

Áp dụng mô hình lai RNN-LSTM trong dự báo tổng sản phẩm quốc nội bình quân đầu người tại Việt Nam

Applying a hybrid RNN-LSTM model in forecasting Vietnam's GDP per capita

Phạm Thị Hà An

Khoa Tài chính Ngân hàng, Trường Đại học Văn Lang, Tp. Hồ Chí Minh

Email: an.pth@vlu.edu.vn

Tóm tắt: Dự báo GDP truyền thống thường gặp hạn chế do các kỹ thuật thống kê phụ thuộc vào dữ liệu điều tra kinh tế và tính dừng của chuỗi thời gian, vốn khó đáp ứng trong thực tế. Những đặc điểm phức tạp như tính mùa vụ, biến động đột ngột và ảnh hưởng từ nhiều yếu tố bên ngoài khiến các mô hình dự báo này thiếu ổn định và độ chính xác thấp. Nghiên cứu này được thực hiện nhằm đề xuất một mô hình lai kết hợp giữa RNN và LSTM nhằm tận dụng tối đa ưu điểm của từng mô hình để cải thiện đáng kể chất lượng dự báo GDP bình quân đầu người tại Việt Nam. Kết quả nghiên cứu cho thấy mô hình lai RNN-LSTM được đề xuất thể hiện khả năng khái quát hóa mạnh mẽ, đưa ra dự đoán ổn định về GDP bình quân đầu người dựa trên các biến số vĩ mô nền kinh tế. Hiệu suất của mô hình này vượt trội hơn đáng kể so với các mô hình CNN, RNN và LSTM.

Từ khóa: GDP bình quân đầu người; Mô hình học máy; RNN-LSTM

Abstract: Traditional GDP forecasting approaches often face limitations due to their reliance on statistical techniques that depend on economic survey data and the stationarity of time series, conditions which are challenging to meet in practice. Complex characteristics such as seasonality, sudden fluctuations, and external influences contribute to instability and low accuracy in these forecasting models. This study aims to propose a hybrid model combining RNN and LSTM to leverage the strengths of each approach, significantly improving the forecasting accuracy of per capita GDP in Vietnam. The research results indicate that the proposed RNN-LSTM hybrid model demonstrates robust generalization capabilities, providing stable predictions of per capita GDP based on macroeconomic variables. Its forecasting performance is significantly superior to CNN, RNN, and LSTM models.

Keywords: GDP per capita; Machine learning model; RNN-LSTM

1. Giới thiệu nghiên cứu

Dự báo GDP truyền thống thường dựa vào các kỹ thuật thống kê thu được từ điều tra kinh tế, chẳng hạn như dự báo xu hướng, mô hình hồi quy và phân tích dữ liệu chuỗi thời gian [1]. Tuy nhiên, các phương pháp tiếp cận truyền thống này có những hạn chế. Các kỹ thuật dự báo xu hướng dựa trên nghiên cứu thực nghiệm thường đòi hỏi nhiều công sức và mất nhiều thời gian, trong khi kết quả từ các cuộc điều tra kinh tế thường bị trễ so với các diễn biến thực tế. Một rào cản thực tế lớn nảy sinh khi dữ liệu chuỗi thời gian, chẳng hạn như GDP và kiều hối, thường không đáp ứng các giả định lý tưởng của phân tích chuỗi thời gian truyền thống và phương pháp phát hiện quan hệ nhân quả. Những giả định này thường là giả định về tính dừng của các chuỗi thời gian này. Thực tế, giả định này thường bị vi phạm do có các biến động theo mùa, các sự thay đổi đột ngột và những đặc điểm phức tạp khác [2]. Chẳng hạn, sự thay đổi trong lượng

kiều hối vào các thời điểm khác nhau thách thức giả định về tính dừng. GDP chịu ảnh hưởng của nhiều yếu tố bên ngoài, chẳng hạn như mức độ phát triển kinh tế, ưu tiên chính sách công, điều kiện khí hậu địa phương và thu nhập bình quân. Hơn nữa, GDP có các đặc điểm phức tạp như biến động, vòng lặp phản hồi và tính chu kỳ, khiến các phương pháp dự báo truyền thống gặp nhiều khó khăn.

