

ỨNG DỤNG MÔ HÌNH MULTIVARIATE LSTM ĐỂ DỰ BÁO NHIỆT ĐỘ VÀ LƯỢNG MƯA USING MULTIVARIATE LSTM MODEL IN FORECASTING TEMPERATURE AND RAINFALL

DƯƠNG THỊ HÀ¹, NGUYỄN THÁI NGHE²

¹Học viên Cao học, Trường Đại học SPKT Vĩnh Long

²Trường Đại học Cần Thơ

Email: laphathu80@gmail.com, ntnghe@cit.ctu.edu.vn

Nhận bài (Received): 26/4/2022; Phản biện (Reviewed): 03/6/2022; Chấp nhận (Accepted): 20/6/2022

TÓM TẮT:

Dự báo nhiệt độ và lượng mưa là vấn đề quan trọng trong lĩnh vực nông nghiệp nhằm hỗ trợ người dân có kế hoạch gieo trồng phù hợp. Một số kỹ thuật trước đây đã được đề xuất dựa trên phân tích thống kê, máy học và kỹ thuật học sâu. Bài viết này đề xuất phương pháp xây dựng mô hình dự báo nhiệt độ và lượng mưa hàng tháng bằng mô hình đa biến bộ nhớ ngắn hạn dài hạn (Multivariate Long Short-Term Memory- MLSTM). Các tham số của mô hình được điều chỉnh sao cho phù hợp với bài toán đặt ra. Mô hình được đánh giá thông qua đo độ lỗi RMSE và MAE. Bên cạnh đó, các mô hình dự báo khác (như LSTM, MLP, SVR) cũng được sử dụng nhằm so sánh hiệu quả của mô hình đề xuất. Kết quả thực nghiệm trên tập dữ liệu nhiệt độ và lượng mưa trung bình hàng tháng tại Việt Nam từ 1901 đến 2015 cho thấy mô hình MLSTM đạt hiệu quả khá tốt với độ lỗi RMSE trên tập nhiệt độ là 1.311 và MAE là 1.051, tương ứng trên tập dữ liệu lượng mưa là 2.299 và 2.450.

Từ khóa: Dự báo nhiệt độ, Dự báo lượng mưa, Máy học, học sâu, Mô hình Multivariate LSTM, LSTM.

ABSTRACT:

Forecasting temperature and rainfall is an important issue in the agricultural sector to assist farmers in planning appropriate planting. Several techniques have previously been proposed based on statistical analysis, learning machine and deep learning techniques. This paper proposes a method to build a model to forecast monthly temperature and rainfall using the Multivariate Long Short-Term Memory (MLSTM) model. The parameters of the model are adjusted to suit the proposed problem. The model is evaluated using RMSE and MAE measurements. Besides, other forecasting models (such as LSTM, MLP, and SVR) are also used to compare the effectiveness of the proposed model. Experimental results on the data set of monthly average temperature and rainfall in Vietnam from 1901 to 2015 show that the MLSTM model is quite effective with the RMSE on the temperature set of 1.311 and the MAE of 1.051, respectively on the rainfall data set of 2.299 and 2.450.

Keywords: Temperature forecasting, rainfall forecasting, learning machine, deep learning, Multivariate LSTM, LSTM model.

1. GIỚI THIỆU

Biến đổi khí hậu ảnh hưởng đến mọi mặt đời sống kinh tế - xã hội của con người. Vì vậy việc dự báo các yếu tố khí hậu đang ngày càng quan trọng và cần thiết, trở thành mối quan tâm lớn của tất cả các quốc gia trên thế giới. Do ảnh hưởng của biến đổi khí hậu ngày càng sâu sắc đến hầu hết các nước trên thế giới trong đó Việt Nam cũng là một trong những quốc gia chịu ảnh hưởng nặng nề của biến đổi khí hậu. Nhiệt độ và lượng mưa là hai yếu tố chịu tác động lớn trong hệ thống khí hậu. Nó có ảnh hưởng trực tiếp đến hệ sinh thái, nước, tài nguyên, các hoạt động nông nghiệp và nuôi trồng thủy sản là những yếu tố chính dẫn đến sự phát triển của vùng đồng bằng Sông Cửu Long nói riêng và các tỉnh miền trung nói chung. Theo thống kê Báo điện tử, 'các trận mưa bão lũ năm 2020-2021 đã gây ra nhiều hậu quả nặng nề cho người dân miền trung với lượng mưa 3000 mm, có nơi lên đến 4526 mm, nhiều ngôi nhà bị sập, hư hỏng nặng, về nông nghiệp, nhiều héc-ta lúa, hoa màu bị thiệt hại nặng nề, về hạ tầng đê điều, thủy lợi, giao thông nhiều ki-lô-mét đường bộ, sông, bị hư hỏng nặng. Về giáo dục, y tế nhiều trường học bị thiệt hại nặng nề. Tổng thiệt hại kinh tế khoảng 30000 tỷ đồng. Từ những nguyên nhân trên, việc dự báo nhiệt độ và lượng mưa là vấn đề cần thiết nhằm hỗ trợ người dân ứng phó biến đổi khí hậu kịp thời.

Bài viết này đề xuất kỹ thuật học sâu với mô hình MLSTM để dự báo các chỉ số về nhiệt độ và lượng mưa trung bình hàng tháng ở Việt Nam. Nghiên cứu đã đề xuất sử dụng mô hình dự báo dựa trên mạng bộ nhớ ngắn hạn dài hạn LSTM.

