ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

XÂY DỰNG MÔ HÌNH ĐIỀN KHUYẾT DỮ LIỆU CHUỖI THỜI GIAN SỬ DỤNG MẠNG ĐỐI NGHỊCH TẠO SINH (GAN)

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Trần Hoàng Anh	20521079
2	Nguyễn Huỳnh Vương Quốc	20521813
3	Ngô Huỳnh Trưởng	20522085
4	Nguyễn Hữu Minh Tâm	20521871

TP. HỒ CHÍ MINH – 12

1. GIỚI THIỆU

"80 percent of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis."

Theo IBM Data Analytics, chúng ta có thể dành tới 80% thời gian để tiền xử lí dữ liệu. Để có thể làm một tác vụ học máy hay học sâu cụ thể thì chúng phải phải cần có một bộ dữ liệu đủ tốt, nhưng mọi chuyện vốn không đơn giản như vậy. Việc dữ liệu không được tiền xử lí đủ tốt có thể dẫn đến sự sụp đổ của cả mô hình. Và một trong những vấn đề lớn nhất trong việc làm sạch dữ liệu đó chính là việc dữ liệu bị khuyết. Với đồ án này, nhóm chúng em đề xuất xây dựng và thử nghiệm một mô hình điền khuyết cho bộ dữ liệu timeseries Weather Forecast, được lấy trong khoảng thời gian 5 năm gần đây. Đề tài của nhóm tập trung vào vấn đề điền khuyết cho dữ liệu bị khuyết hoàn toàn ngẫu nhiên.

Weather Forecast là bộ dữ liệu được cung cấp bởi NASA Power. Bộ dữ liệu này gồm các chỉ số về thời tiết như nhiệt độ, lượng mưa, tốc độ gió… theo thời gian (hay còn gọi là Time Series). Với đề tài này, chúng em sẽ tiến hành xây dựng mô hình học sâu sử dụng kiến trúc mạng đối nghịch tạo sinh (GAN) áp dụng lên dữ liệu chuỗi thời gian ở Việt Nam. Sau đó so sánh và đánh giá kết quả phương pháp này với các phương pháp cổ điển khác như điền khuyết bằng mean, median, random hoặc phương pháp sử dụng học máy như KNN. Kết quả cuối cùng cho thấy việc áp dụng mô hình GAN có tiềm năng và triển vọng cao, tuy nhiên hiện tại vẫn chưa đạt được kết quả quá tốt.

Với ý tưởng ban đầu của mạng đối nghịch tạo sinh là dành cho tác vụ tạo sinh hình ảnh giả sao cho giống thật nhất. Về kiến trúc, GAN sử dụng 2 mạng con bên trong có nhiệm vụ đối kháng nhau. Generator có vai trò sinh dữ liệu sao cho giống thật nhất, trong khi mạng phân biệt cố gắng phân biệt giữ dữ liệu thật hay dữ liệu được sinh ra bởi Generator. Bằng cách triển khai Generator sử dụng GRU (Một biến thể của RNN) ở cả Generator và mạng phân biệt mô hình có thể điền khuyết ở một số bộ dữ liệu cụ thể.

Với những bộ dữ liệu đã được điền khuyết, thử nghiệm xây dựng các mô hình cho tác vụ dự đoán nhiệt độ dựa vào thông số của những ngày trước. Các mô hình học máy (Linear Regression, Random Forest, Support Vector Machine) và học sâu (Artificial Neural Network) được sử dụng để dự đoán và đánh giá các bộ dữ liệu được điền khuyết.