Trong kỷ nguyên dữ liệu lớn, có những cách tiếp cận thay thế để dự báo GDP, chẳng hạn như sử dụng các mô hình dữ liệu và thuật toán trí tuệ nhân tạo (AI). Các mô hình này giúp xác định và lựa chọn các yếu tố tác động quan trọng, dẫn đến việc xây dựng các mô hình dự báo GDP mạnh mẽ hơn [3]. Nhiều kỹ thuật AI và học máy đã được áp dụng để dự báo các chỉ số kinh tế quan trọng, chẳng hạn như tăng trưởng GDP, chỉ số giá tiêu dùng (CPI), tỷ lệ thất nghiệp, mức tiêu thụ năng lượng, xuất khẩu và lãi suất. Các phương pháp này bao gồm mạng nơ-ron nhân tạo (ANNs),

hồi quy vector hỗ trợ (SVR),... [4]. Do những hạn chế của các kỹ thuật học máy truyền thống, cộng đồng nghiên cứu ngày càng hướng tới các mô hình học sâu như RNN, LSTM, CNN, GRU và SVM [5]. Tuy nhiên, khi được sử dụng riêng lẻ các mô hình này cũng tuân theo cách tiếp cận huấn luyện biệt lập, trong đó mỗi mô hình được đào tạo độc lập cho một nhiệm vụ cụ thể. Để khắc phục những hạn chế này, trong nghiên cứu này, tác giả sử dụng một mô hình lai giữa RNN và LSTM để tận dụng lợi thế của từng mô hình trong việc dự báo.

Phần còn lại của nghiên cứu này được tổ chức như sau: Phần 2 trình bày cơ sở lý thuyết nhằm tổng quan các nghiên cứu trước đây về phương pháp dự báo GDP, Phần 3 trình bày dữ liệu và phương pháp nghiên cứu đề xuất, Phần 4 trình bày kết quả và thảo luận, Phần 5 đưa ra kết luận và khuyến nghị cho nghiên cứu.

2. Cơ sở lý thuyết

Nhiều nghiên cứu đã khám phá dự báo GDP bằng máy học, cung cấp nhiều cách tiếp cận đa dạng và góc nhìn có giá trị. Dự đoán dữ liệu tuần tự là điều cần thiết trong nhiều lĩnh vực, bao gồm khí tượng học, địa chấn học, kinh tế học và tài chính. Khi tác giả đi sâu vào dự đoán chuỗi thời gian, đã gặp nhiều chiến lược khác nhau: dự báo thống kê, khám phá máy học và sự xuất hiện của các mô hình lai [6]. Dữ liệu chuỗi thời gian trong thế giới thực thường thể hiện tính phi tuyến tính và tính không dừng, thách thức các phương pháp dự đoán tuyến tính thông thường. Có hai cách tiếp cận nổi bật trong dự báo chuỗi thời gian tài chính: các phương pháp kinh tế lượng thông thường và các kỹ thuật máy học [7].

Tuy nhiên, các mô hình kinh tế lượng thông thường với các tham số cố định thường gặp khó khăn trong việc phân tích và làm phức tạp dữ liệu tài chính đa chiều và dễ biến động. Để ứng phó với những hạn chế này, mạng nơ-ron đã trở nên phổ biến trong dự đoán GDP do khả năng rút ra hiểu biết sâu sắc từ dữ liệu thô tần suất cao mà không cần giả định trước [3].

