệu, giúp phát hiện các mối quan hệ phức tạp trong chuỗi dữ liệu thời tiết lịch sử (Saleh & cs., 2024; Mojtahedi & cs., 2025). Các mô hình này có khả năng xử lý tốt tính phi tuyến và dữ liệu bất thường, góp phần nâng cao độ chính xác của dự báo. Tuy vậy, do đặc tính biến động và khó đoán của lượng mưa, việc dự báo chính xác vẫn là một thách thức lớn (Kumar & Yadav, 2020).

Trong những năm gần đây, các phương pháp học máy đã chứng minh hiệu quả vượt trội so với các phương pháp dự báo truyền thống. Abbott & Marohasy (2014) sử dụng mạng nơron nhân tạo (ANN) để dự báo lượng mưa liên tục dựa trên các chỉ số khí hậu, giúp cải thiện đáng kể độ chính xác. Yu & cs. (2017) so sánh các mô hình học máy như Random Forest (RF) và SVM trong dự báo lượng mưa thời gian thực,

cho thấy các mô hình đơn lẻ như RF và SVM hoạt động hiệu quả hơn mô hình tích hợp trong dự báo ngắn hạn. Cramer & cs. (2017) đánh giá nhiều mô hình học máy gồm SVR, NN, GP và KNN, kết quả cho thấy chúng vượt trội hơn các mô hình dự báo tiên tiến khác và thể hiện mối tương quan giữa độ chính xác và điều kiện khí hậu. Rahman & cs. (2022) đề xuất hệ thống dự báo lượng mưa thời gian thực cho thành phố thông minh, kết hợp logic mờ và bốn mô hình học máy, đạt hiệu suất cao hơn các mô hình đối chứng. Endalie & cs. (2022) cũng khẳng định khả năng cải thiện độ chính xác của các mô hình học máy trong dự báo mưa. Gần đây, nhiều nghiên cứu như Pirone & cs. (2023) với ANN, Baig & cs. (2024) với Boosted Regression Trees, Kumar & cs. (2023) với SVM, và Baljon & Sharma (2023) với cây quyết định (DT) đều cho thấy phương pháp học máy có hiệu quả vượt trội so với các mô hình thống kê truyền thống trong ước lượng và dự báo lượng mưa.

Các mô hình học máy cho thấy hiệu quả vượt trội so với mô hình truyền thống trong xử lý dữ liệu phức tạp, song vẫn tồn tại hạn chế như phụ thuộc vào đặc trưng và chất lượng dữ liệu đầu vào. Dữ liệu thời tiết, đặc biệt là lượng mưa, mang tính phức tạp, biến động lớn theo vùng miền và khu vực. Do đó, việc ước lượng và dự báo chính xác lượng mưa vẫn là một thách thức cần được tiếp tục nghiên cứu và cải thiện để nâng cao độ tin cậy và khả năng ứng dụng trong thực tiễn.

Nghiên cứu này như một phần mở rộng của nghiên cứu đã được tác giả Hai & cs. (2025) công bố. Nghiên cứu mở rộng theo hướng tập trung phân tích, khai thác khía cạnh các mối quan hệ thời gian phức tạp của các dữ liệu chuỗi thời gian khí tượng.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Dữ liệu và khu vực nghiên cứu

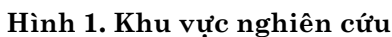
Nghiên cứu này tập trung vào khu vực thành phố Vinh, trung tâm kinh tế - văn hoá - chính trị của tỉnh Nghệ An, nằm ở tọa độ khoảng $19^{\circ}20'$ vĩ Bắc và $104^{\circ}50'$ kinh Đông, thuộc vùng Bắc Trung bộ của Việt Nam. Nhờ vị

trí tiếp giáp biển Đông và nằm trong vùng khí hậu nhiệt đới gió mùa, thành phố Vinh mang đặc trưng của khí hậu nóng ẩm, với hai mùa rõ rệt: mùa nóng ẩm và mùa lạnh khô với sự biến động mạnh về nhiệt độ và lượng mưa giữa các thời kỳ trong năm.

Nhiệt độ trung bình năm của khu vực dao động từ $23-24^{\circ}\text{C}$, với sự chênh lệch rõ rệt giữa các mùa. Mùa hè (tháng 5-8) chịu ảnh hưởng của gió Tây Nam khô nóng, nhiệt độ trung bình khoảng 33°C và có thể vượt 42°C ; trong khi mùa đông (tháng 11-2) bị tác động bởi gió Đông Bắc lạnh ẩm, nhiệt độ trung bình chỉ khoảng 19°C , đôi khi dưới 10°C . Lượng mưa trung bình năm đạt 1.200-2.000mm, tập trung vào mùa hè, thường kèm 5-8 cơn bão hoặc áp thấp nhiệt đới. Thành phố Vinh có độ ẩm trung bình 85-90% và khoảng 1.500-1.700 giờ nắng mỗi năm, thuận lợi cho nông nghiệp. Tuy nhiên, khí hậu khắc nghiệt với mùa hè nóng kéo dài, mùa đông lạnh ẩm và lượng mưa phân bố không đều khiến khu vực dễ bị tổn thương trước biến đổi khí hậu, đặc biệt là hạn hán, lũ lụt và bão mạnh.