Các công trình trước đây cho thấy học sâu với LSTM có thể là một phương pháp

thích hợp cho dữ liệu chuỗi thời gian, Lim and Zohren (2020) đã nghiên cứu mô hình học sâu kết hợp mô hình thống kê và các thành phần mạng nơ-ron để cải thiện các phương pháp dự báo chuỗi thời gian. Ikram et al. (2019) đã trình bày một nghiên cứu sử dụng mạng nơ-ron tuần hoàn (RNN) với kiến trúc LSTM để dự đoán nhiệt độ môi trường xung quanh (TA). Nghiên cứu của Poornima et al. (2019) sử dụng mạng lưới thần kinh nhân tạo dựa trên bộ nhớ ngắn hạn tăng cường (LSTM tăng cường) để dự đoán lượng mưa cho kết quả đánh giá theo phương pháp so sánh có độ chính xác 87,99%. Một nghiên cứu khác của Xingjian et al. (2015) đã đề xuất LSTM tích hợp (ConvLSTM) và sử dụng nó để xây dựng mô hình có thể huấn luyện cho bài toán dự báo lượng mưa bằng tiếp cận máy học. Ngoài ra, còn có các nghiên cứu khác như Kratzert et al. (2018), tác giả đã dùng nhiều lưu vực của bộ dữ liệu CAMELS đề xuất mô hình dự báo lượng mưa theo phương pháp LSTM đạt độ chính xác tương tự như mô hình SAC-SMA. Zhang et al. (2017) sử dụng mạng lưới thần kinh lặp lại và bộ nhớ dài - ngắn hạn dự đoán nhiệt độ bề mặt nước biển cho kết quả mang lại độ chính xác khá cao.

2. PHƯƠNG PHÁP ĐỀ XUẤT

Trong bài báo này sẽ nghiên cứu trình bày phương pháp hướng tiếp cận dựa trên các mô hình máy học như: Mạng nơ-ron đa tầng (MLP), hồi quy vector hỗ trợ (SVR), bộ nhớ dài-ngắn hạn (LSTM). Ngoài ra, bài viết đề xuất mô hình đa biến bộ nhớ dài-ngắn hạn (MLSTM) với chuỗi thời gian có nhiều hơn một biến làm dữ liệu đầu vào cho mạng LSTM nhằm giải quyết được vấn đề của các phương pháp dự báo cũ (phụ thuộc xa của dữ liệu chuỗi thời gian trong mạng nơ-ron thông thường, nâng cao độ chính xác dự báo). Để dự báo nhiệt độ và lượng

1. <https://baochinhphu.vn/thiet-hai-30000-ty-dong-do-thien-tai-di-thuong-o-mien-trung-102283633.htm>

mưa trong tháng với các phương pháp đề xuất được đánh giá trên hai tập dữ liệu, tập dữ liệu dự báo theo tuần, theo tháng của ICRISAT_Weather.csv² từ năm 1978 đến năm 2018 và tập dữ liệu Temper_Rainfall.csv³ thể hiện dữ liệu lịch sử về nhiệt độ và lượng mưa hàng tháng tại Việt Nam từ năm 1901 đến năm 2015. Hình 1 mô hình tổng quan phương pháp đề xuất.

Tổng quan các bước từ phương pháp

đề xuất

Bước 1: Tiền xử lý dữ liệu

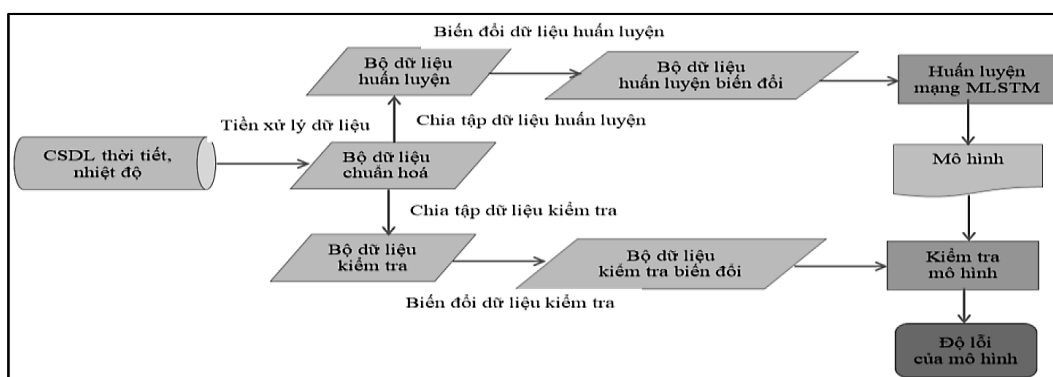
Bước 2: Chia tập dữ liệu ra (dữ liệu huấn luyện, dữ liệu kiểm tra)

Bước 3: Biến đổi dữ liệu (dữ liệu huấn luyện, dữ liệu kiểm tra).

Bước 4: Huấn luyện mô hình.

Bước 5: Kiểm tra mô hình

Bước 6: Đưa ra độ lỗi



Hình 1. Mô hình tổng quan phương pháp đề xuất

• **Dữ liệu:** Bao gồm các tập dữ liệu về nhiệt độ và lượng mưa và một số thuộc tính liên quan khác, được chia thành 2 tập: tập huấn luyện (training set) và tập kiểm tra (testing set).

• **Dữ liệu huấn luyện:** là dữ liệu để huấn luyện các mô hình máy học

• **Dữ liệu kiểm tra:** Dùng để đánh giá sự hiệu quả của các mô hình với nhau. Từ đó xác định mô hình nào hiệu quả.

• **Huấn luyện các mô hình MLSTM, LSTM, MLP, SVR.**

• **Kết quả:** Để so sánh kết quả các mô hình với nhau, bài báo sử dụng hai độ đo phổ biến là RMSE (Root Mean Square Error) và MAE (Mean Absolute Error).