2. NỘI DUNG

2.1. Bộ dữ liệu Weather Forecast

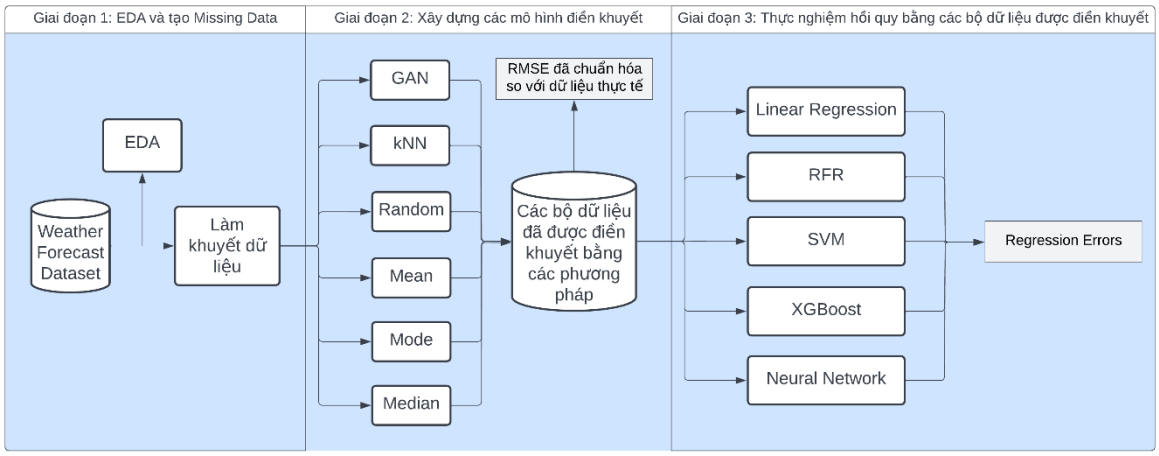
Bộ dữ liệu Weather Forecast được cung cấp bởi tổ chức được thu thập bằng phương pháp thủ công từ website chính thức của NASA. Bộ dữ liệu chứa các thông tin chi tiết về thời tiết và khí hậu tại khu vực Thành phố Thủ Đức trong 5 năm, từ ngày 1 tháng 1 năm 2017 đến ngày 31 tháng 12 năm 2021. Bộ dữ liệu đầy đủ, không có missing values và missing timestamp, bao gồm 1826 dòng và 10 thuộc tính như sau (Bảng 1):

Thuộc tính	Ý nghĩa của thuộc tính	Loại thuộc tính	Khoảng giá trị
YEAR	Năm	Định tính	2017 – 2021
MO	Tháng	Định tính	1 – 12
DY	Ngày	Định tính	1 – 31
Temperature	Nhiệt độ	Định lượng	19.9 – 33.7
Relative_Humidity	Độ ẩm tương đối	Định lượng	43 – 95.3
Specific_Humidity	Độ ẩm cụ thể	Định lượng	9.28 – 20.94
Precipitation	Lương mưa	Định lượng	0 – 143.1
Pressure	Áp suất	Định lượng	99.87 – 101.6
Wind_Speed	Tốc độ gió	Định lượng	0.68 – 7.05
Win_Direction	Hướng gió	Định lượng	22.31 – 343.56

Bảng 1. Các thuộc tính và mô tả của các thuộc tính.

Ngoài các đặc điểm thời tiết cơ bản như độ ẩm, lượng mưa, áp suất, tốc độ gió… chúng em dự đoán rằng hướng gió cũng sẽ có tác động đáng kể đến việc dự báo thời tiết (cụ thể ở đây biến target là nhiệt độ). Với các hướng gió khác nhau, sẽ có những tác động khác nhau, ví dụ như gió Tây Nam sẽ gây mưa lớn, gió tây nóng và khô khiến nhiệt độ tăng cao, và gió mùa Đông Bắc mang theo không khí lạnh và khô, điều này sẽ làm cho độ ẩm giảm, ảnh hưởng đến lượng mưa... Trong dữ liệu sẽ có những điểm outlier, đó là do các kiểu thời tiết bất ngờ gây nên, có thể là ảnh hưởng của bão hoặc áp thấp tại các khu vực lân cận.

Phương pháp phân tích dữ liệu và xây dựng mô hình:



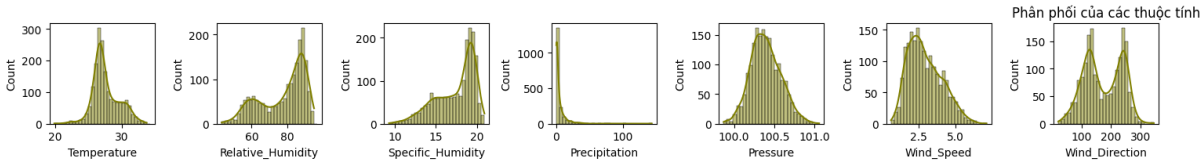
Hình 1. Quy trình phân tích dữ liệu và xây dựng mô hình.

Mô tả quy trình:

- Bước 1: Phân tích thăm dò dữ liệu (EDA) để hiểu thêm về dữ liệu (các thuộc tính, phân phối của dữ liệu…), đồng thời tạo missing values bằng cách "đục lỗ" ngẫu nhiên.
- Bước 2: Xây dựng mô hình điền khuyết sử dụng GAN đồng thời triển khai các mô hình máy học khác như kNN và phương pháp thống kê như Mean, Mode, Median và điền khuyết ngẫu nhiên.
- Bước 3: Các bộ dữ liệu điền khuyết sẽ được làm đầu vào cho các mô hình hồi quy như Linear Regression, RFR, SVM, NN để đánh giá mức độ hồi quy của dữ liệu bằng các chỉ số như RMSE, MAE,…
- Bước 4: Rút ra kết luận.