Nhìn chung, khí hậu thành phố Vinh thể hiện rõ đặc trưng của vùng Bắc Trung bộ - nhiệt đới gió mùa, mưa nhiều, nắng gắt, dao động mạnh theo mùa - vừa là lợi thế cho phát triển nông nghiệp, vừa đặt ra yêu cầu cao về thích ứng trong quy hoạch đô thị và sản xuất bền vững.

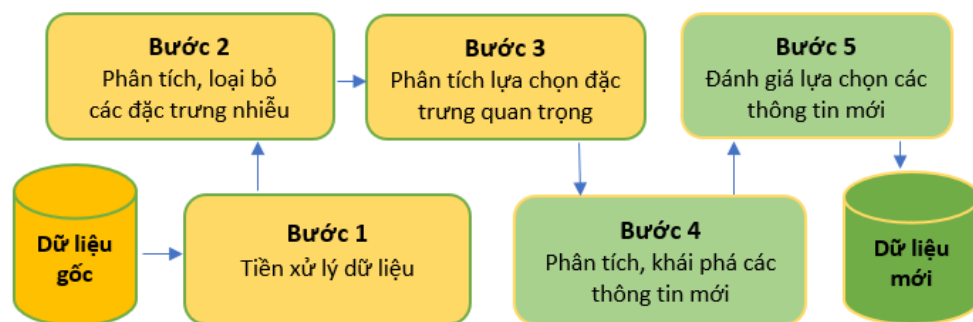
Nghiên cứu này sử dụng dữ liệu khí tượng giai đoạn 2014-2022 tại khu vực thành phố Vinh - Nghệ An, bao gồm các yếu tố như nhiệt độ, độ ẩm, lượng mưa, bốc hơi, hướng và tốc độ gió, lượng mây, số giờ nắng, được ghi nhận theo các thời điểm khác nhau trong ngày (độc giả quan tâm đến nghiên cứu về lượng mưa có thể liên hệ tác giả để có thêm thông tin). Dữ liệu lượng mưa hàng ngày cũng được tích hợp, với số bản ghi gần gấp đôi so với nghiên cứu của Hai & cs. (2025). Nghiên cứu dự báo lượng mưa được thực hiện như bài toán phân lớp nhị phân với nhãn 0 và 1 (trời mưa vào ngày mai). Mục tiêu nghiên cứu là phân tích xu hướng thời tiết, đặc biệt là lượng mưa tại khu vực Vinh và vùng lân cận, nhằm hướng đến phát triển, cải tiến các mô hình dự báo lượng mưa chính xác và hiệu quả hơn cho khu vực.



Phương pháp học máy (machine learning) đã đạt được nhiều bước tiến đáng kể trong nhiều lĩnh vực dự báo khác nhau, bao gồm dự báo giá cổ phiếu, lưu lượng dòng chảy, năng lượng gió và lượng mưa, đặc biệt trong bối cảnh trí tuệ nhân tạo (AI) ngày càng phát triển mạnh mẽ. Nghiên cứu này sử dụng nhiều mô hình học máy đa dạng, bao gồm Rừng ngẫu nhiên (Random Forest), Hồi quy logistic (LR), Máy vector hỗ trợ (SVM), Hồi quy Gaussian (GP), mạng nơron nhiều lớp (MLP), cùng các thuật toán tăng cường (boosting) như AdaBoost, Gradient Boosting và Extreme Gradient Boosting (XGBoost), nhằm ước lượng và dự báo lượng mưa hàng ngày, đồng thời đánh giá hiệu quả giải pháp cải tiến dữ liệu đa bước đề xuất. Với đặc tính phi tuyến và mức độ nhiễu cao trong chuỗi thời gian khí tượng, đặc biệt là dữ liệu lượng mưa, nghiên cứu này tập trung khai thác mô hình có khả năng mô phỏng tốt mối quan hệ phi tuyến phức tạp, tiêu biểu như là thuật toán Random Forest và các thuật toán boosting (Kundu & cs., 2023).

Thực tế, dữ liệu đầu vào quyết định mạnh đến chất lượng dự báo của các mô hình học máy. Để giảm nhiễu và tăng khả năng phản ánh đặc trưng thời tiết, chúng tôi đề xuất cải tiến dữ liệu

Bước 5. Bộ lọc tương quan được dùng để đánh giá thông tin mới và loại bỏ các thông tin nhiễu, giữ lại những thông tin hữu ích. Nhờ đó, dữ liệu đầu vào phản ánh tốt hơn điều kiện thời tiết, góp phần cải thiện hiệu suất của mô hình dự báo.



Hình 2. Sơ đồ tổng quan giải pháp cải tiến dữ liệu đa bước

2.4. Xử lý dữ liệu và thiết lập thực nghiệm

Dữ liệu lượng mưa hàng ngày được trích xuất và tiền xử lý cho các biến thứ tự, phân loại như hướng gió, mức độ mây che phủ; sau đó chuẩn hóa bằng Min-Max Scaling. Tập dữ liệu được chia thành 80% cho huấn luyện - xác thực và 20% cho đánh giá độc lập. Phương pháp kiểm định chéo 5 lần (5-fold cross-validation) được áp dụng để đảm bảo tính khách quan trong đánh giá hiệu suất mô hình. Cuối cùng, hiệu năng của mô hình và tính hiệu quả của giải pháp cải tiến dữ liệu được đánh giá toàn diện thông qua các chỉ số Accuracy, F1-score, Weighted F1-score, Recall và AUC, phản ánh độ chính xác và khả năng phân loại của mô hình.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Phân tích thống kê các đặc trưng thời tiết

Phần này trình bày phân tích thống kê các đặc trưng thời tiết quan trọng như lượng mưa, nhiệt độ, độ ẩm, áp suất và số giờ nắng, nhằm phản ánh biến động khí hậu và xu hướng lượng mưa. Hình 3 cho thấy nền nhiệt giai đoạn 2014-2022 có xu hướng tăng nhẹ, với mức cao nhất trên 40°C, tập trung vào các tháng mùa hè nắng nóng kéo dài. Hình 4 mô tả tổng lượng mưa theo tháng của giai đoạn 9 năm bắt đầu từ 2014.