2.1. Xây dựng các mô hình dự báo

2.1.1. Xây dựng mô hình dự báo LSTM

Mạng bộ nhớ dài - ngắn hạn (Long Short Term Memory networks)⁴, thường được gọi tắt là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter and Schmidhuber (1997). Điểm khác biệt khi xây dựng mô hình LSTM với MLSTM là ở dữ liệu đầu vào của mạng. Ở đây, đã sử dụng giá trị ở thời gian (t-1) trước đó để dự báo cho giá trị ở thời gian t. Dữ liệu đơn biến đầu vào mạng LSTM được mô tả trong ví dụ ở Hình 2. Trong đó $var1(t)$ là thuộc tính cần dự báo và $var1(t-1)$ là giá trị của thuộc tính $var1$ ở thời điểm t-1 trước đó dùng để làm thuộc tính đầu vào cho mô hình.

2. <http://data.icrisat.org/icrisatweather/>

3. <https://data.opendevopmentmekong.net/vi/dataset/nhi-t-d-va-lu-ng-mua-trung-binh-hang-thang-t-i-vi-t-nam-t-1901-2015?type=dataset>

4. <https://dominhhai.github.io/vi/2017/10/what-is-lstm/>

var1(t-1)	var1(t)
0.0	1
1.0	2
2.0	3
3.0	4
4.0	5
5.0	6
6.0	7
7.0	8
8.0	9

Hình 2. Dữ liệu đơn biến

Kiến trúc mạng LSTM đơn biến sử dụng dữ liệu đầu vào là luồng dữ liệu bước thời gian có trình tự, tầng LSTM có 50 nút và một lớp ẩn có 1 nút. Kết quả của dự báo sử dụng kỹ thuật Early Stopping cho 5 epoch liên tục và huấn luyện tối đa 100 epochs.

2.1.2. Xây dựng mô hình dự báo MLSTM

Mô hình đa biến đầu vào LSTM (Multivariate LSTM) sử dụng chuỗi thời gian có nhiều hơn một biến làm dữ liệu đầu vào cho mạng LSTM. Phương pháp trong dự báo chuỗi thời gian là sử dụng các quan sát có độ trễ (t-1) làm biến đầu vào để dự báo bước thời gian hiện tại (t). Với

các quan sát có nhiều hơn một thuộc tính, ta sử dụng tất cả các giá trị thuộc tính có độ trễ (t-1) để làm đầu vào. Ví dụ dưới đây sử dụng giá trị var1(t-1) và var2(t-1) để dự báo cho giá trị var1(t), với t là bước thời gian. Hình 3 mô tả ví dụ với 2 biến đầu vào cho mạng LSTM, cột cần dự báo được đóng khung màu đỏ.

var1(t-1)	var2(t-1)	var1(t)
0.0	50.0	1
1.0	51.0	2
2.0	52.0	3
3.0	53.0	4
4.0	54.0	5
5.0	55.0	6
6.0	56.0	7
7.0	57.0	8
8.0	58.0	9

Hình 3. Đa biến đầu vào cho mạng LSTM

Trong bài toán dự báo lượng mưa và nhiệt độ, dữ liệu trước khi đưa vào mô hình sẽ được chuẩn hóa từ dữ liệu đầu vào. Việc chuẩn hóa dữ liệu đầu vào được minh họa trong các Hình 4 và 5.

Bước 1: Chuyển đổi dữ liệu về dạng dữ liệu theo tuần.

	AvGT	RH1	RH2	Wind	Rain	SSH	Evap	Radiation	FA056_ET
0	22.528571	82.000000	40.857143	8.571429	0.000000	8.828571	4.571429	15.671429	3.842857
1	20.457143	90.714286	42.857143	8.900000	2.457143	6.942857	3.557143	15.500000	3.485714
2	21.271429	79.714286	25.285714	8.242857	0.000000	10.128571	5.128571	18.042857	4.528571
3	22.550000	85.857143	37.428571	9.500000	0.000000	9.542857	5.085714	17.771429	4.371429
4	23.521429	79.428571	37.714286	11.085714	0.400000	8.385714	5.985714	18.171429	4.900000
...	...	...	...	...	...	...	...	...	...
2117	27.850000	82.142857	54.714286	11.800000	0.757143	5.000000	6.985714	16.428571	4.957143
2118	27.314286	81.714286	59.714286	11.071429	5.128571	3.057143	5.485714	11.542857	3.914286
2119	24.907143	87.000000	79.714286	11.957143	5.385714	0.228571	2.542857	8.028571	2.414286
2120	25.500000	88.428571	74.571429	12.128571	3.157143	3.071429	4.157143	12.428571	3.271429
2121	26.325000	89.666667	63.166667	8.600000	5.283333	4.933333	4.200000	15.933333	4.000000

2122 rows x 9 columns

Hình 4. Dữ liệu theo tuần

Bước 2: Dữ liệu sau khi chuẩn hóa sẽ được chuyển thành dữ liệu đa biến đầu vào, sử dụng 3 bước thời gian (việc lựa chọn dựa trên phương pháp tìm kiếm siêu tham số cho từng tập dữ liệu). Nghĩa là các giá trị ở thời

gian (t-1, t-2, t-3) kết hợp với các thuộc tính khác ở bước thời gian t sẽ được sử dụng để dự báo thuộc tính đích (target/class attribute) ở thời gian t. Trong ví dụ này cột dữ liệu cần dự báo là var9(t) như trong Hình 5.

