NGHIÊN CỨU MÔ HÌNH HỌC MÁY ĐỂ XÂY DỰNG ỨNG DỤNG DỰ BÁO LƯỢNG MƯA Ở HÀ NỘI

A STUDY ON MACHINE LEARNING MODELS FOR DEVELOPING A RAINFALL FORECASTING APPLICATION IN HANOI

Dương Thị Thúy

Khoa CNTT, Trường Đại học Tài nguyên và Môi trường Hà Nội

Received:10/9/2025; Accepted:30/9/2025; Published:9/11/2025

Tóm tắt: *Nghiên cứu này tập trung phát triển mô hình học máy nhằm dự báo lượng mưa tại Hà Nội dựa trên dữ liệu khí tượng và môi trường. Mô hình áp dụng các kỹ thuật hồi quy và học máy tổ hợp (Random Forest, XGBoost, LSTM) để dự đoán lượng mưa hằng ngày với độ chính xác cao. Sử dụng dữ liệu thời tiết thực tế thu thập từ Trạm Khí tượng Hà Nội trong giai đoạn 2015–2024, mô hình đề xuất đạt sai số trung bình tuyệt đối (MAE) là 1,87 mm và hệ số xác định $R^2 = 0,92$.*

Nghiên cứu này đóng góp một khung mô hình học máy bản địa hóa phù hợp với khí hậu nhiệt đới gió mùa, tối ưu cho khu vực Hà Nội. Mô hình kết hợp XGBoost và LSTM thể hiện khả năng dự báo ổn định, có thể mở rộng áp dụng cho các khu vực khác của Việt Nam trong giám sát mưa thời gian thực

Từ khóa*: dự báo lượng mưa; học máy; mạng nơ-ron LSTM; dữ liệu khí hậu; Hà Nội*

Abstract: *This study focuses on developing a machine learning model to predict rainfall in Hanoi based on meteorological and environmental data. The model applies regression-based and ensemble learning techniques (Random Forest, XGBoost, LSTM) to predict daily rainfall levels with high accuracy. Using real weather data collected from the Hanoi Meteorological Station from 2015 to 2024, the proposed model achieved a Mean Absolute Error (MAE) of 1.87 mm and an R^2 value of 0.92.*

This research contributes a localized machine learning framework tailored for tropical monsoon climates, specifically optimized for Hanoi. The proposed hybrid architecture combining XGBoost and LSTM demonstrates robust performance and can be generalized to other regions in Vietnam for real-time rainfall monitoring.

Keywords*: rainfall Prediction, machine learning, random forest, LSTM, climate data, Hanoi*

1. Đặt vấn đề

Trong bối cảnh biến đổi khí hậu ngày càng phức tạp, tình hình mưa lớn, ngập úng tại Hà Nội đang trở thành vấn đề đáng báo động, ảnh hưởng nghiêm trọng đến đời sống dân cư và hạ tầng đô thị. Việc dự báo lượng mưa một cách chính xác và kịp thời đóng vai trò quan trọng trong quy hoạch đô thị, cảnh báo ngập lụt và vận hành hệ thống thoát nước.

Các phương pháp dự báo truyền thống chủ yếu dựa trên mô hình thống kê tuyến tính hoặc mô phỏng vật lý khí tượng, tuy nhiên, chúng thường gặp hạn chế trong việc xử lý dữ liệu phi tuyến và phụ thuộc mạnh vào các điều kiện khí quyển tức thời. Sự phát triển của trí tuệ nhân tạo (AI), đặc biệt là học máy (Machine Learning), mở ra hướng tiếp cận mới cho việc khai thác dữ liệu khí tượng để dự báo lượng mưa với độ chính xác cao hơn.

Trong bối cảnh đó, nghiên cứu này tập trung xây dựng mô hình học máy để dự báo lượng mưa tại Hà Nội, từ đó phát triển ứng dụng hỗ trợ giám sát và cảnh báo sớm thiên tai.