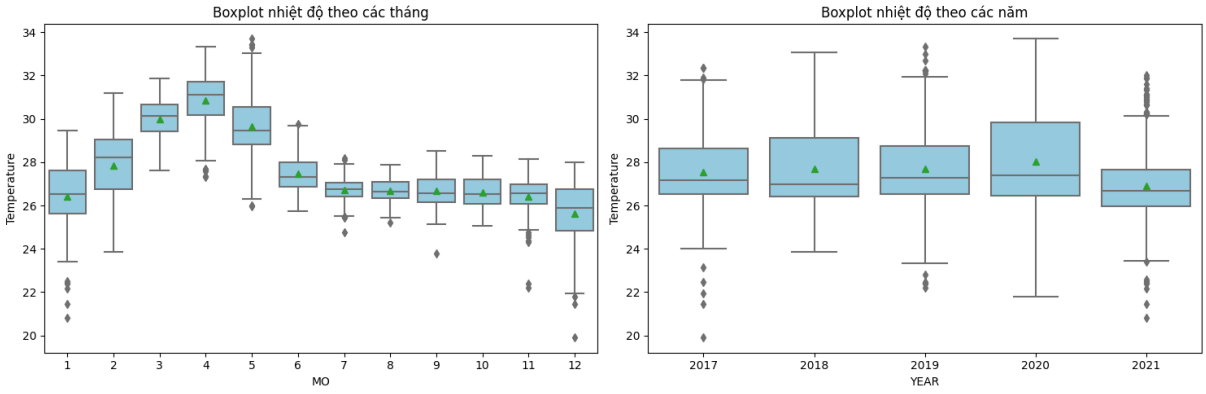
2.2. Phân tích thăm dò dữ liệu

Đầu tiên, về mặt phân phối dữ liệu của các thuộc tính theo hình 2 ta có thể thấy biến phụ thuộc nhiệt độ có phân phối khá giống chuẩn, bên cạnh đó là biến áp suất và tốc độ gió cũng tương tự.



Hình 2. Phân phối dữ liệu của các thuộc tính.

Ngoài các đặc điểm thời tiết cơ bản như độ ẩm, lượng mưa, áp suất, tốc độ gió… chúng em dự đoán rằng hướng gió cũng sẽ có tác động đáng kể đến việc dự báo thời tiết (cụ thể ở đây biến phụ thuộc là nhiệt độ). Với các hướng gió khác nhau, sẽ có những tác động khác nhau, ví dụ như gió Tây Nam sẽ gây mưa lớn, gió tây nóng và khô khiến nhiệt độ tăng cao, và gió mùa Đông Bắc mang theo không khí lạnh và khô, điều này sẽ làm cho độ ẩm giảm, ảnh hưởng đến lượng mưa... Trong dữ liệu sẽ có những điểm outlier, đó là do các kiểu thời tiết bất ngờ gây nên, có thể là ảnh hưởng của bão hoặc áp thấp tại các khu vực lân cận.



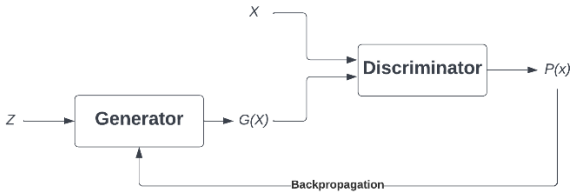
Hình 3. Boxplot của nhiệt độ theo các tháng và các năm.

Bên cạnh đó hình 3 cũng cho thấy rằng nhiệt độ giữa các tháng trong năm khá cao vào đầu năm và thấp vào những tháng cuối năm, tuy nhiên nhiệt độ gần như không có khác biệt giữa khác biệt với p-value tính được là 0.0003 < 0.05, có ý nghĩa thống kê. Tương tự p-value cho rằng nhiệt độ giữa các năm không có sự khác biệt là 2.2e^-13 < 0.05, có ý nghĩa thống kê.

2.3. Mô hình GAN điền khuyết cho dữ liệu chuỗi thời gian Weather Forecast

2.3.1. Giới thiệu ý tưởng GAN ban đầu sử dụng trong việc sinh ảnh

GAN là viết tắt "Generative Adversarial Network", hướng tới việc sinh ra dữ liệu mới sau quá trình học (Hình 4). GAN có thể tự sinh ra một khuôn mặt mới, một con người, một đoạn văn, chữ viết, bản nhạc giao hưởng hay những thứ tương tự thế. GAN được kết hợp từ 2 model: generator – G và discriminator – D.