Trong những năm gần đây, cộng đồng nghiên cứu ngày càng có xu hướng chuyển từ các mô hình dự báo truyền thống sang các kỹ thuật học sâu nhằm nâng cao độ chính xác và

khả năng xử lý tính phi tuyến, không dừng của dữ liệu kinh tế vĩ mô. Các mô hình học sâu như RNN, LSTM, GRU, CNN và SVM đã được áp dụng rộng rãi trong dự báo GDP, cho thấy tiềm năng vượt trội so với các mô hình kinh tế lượng cổ điển. Đặc biệt, nhiều nghiên cứu ghi nhận hiệu suất dự đoán nổi bật khi kết hợp các mạng nơ-ron với kỹ thuật tối ưu như di truyền (GA), thuật toán bầy đàn (PSO), hoặc các phương pháp xử lý tín hiệu như CEEMDAN, cho ra đời các mô hình lai (hybrid models) như PC-LSTM-RNN, LSTM-GCN-CEEMDAN, RBFNN-GA, và DQN-GRU-TCN-DBN. Những mô hình lai này không chỉ cải thiện đáng kể độ chính xác, chẳng hạn R^2 đạt đến 99,99% như trong nghiên cứu của Shams et al. [2], mà còn chứng minh được tính thích nghi cao với các tập dữ liệu phức tạp ở các khu vực khác nhau.

Tuy nhiên, dù có nhiều thành tựu nổi bật, việc nghiên cứu về các mô hình lai vẫn còn hạn chế. Phần lớn các công trình hiện nay chỉ dừng lại ở so sánh hiệu suất giữa các mô hình đơn lẻ hoặc đơn giản kết hợp hai mô hình, mà chưa thật sự đi sâu vào hệ thống hóa, tối ưu hóa và khai thác toàn diện tiềm năng của các mô hình lai trong dự báo GDP. Khoảng trống này mở ra một hướng nghiên cứu giàu tiềm năng, thúc đẩy các công trình tương lai khai thác mạnh mẽ hơn sự kết hợp giữa các kiến trúc học sâu tiên tiến và các chiến lược xử lý dữ liệu thông minh nhằm nâng cao hiệu quả dự báo kinh tế vĩ mô trong bối cảnh biến động toàn cầu.

3. Dữ liệu và mô hình nghiên cứu

Phần này mô tả về khung nghiên cứu mô hình lai đề xuất (RNN-LSTM). Đầu tiên, tập dữ liệu thu thập được làm sạch và thực hiện các thống kê mô tả dữ liệu, áp dụng phương pháp ước lượng trung bình và kiểm tra chất lượng dữ liệu thô để chuẩn bị cho việc mô hình hóa. Trong nghiên cứu này phương pháp kỹ thuật ước lượng trung bình và chuẩn hóa dữ liệu được sử dụng.

3.1. Dữ liệu nghiên cứu

Trong nghiên cứu này, để đảm bảo phân tích toàn diện, nghiên cứu đã kết hợp các biến vĩ mô như FDI, lạm phát, tỷ giá hối đoái và hỗ trợ phát triển chính thức (ODA), cùng với các yếu tố khác để dự báo GDP bình quân đầu người của Việt Nam. Chi tiết các biến đầu vào

và mô tả được trình bày trong bảng bên dưới. Để huấn luyện mô hình lai, nghiên cứu đã thực hiện thu thập dữ liệu tự động từ cơ sở dữ liệu Chỉ số Phát triển của Ngân hàng Thế giới

thông qua API. Dữ liệu nghiên cứu được lấy theo năm trong khoảng thời gian từ 1999 đến 2023. Bảng 1 mô tả các đặc điểm chính của các tập dữ liệu.

Bảng 1. Bảng tổng hợp các biến trong nghiên cứu

Biến	Mô tả
year	Các năm dữ liệu được ghi nhận trong khoảng thời gian từ 1999 đến 2023
Country	Biến ghi nhận quốc gia là Việt Nam
GDP	GDP bình quân đầu người của Việt Nam, tính theo USD
FDI	Đầu tư trực tiếp nước ngoài, dòng vào, tính theo % GDP
INF	Lạm phát được đo bằng chỉ số giá tiêu dùng.
Remittances	Kiểu hối cá nhân, tính theo % GDP
migration	Tổng số người di cư ròng trong giai đoạn này, tính bằng số người nhập cư trừ đi số người di cư, bao gồm cả công dân và người không phải công dân.
Population	Tổng số người sống trong cả nước
Trade	Độ mở thương mại, tính theo % GDP
GOV	Chi tiêu của chính phủ cho giáo dục, tính theo % chi tiêu của chính phủ