Hình 4 cho thấy lượng mưa tăng rõ rệt, đặc biệt sau năm 2018, với nhiều tháng vượt 600mm và đỉnh điểm hơn 1.300mm vào cuối năm 2020, chủ yếu do tác động của bão, áp thấp nhiệt đới và hoàn lưu sau bão, phản ánh thời tiết ngày

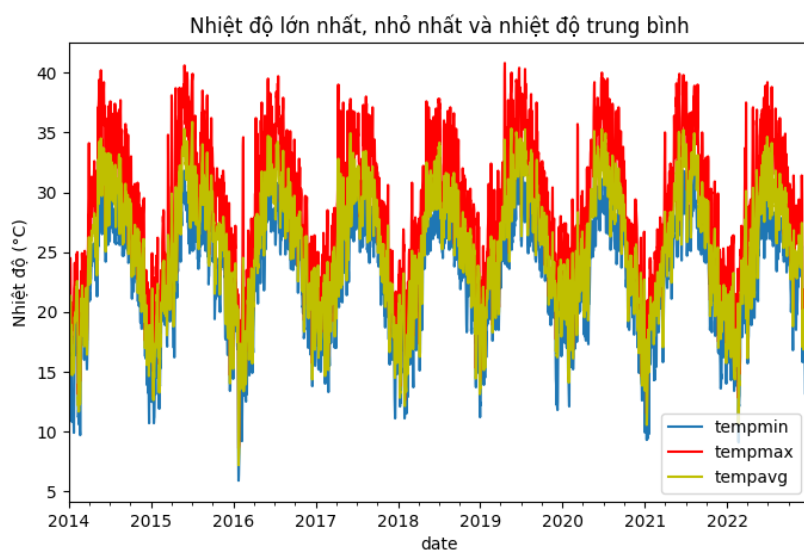
càng cực đoan với mưa lớn, bão và lũ lụt. Hình 5 minh họa tổng lượng mưa theo năm giai đoạn 2014-2022, cho thấy lượng mưa hàng năm tăng mạnh, đặc biệt từ 2020 trở đi đều vượt 2.500mm, cao nhất gần 3.500mm, tiếp tục khẳng định tác động rõ nét của sự kiện cực đoan mưa lớn, bão.

Từ hình 5, giai đoạn 2019-2022 cho thấy thời tiết biến đổi rõ rệt với các hiện tượng cực đoan như mưa bão, lũ lụt diễn ra bất thường, mạnh hơn và nhiều hơn trước, dẫn đến lượng mưa cực lớn ở nhiều tháng, nhiều năm. Mặc dù lượng mưa lớn thuận lợi cho sản xuất, các trận mưa cực đoan cũng gây nhiều thách thức, đặc biệt cho nông nghiệp, giao thông và quy hoạch đô thị, có nguy cơ ngập úng, ngập lụt diện rộng tại các khu vực sản xuất và đô thị.

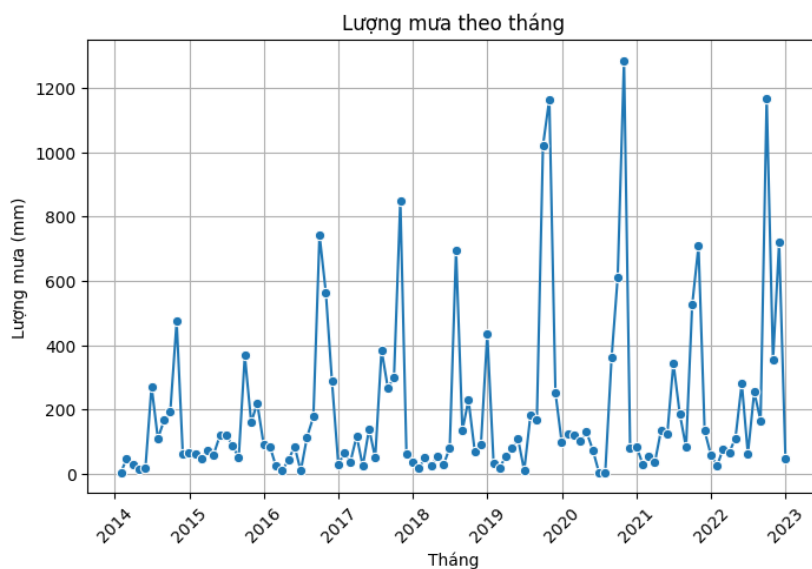
Hình 6, thể hiện mối tương quan giữa lượng mưa theo ngày với một số đặc trưng thời tiết quan trọng như nhiệt độ trung bình, độ ẩm, lượng bốc hơi và số giờ nắng trong ngày. Hình 7 thể hiện mối tương quan giữa các đặc trưng thời tiết được nghiên cứu như là nhiệt độ, độ ẩm, áp suất, lượng bốc hơi, hướng gió và tốc độ gió.