Hình 4. Mô tả mô hình GAN.

GAN giống như 1 trò chơi minimax, trò cảnh sát tội phạm: tội phạm G tạo ra tiền giả, cảnh sát D học cách phân biệt thật giả. Cảnh sát càng cố gắng phân biệt tiền thật-giả thì tội phạm lại dựa vào feedback của cảnh sát để cải thiện khả năng tạo tiền giả của mình, cố gắng khiến cảnh sát phân biệt nhầm.

Mục đích cuối cùng của GAN là giúp G học cách tạo ra tiền giả, khiến cảnh sát D phân vân không thể đưa ra được quyết định chính xác.

➔ Từ ý tưởng trên, nhóm quyết định sử dụng mô hình GAN cho việc điền khuyết cho bộ dữ liệu chuỗi thời gian.

2.3.2. Generator (G)

Cốt lõi của Generator (G) trong đề tài của nhóm là mô hình GRU. Đảm nhiệm việc sinh ra dữ liệu được điền khuyết.

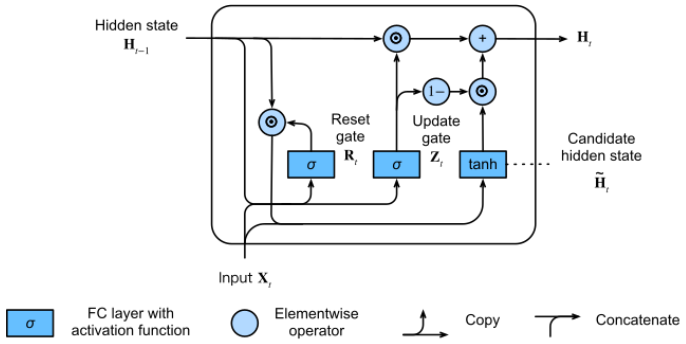
Lý do tại sao lại chọn GRU:

- Thứ nhất, GRU là mô hình học sâu chuyên sử dụng cho các tác vụ liên quan đến việc xử lí dữ liệu dạng chuỗi (sequence), đặc biệt là chuỗi thời gian (timeseries).
- Thứ hai, so với các mô hình biến thể khác của RNN, thì GRU có khả năng tính toán nhanh hơn, cũng như là hiệu suất cao hơn so với mô hình RNN gốc.

Đầu vào và đầu ra của GRU trong thiết kế (Hình 5):

- Đầu vào: Là một chuỗi dữ liệu chứa N phần tử (N là một siêu tham số), mỗi phần tử chứa M thuộc tính. Nếu ánh xạ sang bộ dữ liệu của nhóm thì sẽ là một chuỗi N ngày liên tục nhau, với 6 thuộc tính của ngày hôm đó (Ở đây 6 thuộc tính là các thuộc tính đã được phân tích ở phần phần tích thăm dò trừ đi thuộc tính target là nhiệt độ).
- Đầu ra: Vì là Generator (G) cho việc điền khuyết dữ liệu, nên kích thước của đầu ra sẽ tương tự đầu vào. Tuy nhiên khi qua GRU thì dữ liệu sẽ bị thay đổi giá trị, ta không thể biết được rằng mô hình đang làm việc gì ở bên trong. Vì đa phần các mô hình học sâu là một blackbox

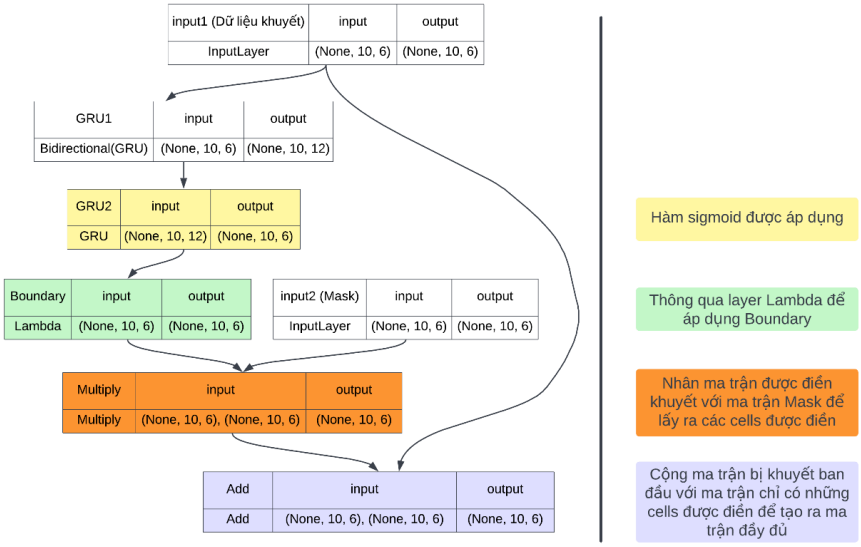
➔ Cho nên chúng ta cần một thứ gì đó để có thể kiểm soát được đầu ra.