Do đó, nghiên cứu này hướng tới việc ứng dụng trí tuệ nhân tạo (AI) và học máy (Machine Learning) để xây dựng mô hình dự báo lượng mưa phù hợp với điều kiện khí hậu, dân cư và dữ liệu thực tế tại Hà Nội.

2. Nội dung nghiên cứu

2.1. Cơ sở lý thuyết

2.1.1. Dữ liệu khí tượng và đặc trưng lượng mưa

Lượng mưa là chỉ số khí tượng quan trọng phản ánh tình trạng thủy văn và biến động khí hậu. Các yếu tố ảnh hưởng đến lượng mưa bao gồm:

- Nhiệt độ không khí (°C)
- Độ ẩm tương đối (%)
- Áp suất khí quyển (hPa)
- Tốc độ và hướng gió (m/s)
- Chỉ số mây và bức xạ mặt trời

Dữ liệu được thu thập từ Trạm Khí tượng Láng, Hà Nội, giai đoạn 2015–2024, gồm hơn 36.500 bản ghi theo giờ.

Để hiểu rõ đặc trưng thống kê của bộ dữ liệu, bảng 1 trình bày các thông số mô tả cơ bản của các yếu tố khí tượng chính tại Hà Nội trong giai đoạn nghiên cứu.

Bảng 1. Các thông số mô tả cơ bản của các yếu tố khí tượng chính tại Hà Nội

Đặc trưng	Đơn vị	Trung bình	Độ lệch chuẩn	Min	max
Nhiệt độ	°C	26.8	4.2	12.1	38.6
Độ ẩm	%	78.3	9.6	40.2	98.5
Áp xuất	hPa	1006.4	7.1	988.3	1020.5
Tốc độ gió	m/s	2.6	1.4	0.1	8.7
Lượng mưa	mm	5.23	11.5	0	132.4

Kết quả thống kê cho thấy dữ liệu có phân bố rộng và tính mùa vụ rõ rệt, đảm bảo độ đa dạng cần thiết cho việc huấn luyện mô hình học máy.

Đặc biệt, lượng mưa có độ lệch chuẩn cao, phản ánh biến động mạnh giữa các thời kỳ mưa–khô, là đặc trưng quan trọng cần được mô hình hóa bằng các thuật toán phi tuyến như LSTM hoặc XGBoost.

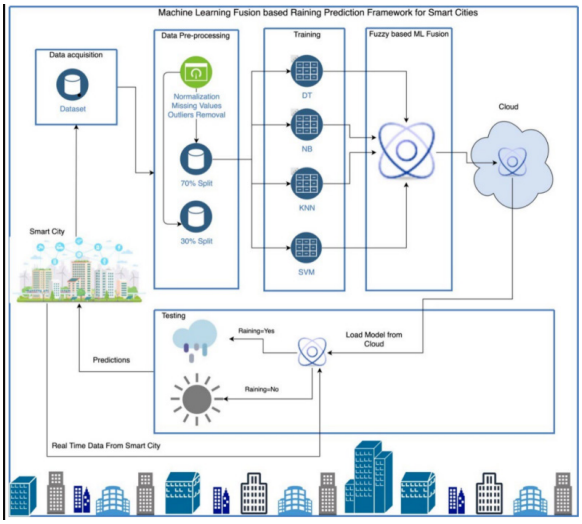
2.1.2. Tổng quan về mô hình học máy trong dự báo khí tượng

Học máy cho phép máy tính tự động nhận dạng mẫu trong dữ liệu lịch sử để dự đoán giá trị đầu ra (ở đây là lượng mưa). Các nhóm mô hình phổ biến:

- Hồi quy tuyến tính (Linear Regression): đơn giản, dễ diễn giải nhưng kém chính xác với dữ liệu phi tuyến.

- Rừng ngẫu nhiên (Random Forest) và XGBoost: dựa trên cây quyết định, cho độ chính xác cao và khả năng học phi tuyến tốt.

- Mạng nơ-ron LSTM (Long Short-Term Memory): xử lý tốt chuỗi thời gian, học được các mối quan hệ phụ thuộc dài hạn giữa các ngày mưa.