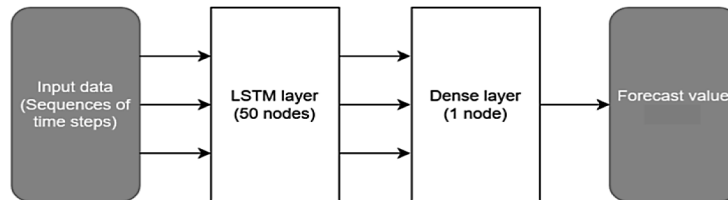
	var1(t-3)	var2(t-3)	var3(t-3)	var4(t-3)	var5(t-3)	var6(t-3)		var4(t)	var5(t)	var6(t)	var7(t)	var8(t)	var9(t)
3	22.528571	82.000000	40.857143	8.571429	0.000000	8.828571		9.500000	0.000000	9.542857	5.085714	17.771429	4.371429
4	20.457143	90.714286	42.857143	8.900000	2.457143	6.942857		11.085714	0.400000	8.385714	5.985714	18.171429	4.900000
5	21.271429	79.714286	25.285714	8.242857	0.000000	10.128571		12.128571	1.457143	7.228571	4.357143	16.314286	4.371429
6	22.550000	85.857143	37.428571	9.500000	0.000000	9.542857		10.828571	0.000000	10.542857	7.885714	20.842857	6.028571
7	23.521429	79.428571	37.714286	11.085714	0.400000	8.385714		10.171429	1.842857	7.385714	5.000000	17.400000	4.414286
...	...	...	...	...	...	...		...	...	...	...	...	...
2117	24.157143	86.428571	82.000000	12.357143	7.657143	0.157143		11.800000	0.757143	5.000000	6.985714	16.428571	4.957143
2118	25.485714	84.571429	73.714286	12.400000	2.128571	1.957143		11.071429	5.128571	3.057143	5.485714	11.542857	3.914286
2119	27.357143	82.714286	60.571429	12.228571	0.000000	3.671429		11.957143	5.385714	0.228571	2.542857	8.028571	2.414286
2120	27.850000	82.142857	54.714286	11.800000	0.757143	5.000000		12.128571	3.157143	3.071429	4.157143	12.428571	3.271429
2121	27.314286	81.714286	59.714286	11.071429	5.128571	3.057143		8.600000	5.283333	4.933333	4.200000	15.933333	4.000000

2119 rows x 36 columns

Hình 5. Dữ liệu đã được chuẩn hóa

Sau các bước chuẩn hóa dữ liệu, bài viết đã xây dựng mô hình MLSTM với sự hỗ trợ của thư viện Keras. Kiến trúc mạng MLSTM gồm có dữ liệu đầu vào (Input data) là luồng dữ liệu bước

thời gian có trình tự (Sequences of time steps), tầng LSTM (LSTM layer) có 50 nút (node) và một tầng ẩn (Dense layer) có 1 nút cho kết quả của dự báo như mô tả trong Hình 6.



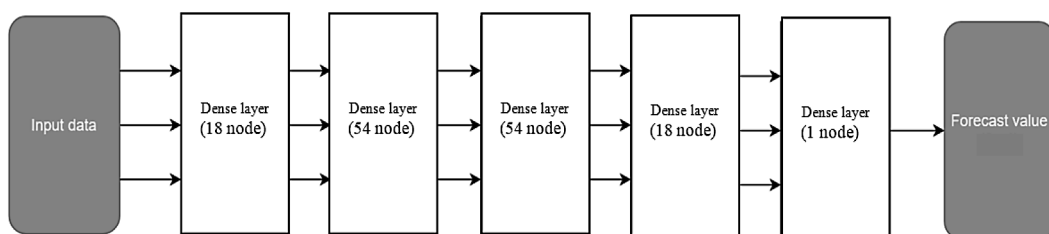
Hình 6. Kiến trúc mạng MLSTM

Để đảm bảo mô hình không học vẹt (overfitting), mô hình đã sử dụng kỹ thuật early stopping khi huấn luyện. Kỹ thuật này sử dụng một hàm gọi lại (callback) sau khi xem xét 5 kỳ huấn luyện (epochs). Trong 5 epochs liên tiếp nếu độ lỗi loss và val_loss không biến động (theo chiều hướng giảm dần) quá 0.01 thì mô hình sẽ dừng việc huấn luyện. Nếu overfitting không xảy ra, quá trình huấn luyện sẽ chạy tối đa 100 epoch. Ngoài ra, kỹ thuật này cũng giúp giảm đáng kể thời gian huấn luyện của mô hình.

2.1.3. Xây dựng mô hình dự báo MLP

MLP là một phiên bản cải tiến của Perceptron để làm việc hiệu quả hơn với dữ liệu phi tuyến tính trong thế giới thực,

nhiều nơ-ron và nhiều lớp ẩn được thêm vào. Cũng sử dụng thư viện Keras, chúng tôi xây dựng mô hình MLP kiến trúc chi tiết gồm dữ liệu đầu vào và 5 tầng ẩn. Tầng ẩn đầu tiên có 18 node sử dụng hàm kích hoạt là ReLU (rectified linear unit). Tầng ẩn thứ 2 và 3 có 54 node, sử dụng hàm kích hoạt là Sigmoid. Tầng ẩn thứ 4 có 18 node, sử dụng hàm kích hoạt là ReLU. Tầng thứ 5 là tầng output có 1 node cho giá trị đầu ra. Do không có quy định chuẩn nào về việc chọn số tầng ẩn, tôi bắt đầu với 2 tầng ẩn và chạy thử nghiệm thấy kết quả thích hợp nhất ở 5 tầng ẩn. Cũng giống như phương pháp MLSTM, tôi sử dụng kỹ thuật Early Stopping trong huấn luyện. Kiến trúc mạng MLP được mô tả trong Hình 8.



Hình 7: Kiến trúc mạng MLP

2.1.4. Xây dựng mô hình dự báo SVR

SVR là một dạng máy học véc tơ hỗ trợ nhưng dành cho việc dự đoán các giá trị liên tục. Với sự hỗ trợ của thư viện học máy Scikit-learn [5], xây dựng model SVR sử dụng hàm nhân Radial Basis Function

(RBF), do không truyền các tham số trực tiếp vào mô hình nên các tham số sẽ lấy mặc định từ thư viện, thông số (kernel) chỉ định kiểu hạt nhân sẽ được sử dụng trong thuật toán, tiếp theo là hệ số (gamma), tham số (C) điều chỉnh độ mạnh yếu của

mô hình, một tham số (**epsilon**) kiểm soát chức năng mất tập luyện để so sánh với giá trị thực, tham số (**cache_size**) chỉ định kích thước của bộ nhớ đệm hạt nhân. cùng một số tham số khác.... Hàm này phù hợp với dữ liệu đa biến và phi tuyến tính như dữ liệu đang thực nghiệm.