3.2. Mô hình RNN-LSTM

Một RNN bao gồm các nơ-ron được kết nối thông qua các trọng số, cho phép xử lý các đầu vào có kích thước khác nhau. RNN rất hữu ích trong việc xử lý dữ liệu tuần tự, chẳng hạn như các biến vĩ mô và GDP, bằng cách duy trì một trạng thái nội bộ mã hóa thông tin về các thời điểm quan sát. LSTM, một loại cụ thể của RNN, cải thiện điều này bằng cách giải quyết vấn đề gradient biến mất và nắm bắt các phụ thuộc dài hạn hiệu quả hơn [2]. Nghiên cứu này áp dụng phương pháp đầu vào tuần tự để trình bày các thuộc tính của mô hình theo thứ tự cụ thể. Mô hình RNN-LSTM lai tích hợp những lợi thế của cả hai khung RNN và LSTM, hiệu quả trong việc nắm bắt các phụ thuộc tạm thời thông qua định dạng tuần tự này. Sự kết hợp của các phương pháp giữ lại bộ nhớ từ RNN và LSTM, kết hợp với cách tiếp cận đầu vào tuần tự, cho phép mô hình nhận diện các phụ thuộc trong dữ liệu một cách hiệu quả hơn, dẫn đến các dự đoán chính xác hơn. Kiến trúc của mô hình RNN-LSTM lai được minh họa trong Hình 3 [8], bao gồm lớp đầu vào, lớp ẩn RNN, lớp ẩn LSTM, lớp kết nối đầy đủ và lớp đầu ra. Nghiên cứu triển khai ba cổng: cổng quên, cổng đầu vào và cổng đầu ra. Lý do chính để sử dụng những mạng này là dự đoán tăng trưởng GDP là một nhiệm vụ hồi quy. Ban đầu, dữ liệu được đưa vào RNN, tạo ra một đầu ra được ký hiệu là

O_t^R trong Phương trình 11. Đối với dự đoán cuối cùng $\hat{y}_t$, đầu ra của RNN được đưa vào lớp LSTM. Kết quả từ mô hình lai RNN-LSTM được thể hiện dưới các phương trình toán học như sau:

$$i_t^H = \sigma(W_{xi}O_t^R + U_{hi}h_{(t-1)}^L + b_i) \quad (11)$$

$$f_t^H = \sigma(W_{xf}O_t^R + U_{hf}h_{(t-1)}^L + b_f) \quad (12)$$

$$o_t^H = \sigma(W_{xo}O_t^R + U_{ho}h_{(t-1)}^L + b_o) \quad (13)$$

$$U_t^H = \tanh(W_{xu}O_t^R + U_{hu}h_{(t-1)}^L + b_u) \quad (14)$$

$$c_t^H = f_t^H \odot C_{(t-1)}^H + i_t^H \odot U_t^H \quad (15)$$

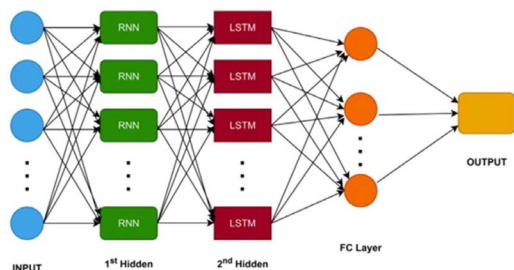
$$h_t^H = \tanh(c_t^H) \odot o_t^H \quad (16)$$

Và cuối cùng, giá trị dự đoán $\hat{y}_t$ được cho thể hiện qua Phương trình 17:

$$\hat{y}_t = \sigma(W_o h_t^H) \quad (17)$$

Trong đó: h_t^H đại diện cho trạng thái ẩn của ô LSTM như được định nghĩa bởi Phương trình 16. Trạng thái này được lấy từ việc sử dụng hàm kích hoạt *tanh* trên trạng thái ô đã sửa đổi, sau đó nhân với cổng đầu ra, O_t^H , dữ liệu được đầu ra từ mô hình RNN là i_t^H (là cổng đầu vào cho mô hình lai tại thời điểm t).