3.2. Khai phá thông tin mới

Như đã đề cập trong phần 2, trước hết, chúng tôi đánh giá mức độ tương quan giữa các đặc trưng thời tiết gốc để loại bỏ nhiễu. Tiếp đó, bộ lọc tương quan được áp dụng nhằm xác định mức độ ảnh hưởng của từng đặc trưng đối với biến dự báo. Cách tiếp cận này giúp nhận diện những đặc trưng quan trọng, từ đó ưu tiên khai thác và phát triển các thông tin mới có giá trị nhằm tăng cường thông tin đầu vào cho các mô hình, góp phần nâng cao hiệu suất mô hình dự báo.



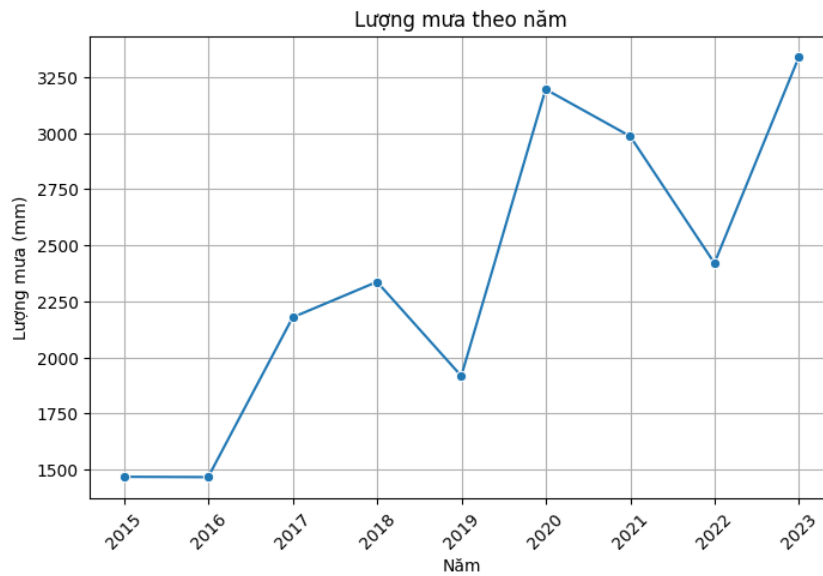
Hình 3. Nhiệt độ lớn nhất, nhỏ nhất và nhiệt độ trung bình theo ngày giai đoạn 2014-2022



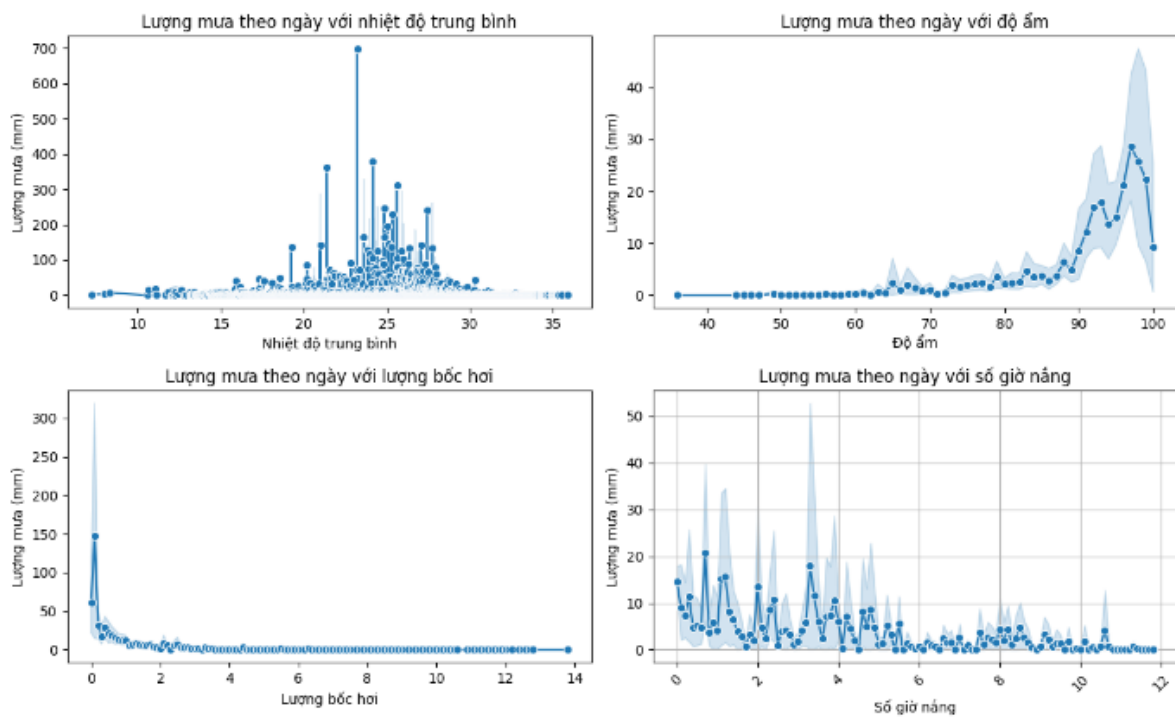
Hình 4. Lượng mưa theo tháng giai đoạn 2014-2022