Hình 5. Cấu trúc 1 cell GRU.

Để kiểm soát được đầu ra, ở lớp cuối của GRU ta sẽ áp dụng hàm sigmoid để đưa toàn bộ giá trị được sinh ra nằm trong khoảng từ 0 đến 1. Sau đó, ta ánh xạ các giá trị đó về một miền giá trị được định nghĩa là boundary.

- Miền giá trị boundary giúp GRU sinh ra dữ liệu nằm trong miền giá trị đó, nhằm kiểm soát được giá trị được sinh ra.
- Boundary có thể là từ phân vị này đến phân vị khác (25% đến 75%), hoặc là từ giá trị min đến max của phân phối dữ liệu (0% đến 100%)
- Với 6 thuộc tính được xét trong bộ dữ liệu, thì boundary sẽ là một ma trận 2 x 6.
- Vì bộ dữ liệu không có giá trị âm nên phương pháp này có thể áp dụng được.
- Đây là một bộ siêu tham số

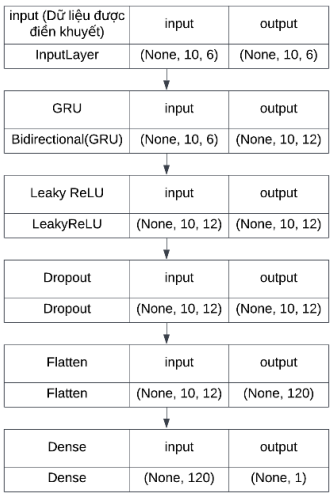


Hình 6. Kiến trúc Generator (G) được sử dụng.

2.3.3. Discriminator (D)

Nhiệm vụ của Discriminator (D) là phân biệt rằng liệu dữ liệu được cho là thật hay là giả (0 hoặc 1). Tuy nhiên, ở Discriminator thì là một mô hình phân loại, nên kiến trúc của nó sẽ đỡ phức tạp hơn của Generator rất nhiều (Hình 6):

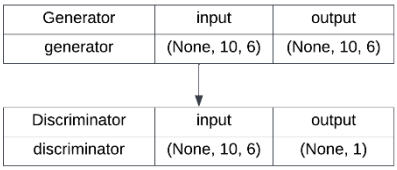
- Áp dụng GRU cho đầu vào, sau đó sử dụng kỹ thuật DropOut để tránh việc overfitting.
- Sau đó kéo dãn ra thành vector, áp dụng hàm sigmoid để phân loại.



Hình 7. Kiến trúc Discriminator (D) được sử dụng.

2.3.4. Mô hình GAN

Mô hình GAN đơn giản là kết nối Generator (G) với Discriminator (D) lại với nhau, Generator sinh dữ liệu, Discriminator phân loại. Tuy nhiên, Discriminator sẽ bị đóng băng bộ tham số. Việc đóng băng bộ tham số của Discriminator để có thể huấn luyện được Generator học được mẫu (pattern) của dữ liệu thật tế.



Hình 8. Kiến trúc GAN được sử dụng.

2.3.5. Huấn luyện mô hình

Đầu tiên, ta lấy ra bộ dữ liệu thật, nhãn 1; dữ liệu giả, nhãn 0.

- Việc sinh ra dữ liệu giả, ta sử dụng Generator để sinh ra dữ liệu. Đầu vào cho Generator sẽ là bộ dữ liệu đã được làm khuyết đi X%.

Huấn luyện Discriminator (D):

- Ta tiến hành gỡ băng cho bộ tham số của Discriminator để có thể huấn luyện.
- Sử dụng bộ dữ liệu vừa được lấy ra, để huấn luyện cho Discriminator.
- Ý nghĩa của việc là để cho Discriminator phân biệt được dữ liệu thật và giả.

Huấn luyện GAN (mục tiêu đánh lừa Discriminator):

- Đầu tiên ta đóng băng Discriminator lại, dùng cho việc phân loại thật giả dựa vào dữ liệu Generator sinh ra.
- Tạo ra nhãn 1, ý nghĩa là nhãn cho dữ liệu thật. Tuy nhiên, ta sẽ sử dụng dữ liệu giả do Generator sinh ra.
- Tiến hành đánh lừa Discriminator bằng cách cung cấp cho Generator dữ liệu bị khuyết, để sinh ra dữ liệu giả.
- Sau đó cho Discriminator phân loại và tinh chỉnh tham số của Generator dựa vào nhãn 1.