2.2 Phương pháp đánh giá kết quả

Để đánh giá kết quả của các phương pháp, nghiên cứu này sử dụng hai độ đo lỗi là RMSE và MAE. Đây là hai trong số các số liệu phổ biến nhất được sử dụng để đo độ chính xác cho các biến liên tục.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

Trong đó: y_j là giá trị thực mẫu thứ j ; $\hat{y}_j$ là giá trị dự đoán mẫu thứ j ; n là số mẫu dùng để đánh giá.

Để tiến hành đánh giá các mô hình dự báo nhiệt độ và lượng mưa, tôi sử dụng ngôn ngữ lập trình Python chạy trên Google Colab và sử dụng các gói thư viện mã nguồn mở Keras và Sklearn. Chương trình bao gồm các mô hình: Đa biến bộ nhớ dài - ngắn hạn (MLSTM), bộ nhớ dài - ngắn hạn (LSTM), Mạng nơ-ron đa tầng (MLP). Hồi

quy vector hỗ trợ (SVR) dự báo nhiệt độ và lượng mưa. Để đánh giá mô hình, nghiên cứu này sử dụng phương pháp tính độ lỗi phổ biến là RMSE và MAE được minh họa như hai công thức bên trên.

3. KẾT QUẢ THỬ NGHIỆM

3.1 Dữ liệu thực nghiệm

Hai tập dữ liệu đã được sử dụng trong thực nghiệm này. Tập dữ liệu thứ nhất của ICRISA là tập dữ liệu nhiệt độ và lượng mưa tại Ấn Độ (gồm 41 năm từ năm 1978 đến 2018, tập dữ liệu gốc theo ngày với 14852 dòng. Mỗi dòng có 9 thuộc tính là nhiệt độ (MaxT, MinT), lượng mưa (rainfall), độ ẩm cao, thấp (RH1, RH2), hướng gió (wind), bốc hơi (Evaporation), chỉ số tính lượng mưa (SSH), bức xạ mặt trời (radiation), lượng hơi nước bốc lên (FAO56_ET). Tập dữ liệu thứ hai Temper_Rainfall là tập dữ liệu nhiệt độ và lượng mưa trung bình hàng tháng tại Việt Nam, gồm 115 năm từ năm 1901 đến 2015, được trích từ tập dữ liệu chuẩn công khai trên trang OpenDevelopment Mê Kông bao gồm 1380 dòng có 2 thuộc tính là nhiệt độ (temperature) và lượng mưa (Rainfall). Chi tiết về hai tập dữ liệu này được mô tả trong Bảng 1.

Bảng 1. Mô tả dữ liệu

Tập dữ liệu gốc	Số mẫu tin gốc	Tập dữ liệu đã xử lý	Số mẫu tin	Mô tả
ICRISAT_Weather	14852	ICRISAT_Weather_Week	2122	Chuyển dữ liệu về chuỗi thời gian theo tuần
		ICRISAT_Weather_Month	488	Chuyển dữ liệu về chuỗi thời gian theo tháng
Temper_Rainfall	1380	Temperature_Rain_Month	1380	Dữ liệu chỉ có 2 cột nhiệt độ và lượng mưa

Tập dữ liệu ICRISAT_Weather dạng chuỗi thời gian theo ngày, có nhiều thuộc tính ảnh hưởng đến dự báo và phù hợp với mô hình MLSTM đa biến. Dữ liệu

được tiền xử lý bằng cách tính như sau: 1 tuần có 7 ngày, ta lấy trung bình mỗi 7 ngày để chuyển về dạng chuỗi thời gian theo tuần, tương tự như vậy ta lấy tất cả

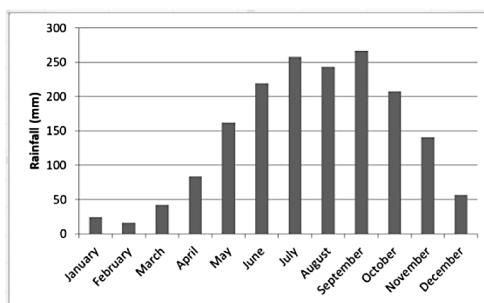
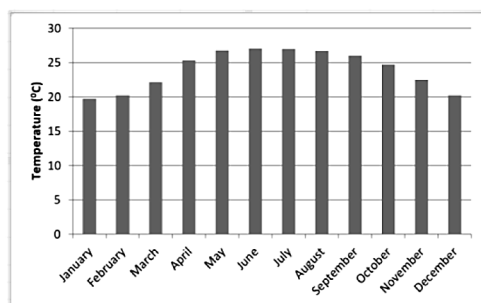
dữ liệu của một tháng nào đó trong năm 2018 chia trung bình ra dữ liệu của 1 tháng) dữ liệu sau khi xử lý mô tả trong Bảng 2.

Bước 2. Dữ liệu sau khi tiền xử lý

Tập dữ liệu	Số mẫu tin	Số cột	Mô tả	Mục đích
ICRISAT_Weather_Week	2122	9	1. Nhiệt độ tối đa, tối thiểu (MaxT, MinT)	- Các thuộc tính ngữ cảnh liên quan dự báo, giúp tăng độ chính xác - Đánh giá sự ảnh hưởng của các thuộc tính môi trường xung quanh đến kết quả dự báo.
ICRISAT_Weather_Month	488		2. Lượng mưa (rainfall) 3. Độ ẩm cao (RH1) 4. Độ ẩm thấp (RH2) 5. Hướng gió (wind) 6. Bốc hơi (Evaporation) 7. Chỉ Số tính lượng mưa (SSH) 8. Bức xạ mặt trời (radiation) 9. Lượng hơi nước bốc lên (FAO56_ET)	
Temperature_Rain_Month	1380	2	1. Nhiệt độ (temperature) 2. Lượng mưa (rainfall)	

Riêng tập dữ liệu Temperature_Rain_Month nhiệt độ và lượng mưa trung bình hàng tháng từ tháng 1 năm 1901 đến tháng 12 năm 2015. Ví dụ: Tổng dữ liệu nhiệt độ và lượng mưa trung bình hàng tháng

cho năm 1901 lần lượt là 2883872°C và 17237291mm, đây là biểu đồ thể hiện nhiệt độ và lượng mưa trung bình hàng tháng năm 1901 tại Việt Nam được thể hiện như Hình 8.