Theo Islam et al. [9], đây là các mô hình mạng nơ-ron tiên tiến có khả năng đạt độ chính xác cao hơn so với các dự đoán dựa trên hồi quy.



Hình 1. Kiến trúc mô hình lai RNN-LSTM

4. Kết quả và thảo luận

Cấu hình của cả phần mềm và phần cứng được sử dụng trong thí nghiệm nghiên cứu được trình bày trong Bảng 2.

4.1. Tiền xử lý dữ liệu

Bảng 2. Thiết lập môi trường thử nghiệm

Processor (CPU)	Core(TM) i7-7600U CPU@2.80GHz
Graphics Card (GPU)	Intel(R) HD Graphics 620
Memory (RAM)	16.00 GB
Programming language	Python 3.12.0
Deep learning framework	TensorFlow v2.16.1

Bảng 3. Kết quả xử lý dữ liệu thiếu

Trước khi xử lý				Sau khi xử lý			
<class 'pandas.core.frame.DataFrame'> Index: 25 entries, 1999 to 2023 Data columns (total 9 columns):				<class 'pandas.core.frame.DataFrame'> Index: 25 entries, 1999 to 2023 Data columns (total 9 columns):			
#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype
0	country	25 non-null	object	0	country	25 non-null	object
1	GDP	25 non-null	float64	1	GDP	25 non-null	float64
2	FDI	25 non-null	float64	2	FDI	25 non-null	float64
3	INF	25 non-null	float64	3	INF	25 non-null	float64
4	Remittances	24 non-null	float64	4	Remittances	25 non-null	float64
5	migration	25 non-null	int64	5	migration	25 non-null	int64
6	Population	25 non-null	int64	6	Population	25 non-null	int64
7	Trade	25 non-null	float64	7	Trade	25 non-null	float64
8	GOV	15 non-null	float64	8	GOV	25 non-null	float64
dtypes: float64(6), int64(2), object(1) memory usage: 2.0+ KB				dtypes: float64(6), int64(2), object(1) memory usage: 2.0+ KB			

4.1.2 Chia tập dữ liệu và chuẩn hóa

Tập dữ liệu được chia thành 2 phần: **80% dữ liệu dành cho huấn luyện, 20% cho kiểm tra**. Sau khi chia tập dữ liệu, tác giả tiếp tục chuẩn hóa các biến bằng hàm **MinMaxScaler**. Việc chuẩn hóa dữ liệu giúp loại bỏ các sai lệch tiềm ẩn bằng cách chuyển đổi tất cả giá trị về một phạm vi chuẩn hóa từ [0,1].

Hàm này điều chỉnh từng cột dữ liệu dựa trên giá trị nhỏ nhất và lớn nhất của dữ liệu, đảm bảo mọi giá trị trong cùng một cột được chuẩn hóa đồng nhất trên toàn bộ tập dữ liệu. Việc chuẩn hóa được thực hiện sau quá trình chia tập dữ liệu nhằm đảm bảo rằng cả tập

4.1.1 Xử lý giá trị thiếu

Để xử lý dữ liệu bị thiếu, hai chiến lược chính thường được áp dụng [10] gồm: (i) bù đắp giá trị thiếu hoặc (ii) loại bỏ các mẫu dữ liệu có chứa giá trị khuyết. Trong trường hợp bù đắp đơn biến, các phương pháp phổ biến bao gồm bù đắp trung bình hoặc trung vị, chuyển tiếp hoặc sử dụng độ trễ, cũng như bù đắp bằng trung bình động [11]. Trong khi đó, bù đắp đa biến thường dựa trên các phương pháp sử dụng KNN hoặc mô hình hồi quy. Nghiên cứu này áp dụng phương pháp bù đắp đơn biến bằng giá trị trung bình để xử lý dữ liệu bị thiếu.