Bảng 1. Tổng quan phân bố các đặc trưng thời tiết quan trọng

	Max	Min	Trung bình	Phương sai	Skewness	Kurtosis
Nhiệt độ trung bình	33,7	7,2	23,8	4,57	-0,62	-0,16
Lượng mưa (mm)	699	1,0	23,45	48,90	5,85	53,22
Lượng bốc hơi	9,90	0,0	1,48	1,17	2,21	8,42
Số giờ nắng	11,4	0,0	1,85	2,95	1,54	1,18
Độ ẩm 7	100	49,1	88,10	8,89	-1,10	1,03
Độ ẩm 19	100	41	81,30	11,84	-0,66	-0,10
Áp suất 7	1.035,70	993,20	1.012,58	6,86	-0,04	-0,45
Áp suất 19	1.033,50	991,70	1.009,90	6,73	0,02	-0,60



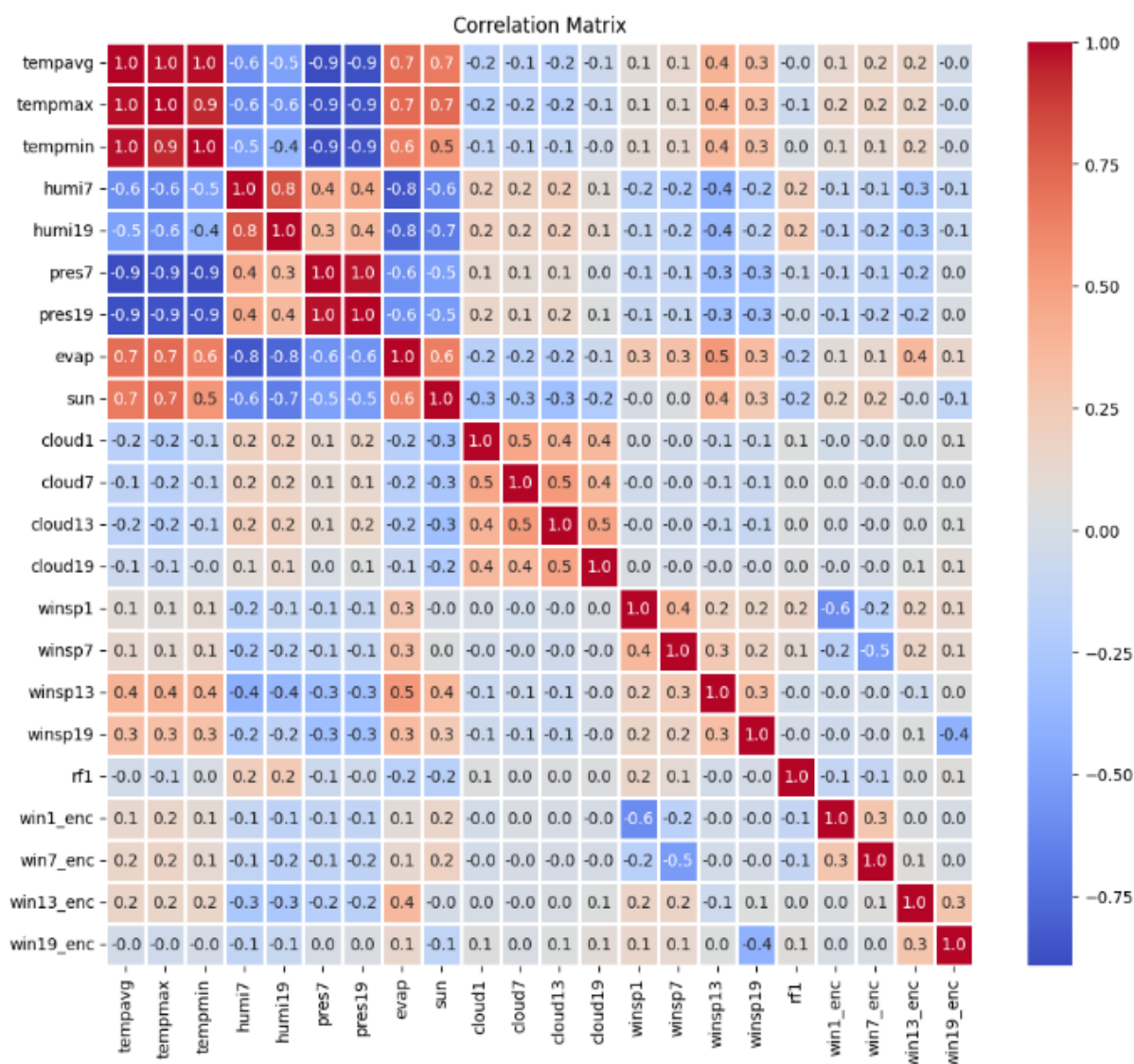
Hình 5. Lượng mưa hàng năm giai đoạn 2014-2022



Hình 6. Tương quan giữa lượng mưa và các đặc trưng thời tiết quan trọng

Nhằm giảm gánh nặng tính toán cho mô hình, các thông tin mới chỉ được khai phá từ một số đặc trưng gốc quan trọng. Kết quả của phân tích tương quan cho thấy các đặc trưng như nhiệt độ trung bình, độ ẩm, lượng bốc hơi, số giờ nắng và hướng gió có mức độ tương quan cao với biến dự báo. Với những đặc trưng gốc quan trọng được

lựa chọn, các thông tin mới được khai thác dựa trên phân tích các đặc trưng thống kê, quy luật phân bố xác suất tích lũy của các đặc trưng. Với các thông tin mới được tạo ra, chúng tôi thực hiện đánh giá mức độ ảnh hưởng của chúng đối với mô hình dự báo bằng cách phân tích mối tương quan giữa chúng và biến dự báo.