Hình 8. Biểu đồ thể hiện nhiệt độ và lượng mưa trung bình hàng tháng

3.2. Kết quả thực nghiệm

Bảng 3 so sánh kết quả dự báo qua các mô hình MLP, SVR, LSTM và MLSTM, kết quả tốt nhất cho từng tập dữ liệu được in đậm. Dữ liệu được lấy ngẫu nhiên 80% các mẫu tin đầu cho huấn luyện và 20% mẫu tin cuối cho kiểm tra (theo trình tự

thời gian). Mục đích của việc chia dữ liệu trên nhằm bám sát theo thực tế, dựa trên các thuộc tính đã học để dự báo cho các thuộc tính tiếp theo. Vì tính ngẫu nhiên của giải thuật, phương pháp MLP không cần trình tự thời gian nên được huấn luyện một lần trên mỗi tập dữ liệu.

Bảng 3. Kết quả dự báo của các mô hình

Tập dữ liệu	Tổng số dòng/cột		RMSE				MAE			
	Dòng	Cột	MLP	SVR	LSTM	MLSTM	MLP	SVR	LSTM	MLSTM
Nhiệt độ										
ICRISAT_weather_week	2122	9	1.566	1.544	1.506	1.335	1.215	1.143	1.143	1.012
ICRISAT_weather_month	488	9	3.386	1.907	1.731	1.311	2.645	1.493	1.446	1.051
Temperature_rain_month	1380	2	1.696	1.634	1.634	1.133	1.207	1.235	1.275	0.863
Lượng mưa										
ICRISAT_weather_week	2122	9	4.371	4.546	4.418	4.190	2.002	2.080	2.688	2.493
ICRISAT_weather_month	488	9	2.475	2.496	2.504	2.299	1.359	1.417	1.786	1.450
Temperature_rain_month	1380	2	67.023	70.427	76.922	54.291	47.011	55.324	59.296	40.495

Từ tập dữ liệu ICRISAT_weather_week gồm chín thuộc tính liên quan đến nhiệt độ và lượng mưa cho thấy mô hình MLSTM đạt hiệu quả với độ lỗi RMSE là **1.335** và MAE là **1.012**, tương ứng trên tập dữ liệu lượng mưa là **4.190** và **2.493**. Trên tập dữ liệu ICRISAT_weather_month cho thấy mô hình MLSTM đạt hiệu quả khá tốt với độ lỗi RMSE trên tập nhiệt độ là **1.311** và MAE là **1.051**, tương ứng trên tập dữ liệu lượng mưa là **2.299** và **1.450**. Với kết quả thu được từ tập dữ liệu dự báo bước thời gian theo tuần, theo tháng từ năm (1978-2018) thì mô hình dự báo MLSTM luôn cho kết quả tốt hơn. Tuy nhiên độ lỗi MAE của lượng mưa lại chưa tốt hơn mô hình MLP. Từ đó cho thấy lượng mưa phân bố không đồng đều, biến động liên tục trong năm cũng ảnh hưởng đến kết quả dự báo. Bên cạnh đó khi sử dụng mô hình đa biến với các thuộc tính ngữ cảnh cũng có

sự ảnh hưởng đến kết quả dự báo. Vì vậy việc xác định các thuộc tính, ngữ cảnh, các dữ liệu liên quan đưa vào huấn luyện rất quan trọng trong bài toán dự báo thời gian với mô hình MLSTM.

Từ tập dữ liệu TEMPER_Rain_Month gồm hai thuộc tính. Nhiệt độ trung bình của tháng và lượng mưa, năm 1901 lần lượt là 2883872°C và 17237291mm, được trình bày trong Hình 8, cho thấy nhiệt độ cao hơn và lượng mưa phổ biến trong thời gian giữa năm. Vào tháng Giêng và tháng Hai, lượng mưa được quan sát thấy rất thấp. Mặt khác, nhiệt độ trung bình tháng dao động quanh năm ít hơn 38 °C. Chúng tôi đã sử dụng dữ liệu thời tiết trong 115 năm và lấy ngẫu nhiên 80% các mẫu tin đầu cho huấn luyện và 20% mẫu tin cuối cho kiểm tra (theo trình tự thời gian). Do độ chênh lệch giữa nhiệt độ và lượng mưa quá lớn nên dẫn đến độ lỗi dự báo của lượng mưa qua mô

hình MLSTM khá cao RMSE là **54.2909** và MAE là **40.4946** so với nhiệt độ chỉ có RMSE là **1.133** và MAE là **0.863**. Với kết quả thu được từ tập dữ liệu dự báo bước thời gian theo tháng từ năm (1901-2015) thì mô hình dự báo MLSTM vẫn chiếm độ lỗi thấp nhất so với các mô hình LSTM, SVR, MLP. Từ đó cho thấy mô hình đa biến đầu vào mạng LSTM (Multivariate LSTM-MLSTM) sử dụng chuỗi thời gian có nhiều hơn một biến làm dữ liệu đầu vào cho kết quả dự báo tốt hơn. Tuy nhiên qua thử nghiệm ta thấy rằng độ lệch chuẩn của dữ liệu quá lớn và thuộc tính dự báo ít cũng ảnh hưởng đến độ lỗi dự báo.