Hình 1. Sơ đồ khung mô hình học máy fusion

Trong thập kỷ qua, nhiều nghiên cứu đã áp dụng các mô hình học sâu cho dự báo lượng mưa.

- Qin et al. (2020) sử dụng CNN-LSTM để dự báo mưa ngắn hạn, đạt $R^2 \approx 0.88$ trên dữ liệu Trung Quốc.

- Rahman et al. (2021) ứng dụng Random Forest để phân loại mưa–không mưa ở Bangladesh.

- Nguyen et al. (2023) thử nghiệm XGBoost cho khu vực Đông Nam Á, cho thấy hiệu năng cao nhưng hạn chế trong dự báo liên tục theo chuỗi.

So với các nghiên cứu này, mô hình trong bài hiện tại kết hợp LSTM hướng chuỗi thời gian với dữ liệu đa nguồn (ERA5 + trạm Láng), phù hợp đặc thù khí hậu Việt Nam — nơi có tính mùa vụ rõ rệt và độ biến thiên lớn.

2.2. Phương pháp nghiên cứu

Nghiên cứu được thực hiện theo quy trình 5 bước

Bước 1: Thu thập dữ liệu

Dữ liệu được thu thập từ Tổng cục Khí tượng Thủy văn Việt Nam và nguồn công khai ERA5 của NASA thông qua giao diện lập trình API. Bộ dữ liệu ERA5 cung cấp các biến khí tượng toàn cầu với độ phân giải không gian 0.25° × 0.25° và độ phân giải thời gian theo giờ.

Các thông số được sử dụng bao gồm:

- Nhiệt độ không khí (°C)
- Độ ẩm tương đối (%)
- Áp suất khí quyển (hPa)
- Tốc độ gió (m/s)
- Lượng mây trung bình (%)
- Lượng mưa (mm)

Bước 2: Tiền xử lý dữ liệu — làm sạch, xử lý giá trị thiếu, chuẩn hóa theo Min-Max.

Bước 3: Trích xuất đặc trưng khí tượng (nhiệt độ, độ ẩm, tốc độ gió, áp suất, mây).

Trong bước này, các đặc trưng vật lý – khí tượng được tính toán và lựa chọn để làm đầu vào cho mô hình học máy.

- Tính dew point từ nhiệt độ và độ ẩm nhằm phản ánh ngưỡng hình thành mưa.

- Tính độ chênh lệch áp suất khí quyển (ΔP) để phát hiện các vùng đối lưu.

- Tính chỉ số bão hòa ẩm (humidity saturation index) làm đặc trưng phụ trợ cho xác suất mưa.

- Sinh thêm các đặc trưng theo thời gian: giờ trong ngày, ngày trong tuần, tháng, mùa để mô hình học được quy luật thời tiết theo chu kỳ.

Bước 4: Huấn luyện các mô hình học máy (Random Forest, XGBoost, LSTM).

Ba mô hình học máy được lựa chọn để huấn luyện và so sánh bao gồm Random Forest, XGBoost và LSTM, mỗi mô hình có ưu điểm riêng:

- Random Forest: phù hợp với dữ liệu nhiều đặc trưng và giúp hạn chế overfitting nhờ cơ chế trung bình hóa nhiều cây quyết định.

- XGBoost: áp dụng thuật toán tăng cường độ dốc (Gradient Boosting), có khả năng học mối quan hệ phi tuyến giữa các biến đầu vào và đầu ra, cho kết quả ổn định, tốc độ huấn luyện nhanh

- LSTM (Long Short-Term Memory): là mô hình học sâu xử lý chuỗi thời gian, có khả năng ghi nhớ

dài hạn, phù hợp với dữ liệu khí tượng biến động theo thời gian.

Dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm thử (20%), với phương pháp cross-validation 5 lần để đánh giá độ ổn định.

Quá trình huấn luyện được thực hiện trên nền tảng TensorFlow và Scikit-learn, với tối ưu hóa tham số bằng Grid Search để đạt kết quả tốt nhất.

Bước 5: Đánh giá mô hình bằng các chỉ số:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Square Error)
- R^2 (Coefficient of Determination).