- Điều này giúp cho Generator học được mẫu (pattern) của dữ liệu thật nó sẽ như thế nào. Từ đó cải thiện việc tạo ra dữ liệu giả ngày càng giống thật.
- Hay nói cách khác là dữ liệu do Generator điền khuyết sẽ ngày càng giống dữ liệu thật.

2.3.6. Phương pháp đánh giá mô hình

Việc đánh giá mô hình Generator nếu chỉ sử dụng thang đo RMSE giữa những cell được điền và so với thực tế thì không mấy có ý nghĩa. Vì mỗi thuộc tính có một miền giá trị khác nhau, nên việc tính RMSE mà không có bước xử lí sẽ mang lại kết quả không chính xác để đánh giá.

Hướng giải quyết:

- Bước 1: Chuẩn hóa toàn bộ cell được điền khuyết và thực tế vào miền giá trị từ 0 – 1.
- Bước 2: Sau đó tính RMSE (Căn bậc 2 của trung bình tổng các giá trị thực trừ đi giá trị dự đoán bình phương).

2.3.7. Kết quả thực nghiệm

Được huấn luyện trên bộ dữ liệu được làm khuyết nhân tạo với tỉ lệ là 80%. Sau nhiều lần thử nhiệm thì khi làm khuyết dữ liệu khoảng 65 đến 80% so với bộ dữ liệu đầy đủ thì GAN cho ra được kết quả tốt nhất. Trong quá trình thực nghiệm và huấn luyện mô hình, chúng em đã thử và tinh chỉnh các siêu tham số và có gặp những trường hợp:

- Huấn luyện càng nhiều epochs thì mô hình cũng không cải thiện được quá nhiều, thậm chí có có phần tăng lỗi.
- Dữ liệu sinh ra giống nhau. Điều này có thể giải thích khi mà Generator tìm ra một điểm dữ liệu đặc biệt mà tại điểm đó Discriminator không thể phân biệt được.

Sau cùng, sau nhiều lần thử thì chúng em tìm ra được bộ siêu tham số tốt nhất thu được kết quả là RMSE là 0.256 đối với mô hình GAN của chúng em (Bảng 2).

Phương pháp	RMSE (đã chuẩn hóa)
GAN	0.256
kNN	0.120

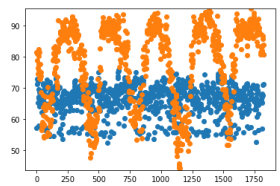
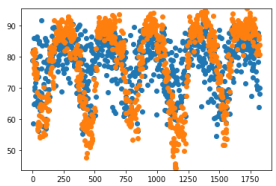
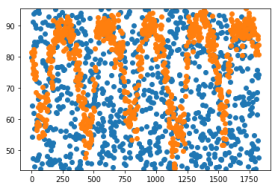
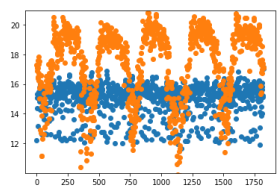
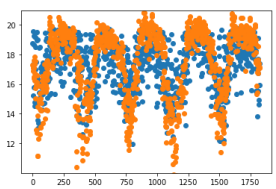
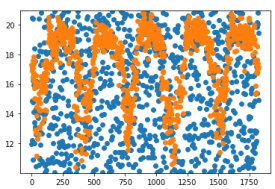
Mean	0.141
Median	0.135
Mode	0.139
Random	0.341