Kết quả so sánh độ lỗi tập dữ liệu ICRISAT_Weather_Month cho thấy, mô

hình LSTM và MLSTM đạt kết quả tốt hơn các mô hình khác trên cùng từng tập dữ liệu. Có thể thấy rằng mạng LSTM hoạt động khá tốt trên luồng dữ liệu bước thời gian có trình tự. Để kiểm tra hiệu suất dự báo, lập đồ thị với trục trung là loss, trục hoành epoch, xem xét sự tổn thất do huấn luyện và số kỷ nguyên trong mỗi mô hình để đánh giá hiệu suất của mô hình trong dự báo. Tổn thất do huấn luyện đối với mô hình dự báo nhiệt độ và lượng mưa được trình bày trong Hình 4.1 và 4.2. Tổn thất đối với dự báo nhiệt độ hội tụ sau 10 kỷ nguyên. Tương ứng với lượng mưa hội tụ sau 25 kỷ nguyên.

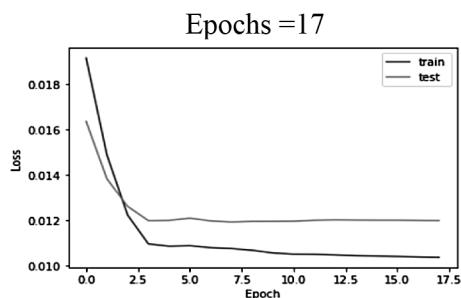
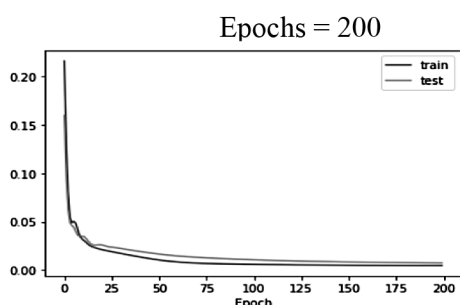
Đây là các tham số được sử dụng cho mô hình trong kết quả thực nghiệm,

Tham số	Giá trị
Epochs dự kiến	200
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	200
loss	0.0046
val_loss	0.0070
RMSE	1.311
MAE	1.051

Tham số	Giá trị
Epochs dự kiến	200
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	17
loss	0.0104
val_loss	0.0123
RMSE	2.353
MAE	1.511

Bảng 4: Tham số huấn luyện nhiệt độ và lượng mưa

Với Epochs dự báo nhiệt độ là 200, Epochs dự báo lượng mưa là 17 thì đồ thị được hiển thị như sau?



Hình 9: Train loss với epoch dự báo nhiệt độ và lượng mưa

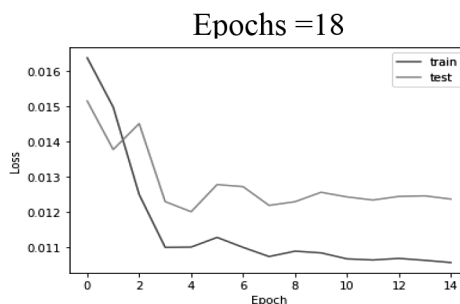
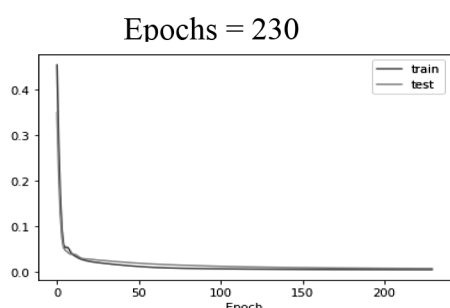
Đây là các tham số được sử dụng cho mô hình trong kết quả thực nghiệm,

Tham số	Giá trị
Epochs dự kiến	230
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	230
loss	0.0047
val_loss	0.0073
RMSE	1.306
MAE	1.038

Tham số	Giá trị
Epochs dự kiến	230
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	18
loss	0.0103
val_loss	0.0123
RMSE	2.355
MAE	1.505

Bảng 5: Tham số huấn luyện nhiệt độ và lượng mưa

Với Epochs dự báo nhiệt độ là 230, Epochs dự báo lượng mưa là 18 thì đồ thị được hiển thị như sau?



Hình 10: Train loss với epoch dự báo nhiệt độ và lượng mưa

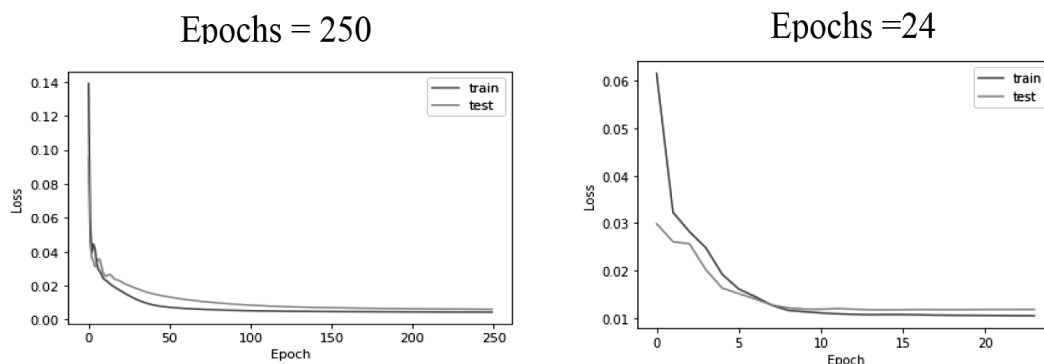
Đây là các tham số được sử dụng cho mô hình trong kết quả thực nghiệm,

Tham số	Giá trị
Epochs dự kiến	250
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	250
loss	0.0044
val_loss	0.0061
RMSE	1.189
MAE	0.943

Tham số	Giá trị
Epochs dự kiến	250
min_delta	0.00006
patience	10
Bactch_size	80
Epochs thực	24
loss	0.0105
val_loss	0.0119
RMSE	2.327
MAE	1.487

Bảng 6: Tham số huấn luyện nhiệt độ và lượng mưa

Với Epochs dự báo nhiệt độ là 230, Epochs dự báo lượng mưa là 18 thì đồ thị được hiển thị như sau?