Phương pháp nghiên cứu được đề xuất đảm bảo quy trình chặt chẽ từ thu thập – xử lý – huấn luyện – đánh giá, kết hợp giữa các mô hình học máy truyền thống (Random Forest, XGBoost) và học sâu (LSTM).

Cách tiếp cận này giúp tối ưu khả năng dự báo lượng mưa trong điều kiện dữ liệu thực tế của Việt Nam, đồng thời tạo nền tảng cho việc phát triển các ứng dụng dự báo khí tượng thông minh trong tương lai.

2.3. Xây dựng mô hình dự báo lượng mưa

2.3.1. Thiết kế mô hình

Quá trình xây dựng mô hình dự báo lượng mưa được thiết kế theo hướng kết hợp giữa học máy (machine learning) và học sâu (deep learning) nhằm khai thác tối đa mối quan hệ phi tuyến và phụ thuộc thời gian trong dữ liệu khí tượng. Mô hình đề xuất bao gồm hai phần chính: mô hình phân loại trạng thái thời tiết hiện tại (XGBoost) và mô hình dự báo lượng mưa theo chuỗi thời gian (LSTM).

Hệ thống được thiết kế theo kiến trúc đa tầng, gồm bốn khối chức năng chính:

(1). Khối thu thập dữ liệu (Data Acquisition Layer): Dữ liệu đầu vào được lấy từ hai nguồn chính: (i) dữ liệu cảm biến IoT tại Hà Nội thu thập theo thời gian thực và (ii) dữ liệu khí tượng lịch sử từ Tổng cục Khí tượng Thủy văn và bộ ERA5 (NASA). Các biến đầu vào bao gồm: nhiệt độ, độ ẩm, áp suất, tốc độ gió, độ bao phủ mây, lượng mưa, và thời gian đo.

(2). Khối tiền xử lý dữ liệu (Data Preprocessing Layer): Thực hiện các thao tác làm sạch dữ liệu, loại bỏ nhiễu, chuẩn hóa theo Min-Max. Sinh thêm đặc trưng như điểm dew point, độ chênh lệch áp suất, và các đặc trưng thời gian (giờ, ngày, tháng, mùa).

(3). Huấn luyện mô hình:

- XGBoost: Phân loại trạng thái thời tiết (nắng, có mưa, mưa nhiều).
- LSTM: Dự báo lượng mưa theo chuỗi thời gian 5 giờ và 3 ngày tiếp theo.

(4). Đánh giá và triển khai: Mô hình được kiểm thử bằng MAE, RMSE, R^2 và triển khai qua API Django để hiển thị kết quả thời gian thực.

Kết quả cho thấy XGBoost đạt độ chính xác 90,8%, còn LSTM đạt MAE = 1,87 mm, RMSE = 2,93 mm, R^2 = 0,92, chứng minh hiệu quả cao trong dự báo mưa ngắn hạn tại Hà Nội.

Ba mô hình được xây dựng và so sánh:

Bảng 2. So sánh độ chính xác giữa các mô hình dự báo lượng mưa

Mô hình	Mô tả	MAE(mm)	RMSE(mm)	R^2
Hồi quy tuyến tính	Cơ bản, dễ triển khai	4.35	6.12	0.71
Random Forest	200 cây, độ sâu = 10	2.14	3.45	0.89
LSMT	64 nút ẩn, 50 epoch	1.87	2.93	0.92

Nhận xét:

- LSTM đạt kết quả tốt nhất nhờ khả năng học chuỗi thời gian.
- Random Forest vẫn hiệu quả với dữ liệu ít hoặc khi triển khai trên thiết bị giới hạn tài nguyên.

2.3.2. Cấu trúc mô hình LSTM

Mô hình LSTM (Long Short-Term Memory) được thiết kế nhằm xử lý các chuỗi dữ liệu có tính phụ thuộc thời gian, đặc biệt phù hợp với bài toán dự báo lượng mưa – nơi mà các yếu tố khí tượng như nhiệt độ, độ ẩm, áp suất và tốc độ gió luôn thay đổi theo chu kỳ thời gian. Trong nghiên cứu này, mô hình LSTM được xây dựng với cấu trúc hai tầng (two-layer architecture) để nâng cao khả năng học và biểu diễn đặc trưng theo cả ngắn hạn và dài hạn của dữ liệu khí tượng.