Hình 7. Ma trận tương quan giữa các đặc trưng thời tiết

Theo giải pháp cải tiến dữ liệu đề xuất, hàng trăm thông tin mới có thể được tạo ra. Do vậy, việc phân tích ảnh hưởng của chúng đối với mô hình dự báo là cần thiết để giữ lại những thông tin mới có hiệu quả và loại bỏ những thông tin mới không có nhiều ý nghĩa. Cách tiếp cận này không chỉ giúp cải thiện độ chính xác của mô hình dự báo, mà còn giảm đáng kể gánh nặng tính toán cho mô hình.

3.3. Đánh giá hiệu quả các mô hình dự báo và giải pháp cải tiến dữ liệu đề xuất

Để phân tích, đánh giá hiệu quả của các mô hình dự báo cũng như đánh giá hiệu quả của giải

pháp cải tiến dữ liệu đề xuất, nghiên cứu đã sử dụng nhiều mô hình học máy khác nhau, kết hợp với các chỉ số đánh giá như độ chính xác (precision), tỷ lệ thu hồi (recall), điểm F1 (F1-score), điểm F1 có trọng số (weighted F1-score), độ chính xác tổng thể (accuracy) và chỉ số AUC (AUC score). Trong đó, bài báo này tập trung phân tích các chỉ số F1-score, F1-score có trọng số, độ chính xác và chỉ số AUC nhằm đánh giá toàn diện hiệu suất của các mô hình dự báo cũng như giải pháp được đề xuất. Bên cạnh đó, độ phức tạp tính toán cũng được xem xét như một yếu tố khi lựa chọn mô hình cho khả năng mở rộng với dữ liệu lớn.

Bảng 2. Kết quả thực nghiệm các mô hình học máy với dữ liệu gốc

Model	Class	Re-call	F1-score Weight-ed	Acc	AUC
RF	0	0,95	0,71	0,75	0,62
	1	0,27			
GP	0	0,96	0,72	0,76	0,62
	1	0,28			
SVC	0	0,98	0,70	0,76	0,60
	1	0,22			
LR	0	0,97	0,72	0,76	0,62
	1	0,27			
AdaBoost	0	0,96	0,69	0,74	0,59
	1	0,22			
XGBoost	0	0,97	0,73	0,76	0,64
	1	0,37			
Gradient Boosting	0	0,97	0,77	0,77	0,63
	1	0,30			
MLP	0	0,92	0,74	0,76	0,66
	1	0,40			

Bảng 2 trình bày kết quả thực nghiệm dự báo với dữ liệu gốc của nhiều mô hình học máy khác nhau từ các mô hình học máy đơn lẻ như SVC, LR đến các mô hình học kết hợp và các kỹ thuật Boosting như AdaBoost, XGBoost và GradientBoosting. Kết quả từ bảng cho thấy, nhìn chung các mô hình học kết hợp như RF và các kỹ thuật Boosting cho hiệu quả dự báo tốt hơn ở hầu hết các chỉ số đánh giá mô hình, đặc biệt các kỹ thuật Boosting cho hiệu quả tốt hơn đáng kể so với các mô hình dự báo còn lại. Về chỉ số độ chính xác (accuracy), hiệu quả các mô hình dự báo khá tương đồng nhau với mức trung bình 76%. Tuy nhiên, về F1-score, các mô hình Boosting (Gradient Boosting, XGBoost) và MLP cho hiệu suất vượt trội so với các mô hình còn lại. Gradient Boosting đạt cao nhất (77%), tiếp theo là MLP (74%) và XGBoost (73%), trong khi SVC và AdaBoost thấp nhất trung bình 70%. Đối với recall của lớp 1, tất cả mô hình đều đạt mức thấp, dao động từ 22%-40%, cho thấy khả năng nhận diện lớp này còn hạn chế. MLP có recall cao nhất (40%), tiếp theo là XGBoost (37%) và Gradient Boosting (30%). Ngược lại, SVC và AdaBoost ghi nhận mức thấp nhất, chỉ khoảng 22%.