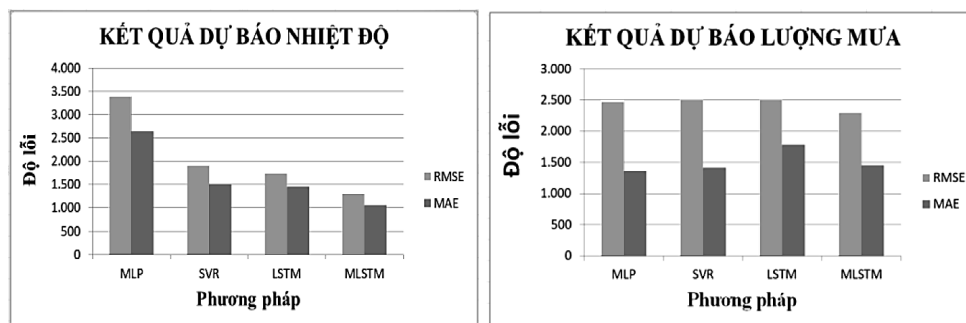


Hình 11: Train loss với epoch dự báo nhiệt độ và lượng mưa

Để so sánh kết quả dự báo trên các tập dữ liệu từ các mô hình MLP, SVR, LSTM và MLSTM, kết quả tốt nhất cho từng tập dữ liệu được in đậm. Dữ liệu được lấy ngẫu nhiên 80% các mẫu tin đầu cho huấn luyện và 20% mẫu tin cuối cho kiểm tra (theo trình tự thời gian). Mục đích của việc chia dữ liệu trên nhằm bám sát theo thực tế, dựa trên các yếu tố đã học để dự báo cho các yếu tố tiếp theo. Vì tính ngẫu nhiên của giải thuật, phương pháp MLP

không cần trình tự thời gian nên được train/test 10 lần và lấy trung bình độ lỗi, các phương pháp còn lại huấn luyện một lần trên mỗi tập dữ liệu.

Kết quả dự báo nhiệt độ và lượng mưa theo tháng với tập dữ liệu ICRISAT_Weather_Month của mô hình MLSTM được trình bày trong Hình 10. Quan sát biểu đồ này, ta có thể thấy rằng mô hình MLSTM dự báo khá tốt theo tiêu chí RMSE so với các mô hình khác



Hình 12. So sánh kết quả dự báo nhiệt độ và lượng mưa

PHẦN KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1.1. KẾT LUẬN

Bài viết này đã đề xuất phương pháp dự báo nhiệt độ và lượng mưa bằng kỹ thuật học sâu nhằm hỗ trợ người dân có kế hoạch gieo trồng phù hợp, góp phần thúc đẩy ngành nông nghiệp tại Việt Nam nói chung cũng như khu vực ĐBSCL nói riêng. Dữ liệu sau khi thu thập, xử lý, tiến hành huấn luyện và kiểm tra với mô hình

đa biến bộ nhớ dài - ngắn hạn MLSTM. Kết quả thực nghiệm được so sánh với các mô hình dự báo khác như (LSTM, MLP, SVR) và cho thấy phương pháp tiếp cận được đề xuất có thể tạo ra các dự báo khá chính xác, có thể áp dụng vào hệ thống thực tế.

1.2. HẠN CHẾ

Tập dữ liệu thu thập chưa đủ lớn để huấn luyện dẫn đến độ chính xác của mô hình dự báo chưa cao.

1.3. HƯỚNG PHÁT TRIỂN

Nghiên cứu này sẽ tiếp tục được cải tiến nhằm nâng cao độ chính xác của mô hình dự báo cũng như phát triển mô hình để có thể dự báo cho nhiều hơn một bước thời

gian đầu vào, đầu ra và hình thành công cụ để thuận tiện cho người dùng cuối sử dụng. Bên cạnh đó, việc so sánh với các nghiên cứu liên quan cũng sẽ được thực hiện trong tương lai.

TÀI LIỆU THAM KHẢO

- [1]. Lim, B., & Zohren, S., (2020). Time Series Forecasting With Deep Learning: *A Survey. arXiv preprint arXiv:2004.13408*.
- [2]. Ikram, B. A. O., Abdelhakim, B. A., Abdelali, A., Zafar, B., & Mohammed, B. (2019, March). Deep Learning architecture for temperature forecasting in an IoT LoRa based system. In *Proceedings of the 2nd International Conference on Networking, Information Systems & Security* (pp. 1-6).
- [3]. Poornima, S., & Pushpalatha, M. (2019). Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. *Atmosphere*, 10(11), 668.
- [4]. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.
- [5]. LIAO, W., YIN, Z., WANG, R., & LEI, X. (2019). RAINFALL-RUNOFF MODELLING BASED ON LONG SHORT-TERM MEMORY (LSTM).
- [6]. Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Computers & Electrical Engineering*, 70, 66-73.
- [7]. Nguyen Thai-Nghe, Nguyen Thanh-Hai and Nguyen Chi Ngon, “Deep Learning Approach for Forecasting Water Quality in IoT Systems” *International Journal of Advanced Computer Science and Applications(IJACSA)*, 11(8), 2020. <http://dx.doi.org/10.14569/IJACSA.2020.0110883>.
- [8]. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005-6022. Zhang,
- [9]. Zhang, Q., Wang, H., Dong, J., Zhong, G., & Sun, X. (2017). Prediction of sea surface temperature using long short-term memory. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1745-1749.