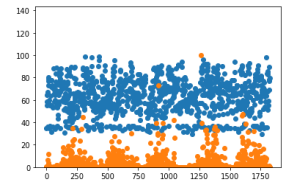
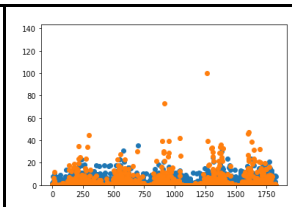
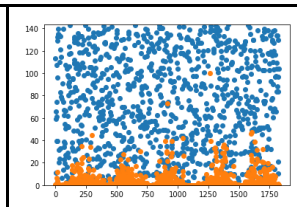
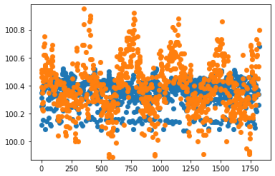
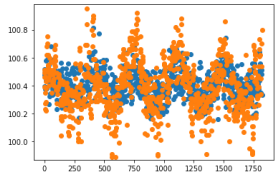
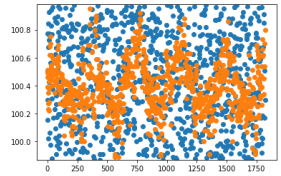
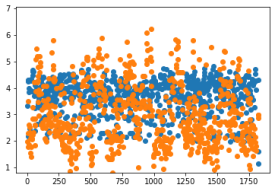
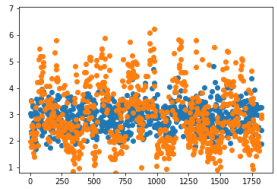
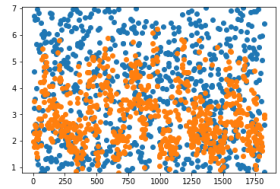
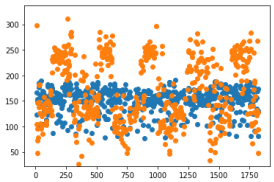
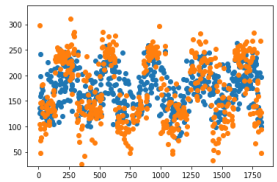
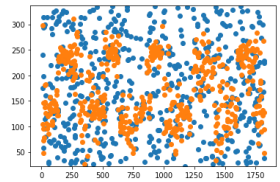
Bảng 2. RMSE đã chuẩn hóa của các phương pháp điền khuyết.

So sánh với các phương pháp điền khuyết truyền thống khác thì mô hình của nhóm có phần không tốt nếu chỉ dựa vào thang đó RMSE. Một số lý do có thể kể ra như:

- Dữ liệu không đủ lớn.
- Kiến trúc mô hình không phù hợp hoặc quá đơn giản.
- Lựa chọn bộ siêu tham số không phù hợp.

Bảng 3 bên dưới trực quan kết quả của các giá trị điền khuyết bởi 3 phương pháp: GAN, kNN và điền ngẫu nhiên trong khoảng giá trị. Cột x của các biểu đồ thể hiện vị trí của điểm dữ liệu được điền khuyết, cột y thể hiện giá trị được điền khuyết (màu xanh) và giá trị thực tế (màu cam). Kết quả cho thấy rằng GAN không phải là điền một cách hoàn toàn ngẫu nhiên so với phương pháp điền khuyết ngẫu nhiên trong khoảng giá trị của nó. Tuy nhiên thì dữ liệu điền khuyết tập trung phần lớn ở giữa khoảng boundary. Có thể giải thích như trên khi mà Generator tìm ra một khoảng dữ liệu đặc biệt mà tại điểm đó Discriminator không thể phân biệt được. Cũng là một hậu quả của việc chọn bộ siêu tham số không phù hợp.

Thuộc tính	**GAN**	**kNN**	**Random**
Relative_Humidity			
Specific_Humidity			

Precipitation			
Pressure			
Wind_Speed			
Wind_Direction			

Bảng 3. So sánh giữa dữ liệu được điền khuyết (xanh) và giá trị thực tế (cam).

Hơn nữa để kiểm định mô hình GAN của chúng em theo một hướng khác, chúng em sẽ sử dụng các bộ dữ liệu được điền khuyết của các phương pháp để xây mô hình hồi quy cho biến còn lại, chính là biến nhiệt độ (Temperature). Các mô hình chúng em sử dụng là Linear Regression, RFR, SVM, XGBoost và Neural Network (Bảng 4).

Dữ liệu	Train				Test			
	R^2	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE
Dữ liệu gốc	**0.99611**	**0.01486**	**0.1219**	**0.08272**	**0.99116**	**0.03512**	**0.18742**	**0.11614**
	RFR	RFR	RFR	RFR	SVM	SVM	SVM	SVM
GAN	**0.79118**	**0.79775**	**0.89317**	**0.36699**	**0.44393**	**2.20889**	**1.48623**	**1.00943**
	RFR	RFR	RFR	LR	RFR	RFR	RFR	RFR
Mean	**0.76513**	**0.89726**	**0.94724**	**0.65394**	**0.45953**	**2.14694**	**1.46524**	**0.97371**
	RFR	RFR	RFR	RFR	RFR	RFR	RFR	RFR
kNN	**0.82272**	**0.67727**	**0.82296**	**0.59118**	**0.36871**	**2.50769**	**1.58357**	**1.0561**
	RFR	RFR	RFR	RFR	XGBoost	XGBoost	XGBoost	XGBoost
Random	**0.42655**	**2.1907**	**1.4801**	**1.1805**	**-0.00173**	**3.97921**	**1.9948**	**1.54578**
	RFR	RFR	RFR	RFR	LR	LR	LR	LR

Bảng 4. Bảng so sánh độ tốt của các bộ dữ liệu được điền khuyết bởi các mô hình hồi quy khác nhau.