Về chỉ số AUC, hiệu suất các mô hình dao động từ 59%-66%. Mô hình MLP đạt cao nhất (66%), tiếp theo là XGBoost (64%) và Gradient Boosting (63%). Nhóm RF, LR và GP ở mức trung bình khoảng 62%, trong khi SVC và AdaBoost thấp nhất, gần 60%. Nhìn chung, với dữ liệu phức tạp như thời tiết và lượng mưa, các mô hình học kết hợp và kỹ thuật Boosting (đặc biệt XGBoost, Gradient Boosting) cùng MLP cho hiệu suất vượt trội hơn so với các mô hình còn lại. Bên cạnh độ chính xác của mô hình dự báo, độ phức tạp tính toán cũng là một khía cạnh cần cân nhắc khi lựa chọn mô hình cho bài toán thực tế với dữ liệu lớn, phức tạp. Về khía cạnh độ phức tạp tính toán, có thể thấy rõ sự đánh đổi giữa hiệu quả mô hình và chi phí. LR có độ phức tạp thấp ($O(n \times d)$) nhưng hiệu quả hạn chế, trong khi SVC có độ phức tạp cao ($O(n^2) - O(n^3)$) nhưng hiệu quả chưa tương xứng. RF có độ phức tạp trung bình ($O(T \times n \times \log n)$) và khả năng mở rộng tốt. Ngược lại, các mô hình Boosting (Gradient Boosting, XGBoost) và MLP có độ phức tạp cao hơn ($O(T \times n \times d)$) nhưng đạt hiệu suất vượt trội hơn. Nhìn chung, hiệu năng cao thường đi kèm chi phí tính toán lớn, do đó cần cân nhắc giữa hiệu quả và tài nguyên khi lựa chọn mô hình.

Bảng 3. Kết quả thực nghiệm các mô hình dự báo với giải pháp cải tiến dữ liệu đề xuất

Model	Class	Dữ liệu gốc				Dữ liệu mới			
		Recall	F1-score Weighted	Acc	AUC	Recall	F1-score Weighted	Acc	AUC
RF	0	0,95	0,71	0,75	0,62	0,92	0,79	0,81	0,72
	1	0,27				0,52			
GP	0	0,96	0,72	0,76	0,62	0,92	0,77	0,78	0,69
	1	0,28				0,46			
SVC	0	0,98	0,70	0,76	0,60	0,94	0,77	0,79	0,69
	1	0,22				0,43			
LR	0	0,97	0,72	0,76	0,62	0,93	0,76	0,78	0,68
	1	0,27				0,42			
AdaBoost	0	0,96	0,69	0,74	0,59	0,91	0,78	0,79	0,71
	1	0,22				0,52			
XGBoost	0	0,97	0,73	0,76	0,64	0,91	0,86	0,86	0,83
	1	0,37				0,75			
Gradient Boosting	0	0,97	0,77	0,77	0,63	0,92	0,83	0,83	0,77
	1	0,30				0,62			
MLP	0	0,92	0,74	0,76	0,66	0,89	0,82	0,83	0,78
	1	0,40				0,66			

Bảng 3 cho thấy khi sử dụng dữ liệu mới, hiệu quả dự báo được cải thiện rõ rệt ở tất cả các mô hình so với dữ liệu gốc. Với dữ liệu gốc, các mô hình chỉ đạt mức hiệu suất trung bình, với accuracy và weighted F1-score dao động 69%-77%, trong đó Gradient Boosting cao nhất, tiếp theo là GP, SVC, LR, XGBoost và MLP. Tuy nhiên, Recall và AUC đều thấp, đặc biệt Recall của lớp 1 chỉ khoảng 22%-40%, hạn chế khả năng nhận diện chính xác các trường hợp thuộc lớp 1 - yếu tố rất quan trọng trong dự báo lượng mưa hiệu quả. Chỉ số AUC trung bình chỉ đạt khoảng 60%.

Theo giải pháp cải tiến dữ liệu đề xuất, như thể hiện trong bảng, mô hình XGBoost đạt hiệu suất dự báo cao nhất, lần lượt đạt 86%, 86% và 83% về độ chính xác, điểm F1 có trọng số và chỉ số AUC tương ứng. Các mô hình Gradient Boosting, MLP đều cho kết quả tốt, mỗi mô hình đều đạt trung bình 83% về độ chính xác và điểm F1 có trọng số. Sự cải thiện nổi bật nhất được ghi nhận ở mô hình XGBoost, với accuracy và weighted F1-score cùng tăng khoảng 15,5% (từ 73% và 76% lên 86%). Đặc biệt, Recall lớp 1

tăng mạnh từ 37% lên 75% ($\approx 103\%$) và AUC tăng từ 64% lên 83% (gần 30%), cho thấy khả năng nhận diện và phân biệt lớp 1 được cải thiện rõ rệt khi sử dụng dữ liệu mới. Các mô hình RF, Gradient Boosting và MLP cũng đạt mức cải thiện đáng kể trên nhiều chỉ số. Nhóm GP, LR và SVC ghi nhận mức cải thiện đáng kể nhưng thấp hơn, với Recall tăng từ 22% lên 52%, weighted F1-score từ 70% lên 77% và AUC từ 62% lên 69%. Nhìn chung, dữ liệu mới giúp nâng cao hiệu quả mô hình, trong đó XGBoost thể hiện ưu thế vượt trội.