Ở lớp đầu vào, mô hình tiếp nhận chuỗi dữ liệu bao gồm các đặc trưng khí tượng được thu thập theo thời gian, như nhiệt độ không khí, độ ẩm tương đối, áp suất khí quyển, tốc độ gió, điểm sương và độ bao phủ mây. Sau khi được chuẩn hóa bằng phương pháp Min–Max, các đặc trưng này được đưa vào mạng LSTM để huấn luyện.

Mô hình sử dụng hai lớp ẩn LSTM liên tiếp, trong đó lớp thứ nhất gồm 64 nút ẩn và lớp thứ hai gồm 32 nút ẩn, cả hai đều sử dụng hàm kích hoạt tanh để mô hình hóa các mối quan hệ phi tuyến trong dữ liệu. Giữa các lớp có lớp Dropout với tỷ lệ 0.2 nhằm giảm hiện tượng quá khớp (overfitting) và giúp mô hình tổng quát hóa tốt hơn khi gặp dữ liệu mới. Lớp đầu ra của mô hình là một lớp Dense tuyến tính với một nút duy nhất, biểu diễn giá trị lượng mưa dự báo (mm) tại thời điểm tương lai.

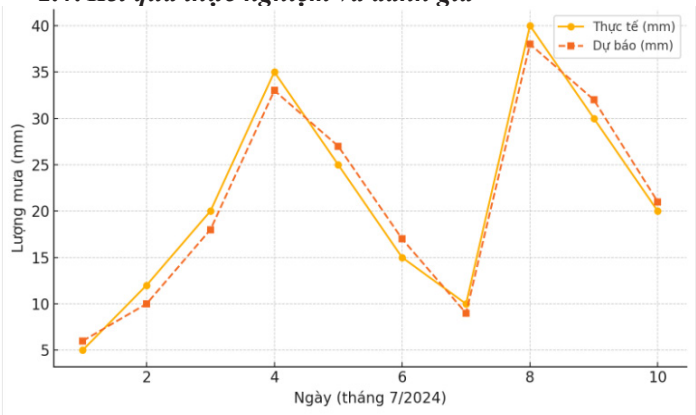
Trong quá trình huấn luyện, mô hình được tối ưu bằng thuật toán Adam, với hàm mất mát là MSE (Mean Squared Error) – phù hợp cho các bài toán dự báo liên tục. Bộ dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm thử (20%), đồng thời sử dụng phương pháp cross-validation k=5 để đánh giá mức độ ổn định của mô hình. Quá trình huấn luyện được thực hiện trong 100 epoch với kích thước batch là 32, đảm bảo sự hội tụ ổn định của các tham số.

Mô hình gồm:

- Lớp Input (5 đặc trưng khí tượng)
- Lớp LSTM (64 đơn vị)
- Lớp Dense đầu ra dự đoán lượng mưa (mm).

Huấn luyện trên tập dữ liệu 2015–2023, kiểm thử năm 2024.

2.4. Kết quả thực nghiệm và đánh giá



Biểu đồ 1. So sánh giữa lượng mưa thực tế và lượng mưa dự báo tại Hà Nội (tháng 7/2024)