Nhìn vào bảng 4, ta thấy rằng tuy RMSE ở giai đoạn điền khuyết dữ liệu thì GAN không đạt hiểu quả bằng kNN, nhưng khi sử dụng dữ liệu do GAN điền khuyết cho tác vụ hồi quy với biến còn lại thì cho thấy được GAN có nhỉnh hơn kNN một phần nhỏ ở tập test. So với dữ liệu gốc và khi áp dụng phương pháp điền khuyết bằng giá trị Mean thì GAN không đạt hiệu quả bằng. Tuy nhiên với kết quả đạt được ở trên, thì ta thấy được rằng việc điền khuyết bằng GAN là một phương pháp mang trong mình một tiềm năng lớn. Vì các lý do giới hạn về mặt kiến thức, cũng như về mặt thời gian, nên phương pháp và kết quả không đạt được kết quả quá tốt. Những phần nào thử nghiệm này cũng mở ra được một phương hướng mới cho việc tiền xử lý dữ liệu, đặc biệt là việc điền khuyết dữ liệu.

3. KẾT LUẬN

Tóm lại, với bộ dữ liệu Weather Forecast của NASA Power chúng em đã tạo ra phần trăm khuyết cho dữ liệu. Sau đó, thực hiện một số thao tác phân tích thăm dò nhằm tìm ra tương quan của các biến với nhau. Tiếp theo, triển khai mô hình GAN với Generator và Discriminator sử dụng kiến trúc GRU. Kết quả sau bước này cho ra dữ liệu được điền khuyết có RMSE đã chuẩn hóa khá cao. Mặc dù vậy, khi đưa dữ liệu này vào các mô hình hồi quy cho biến còn lại thì kết quả cho khả quan hơn, thậm chí hơn cả kNN trên tập dữ liệu test.

Do GAN là kết hợp giữa hai model nên việc train song song 2 model này rất khó và nhạy cảm bởi các siêu tham số như hệ số học (learning rate), số mẫu huấn luyện (batch size), khung thời gian của sample (timesteps), giá trị biên của các biến (boundery)…. Theo như kết quả thực nghiệm cho thấy GAN không phải là điền một cách hoàn toàn ngẫu nhiên so với phương pháp điền khuyết ngẫu nhiên trong khoảng giá trị của nó. Tuy nhiên, việc GAN cho kết quả RMSE chuẩn hóa khá cao chúng em vẫn chưa giải thích được một phần vì mô hình khá nhạy cảm với siêu tham số, mô hình quá đơn giản hoặc không phù hợp hoặc dữ liệu quá ít. Những hướng phát triển trong tương lai của nhóm: tinh chỉnh bộ siêu tham số, tăng cường bộ dữ liệu, thay đổi kiến trúc mô hình sao cho phù hợp hơn.

TÀI LIỆU THAM KHẢO

[1] Yonghong Luo, Ying Zhang, Xiangrui Cai và Xiaojie Yuan , E2GAN: End-to-End Generative Adversarial Network for Multivariate Time Series Imputation, 2019.

[2] Jinsung Yoon, James Jordon, Mihaela van der Schaar, GAIN: Missing Data Imputation using Generative Adversarial Nets, 2018.

[3] NASA Power. Link: https://power.larc.nasa.gov/ (Truy cập ngày: 31/09/2022).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Hoàng Anh	- Chịu trách nhiệm chính cho việc xây dựng mô hình điền khuyết GAN. - Phân công, giám sát, quản lí công việc cho cả nhóm. - Tổng hợp thông tin, kết quả, soạn nội dung báo cáo và tạo dashboard ảnh.
2	Nguyễn Huỳnh Vương Quốc	- Phân tích thăm dò, trực quan hóa dữ liệu. - Xây dựng dashboard có thể tương tác bằng Flask. - Tổng hợp và phân chia source code thành module. - Tham gia vào việc tổng hợp và format báo cáo.
3	Ngô Huỳnh Trưởng	- Xây dựng và phân tích các mô hình hồi quy. - Đưa ra các đánh giá, nhận xét về kết quả xây dựng mô hình hồi quy.
4	Nguyễn Hữu Minh Tâm	- Phân tích thăm dò, trực quan hóa dữ liệu. - Tham gia vào việc tổng hợp và format báo cáo. - Tạo slide báo cáo.