4. KẾT LUẬN

Dự báo lượng mưa chính xác giữ vai trò then chốt trong nhiều hoạt động của con người, như quản lý tài nguyên nước, quy hoạch đô thị, lập kế hoạch sản xuất đặc biệt trong lĩnh vực nông nghiệp. Bài báo này chúng tôi tập trung vào việc nghiên cứu dữ liệu, phân tích đánh giá hiệu quả các mô hình học máy trong dự báo lượng mưa ở khu vực cụ thể để có thể cung cấp góc nhìn chuyên sâu hơn về xu hướng thời tiết

nói chung, xu hướng lượng mưa nói riêng tại khu vực nghiên cứu. Kết quả nghiên cứu cho thấy với các dữ liệu phức tạp như dữ liệu về lượng mưa, các mô hình học kết hợp như RF và các kỹ thuật boosting cho hiệu quả dự báo tốt hơn các mô hình khác ở hầu hết các chỉ số đánh giá, đặc biệt các kỹ thuật boosting. Tuy nhiên, hiệu quả này đi kèm với độ phức tạp và chi phí tính toán cao hơn. Vì vậy, tùy theo yêu cầu bài toán, cần lựa chọn mô hình phù hợp để cân bằng giữa độ chính xác dự báo và chi phí tính toán; trong trường hợp ưu tiên độ chính xác, XGBoost và MLP là những lựa chọn cho hiệu quả dự báo lượng mưa vượt trội.

Bên cạnh đó, để cải thiện hiệu suất các mô hình dự báo, chúng tôi giới thiệu một giải pháp cải tiến dữ liệu đầu vào của các mô hình học máy, cho phép tạo ra các thông tin mới từ dữ liệu gốc nhằm phản ánh tốt hơn mối quan hệ phức tạp, đặc biệt các mối quan hệ thời gian trong chuỗi dữ liệu chuỗi thời gian khí tượng. Kết quả nghiên cứu cho thấy hiệu suất dự báo của các mô hình được cải thiện đáng kể khi sử dụng các thông tin mới so với mô hình chỉ dựa trên dữ liệu ban đầu. XGBoost là mô hình ghi nhận mức cải thiện tốt nhất với 17,8% độ chính xác và 13,2% điểm F1 có trọng số. Nghiên cứu góp phần thúc đẩy việc phát triển các giải pháp nâng cao độ chính xác của mô hình dự báo và ước lượng lượng mưa, một lĩnh vực mà việc dự báo chính xác vẫn còn nhiều thách thức.

Trong tương lai, hướng nghiên cứu có thể mở rộng với các bộ dữ liệu lớn hơn cả về phạm vi không gian và thời gian. Hơn thế, các giải pháp được đề xuất trong nghiên cứu này có thể được ứng dụng trong những lĩnh vực khác có đặc tính dữ liệu tương tự.

TÀI LIỆU THAM KHẢO

- Abbot J. & Marohasy J. (2014). Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmospheric Research*. 138: 166-178.
- Baljon M. & Sharma S.K. (2023). Rainfall prediction rate in Saudi Arabia using improved machine learning techniques. *Water*. 15(4): 826.
- Công Thanh, Nguyễn Như Quý & Mai Văn Khiêm (2018). Đánh giá ước lượng mưa từ độ phản hồi radar Nhà Bè. *Tạp chí Khoa học ĐHQGHN: Các Khoa học Trái đất và Môi trường*. 34: 10-17.
- Cramer S., Kampouridis M., Freitas A.A. & Alexandridis A.K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*. 85: 169-181.
- Dương Thị Hà & Nguyễn Thái Nghe (2022). Ứng dụng mô hình đa biến bộ nhớ dài-ngắn hạn trong dự báo nhiệt độ và lượng mưa. *Tạp chí Khoa học Đại học Cần Thơ*. 58(4): 8-16.
- Endalieu D., Haile G. & Taye W. (2022). Deep learning model for daily rainfall prediction: case study of Jimma, Ethiopia. *Water Supply*. 22(3): 3448-3461.
- Giang N.H., Wang Y., Hieu T.D., Tho Q.T., Phuong L.A. & Tu Do H.N. (2021). Toward rainfall prediction by machine learning in Perfume River Basin, Thua Thien Hue Province, Vietnam. *Open Geosciences*. 13(1): 963-976.
- Hai N.H., Quynh T.D. & Le H.M. (2025). A Novel Feature Engineering Method for Rainfall Prediction. In *International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences*. Cham: Springer Nature Switzerland. pp. 130-141.
- Kundu S., Biswas S.K., Tripathi D., Karmakar R., Majumdar S. & Mandal S. (2023). A review on rainfall forecasting using ensemble learning techniques. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*. 6: 100296.
- Mojtahedi F.F., Yousefpour N., Chow S.H. & Cassidy M. (2025). Deep learning for time series forecasting: Review and applications in geotechnics and geosciences. *Archives of Computational Methods in Engineering*. pp. 1-31.
- Nguyễn Hoàng Minh, Phùng Tiến Dũng, Vũ Thị Thanh Vân, Hoàng Văn Đại, Mai Văn Khiêm & Nguyễn Phương Nhung (2022). Nghiên cứu hiệu chỉnh lượng mưa ước lượng từ radar phục vụ nâng cao chất lượng mô phỏng dòng chảy. *Tạp chí Khí tượng Thủy văn*. 743(11): 23-35.
- Pirone D., Cimorelli L., Del Giudice G. & Pianese D. (2023). Short-term rainfall forecasting using cumulative precipitation fields from station data: a probabilistic machine learning approach. *Journal of Hydrology*. 617: 128949.
- Saleh M.A., Rasel H.M. & Ray B. (2024). A comprehensive review towards resilient rainfall forecasting models using artificial intelligence techniques. *Green Technologies and Sustainability*. 2(3): 100104.
- Yu P.S., Yang T.C., Chen S.Y., Kuo C.M. & Tseng H.W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of hydrology*. 552: 92-104.