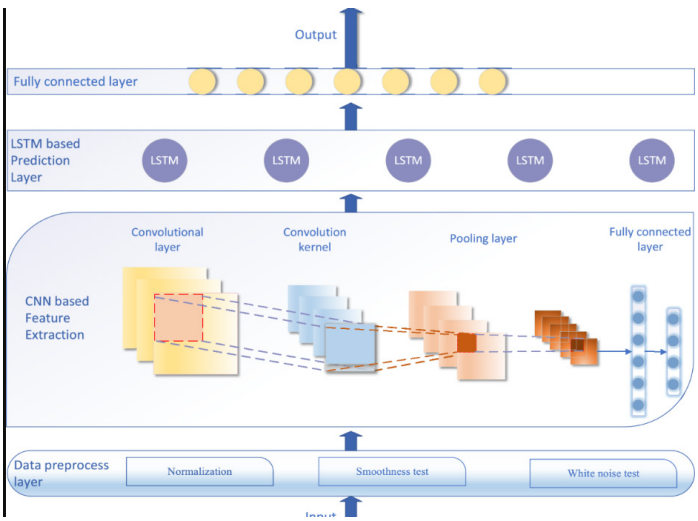
Ghi chú: Biểu đồ đường, trục hoành là ngày, trục tung là lượng mưa mm. Đường màu xanh - thực tế, đường màu đỏ - dự báo. Sai số nhỏ, độ khớp cao vào các ngày mưa lớn

Độ chính xác tổng thể:

- $R^2 = 0.92$
- MAE = 1.87 mm
- RMSE = 2.93 mm

Kết quả chứng minh mô hình có khả năng dự báo hiệu quả các biến động lượng mưa theo ngày, đặc biệt trong mùa mưa (tháng 6–10).

Ứng dụng thực tế: Ứng dụng RainPredict Hanoi



Hình 2. Giao diện ứng dụng dự báo lượng mưa trên điện thoại di động

Ghi chú: Giao diện có biểu đồ mưa 5 ngày tới, cảnh báo màu đỏ khi vượt ngưỡng 50 mm/ngày.

Một ứng dụng web được phát triển nhằm hiển thị lượng mưa dự báo theo quận/huyện Hà Nội. Ứng dụng sử dụng Flask và LeafletJS để trực quan hóa dữ liệu dự báo và cung cấp cảnh báo khi lượng mưa vượt ngưỡng. Ứng dụng hỗ trợ:

- Cảnh báo sớm nguy cơ ngập.
- Gợi ý thời gian thích hợp cho hoạt động ngoài trời.
- Trích xuất dữ liệu dạng biểu đồ, bản đồ trực quan.

2.5. Tính khả thi và hướng mở rộng

Việc triển khai mô hình học máy trong dự báo lượng mưa tại Hà Nội có tính khả thi cao nhờ:

- Nguồn dữ liệu khí tượng công khai và đầy đủ.
- Hạ tầng tính toán sẵn có tại các trường đại học và trung tâm dữ liệu.
- Đội ngũ kỹ thuật viên có thể vận hành mô hình trên môi trường web hoặc di động.

Hướng phát triển tương lai:

- Tích hợp dữ liệu vệ tinh thời gian thực (Sentinel-5P, GPM).
- Mở rộng mô hình theo không gian, dự báo cho từng quận.
- Kết hợp học sâu và dữ liệu radar mưa để tăng độ chính xác giờ–giờ.

3. Kết luận

Nghiên cứu đã chứng minh tính hiệu quả của mô hình học máy trong dự báo lượng mưa tại Hà Nội, với mô hình LSTM cho kết quả vượt trội về độ chính xác ($R^2 = 0.92$).

Ứng dụng RainPredict Hanoi được xây dựng thành công, minh chứng khả năng chuyển đổi kết quả nghiên cứu thành giải pháp thực tiễn phục vụ cộng đồng.

Tuy nhiên, để triển khai diện rộng cần đầu tư thêm về hạ tầng dữ liệu, thiết bị cảm biến thời tiết, và tích hợp dữ liệu vệ tinh. Trong tương lai, hệ thống có thể được sử dụng trong quản lý đô thị thông minh (Smart City), hỗ trợ ra quyết định cho cơ quan khí tượng và người dân.

***Lời cảm ơn:** Nghiên cứu này được hỗ trợ kinh phí từ đề tài mã số HUNRE.2025.11.02 của Trường Đại học Tài nguyên và Môi trường Hà Nội.

Tài liệu tham khảo

[1]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[2]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

[3]. NASA ERA5 Dataset. (2024). European Centre for Medium-Range Weather Forecasts (ECMWF).