huấn luyện và tập kiểm tra đều được chuẩn hóa theo các tham số của tập huấn luyện, phù hợp với nguyên tắc chỉ sử dụng dữ liệu huấn luyện để điều chỉnh mô hình thay vì dữ liệu kiểm tra. Tiếp theo, dữ liệu được chuyển đổi sang định dạng **mảng 3D** để tương thích với các mô hình **LSTM (Long Short-Term Memory)** và **RNN (Recurrent Neural Network)** trong thư viện Keras của Python, được xây dựng trên nền tảng TensorFlow.

Tập dữ liệu huấn luyện được định dạng dưới dạng **(17, 3, 7)** cho tập biến đầu vào là các biến vĩ mô được sử dụng để dự báo GDP bình quân đầu người và dạng **(17,1)** cho biến mục tiêu GDP bình quân đầu người.

Tập kiểm tra dưới dạng (2, 3, 7) cho tập biến đầu vào là các biến vĩ mô được sử dụng để dự báo GDP bình quân đầu người và dạng (2,1) cho biến mục tiêu GDP bình quân đầu người, sử dụng phương pháp cửa sổ trượt (sliding window). Cấu trúc này đại diện cho số lượng mẫu, số bước thời gian (time

steps), và số đặc trưng (features) trong mỗi tập dữ liệu.

Trong đó, 3 bước thời gian phản ánh số quan sát trước đó được sử dụng để dự đoán, và 7 đặc trưng là các biến vĩ mô được sử dụng để dự báo GDP bình quân đầu người.

4.2. Phân tích thống kê

Bảng 4. Thống kê mô tả mẫu

	GDP	FDI	INF	Remittances	migration	Population	Trade	GOV
count	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
mean	1953.39	4.84	5.72	4.46	23336.68	88350250.80	140.89	16.23
std	1303.31	1.51	5.45	1.26	215904.38	7838216.28	22.97	1.11
min	375.99	3.39	-1.71	3.09	-220458.00	76287452.00	102.79	14.47
25%	710.75	3.99	3.16	3.34	-81789.00	81088313.00	124.33	15.45
50%	1950.93	4.31	3.83	4.30	-4979.00	88468314.00	135.41	16.23
75%	2956.11	4.93	7.75	5.14	-2051.00	95176977.00	160.98	16.23
max	4282.09	9.66	23.12	7.98	571113.00	100352192.0	186.68	18.78

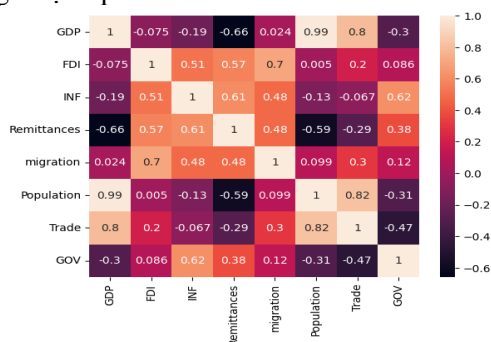
GDP bình quân đầu người của Việt Nam trong giai đoạn nghiên cứu dao động khá rộng, từ mức thấp nhất là 375.99 USD đến mức cao nhất là 4,282.09 USD. Giá trị trung bình là 1,953.39 USD, trong khi độ lệch chuẩn ở mức 1,303.31 cho thấy sự biến động lớn qua các năm. Phân vị 25%, 50% và 75% lần lượt là 710.75, 1,950.93 và 2,956.11 USD, phản ánh sự gia tăng đều đặn qua thời gian. FDI được tính theo tỷ lệ phần trăm GDP, với mức trung bình là 4.84% và độ lệch chuẩn tương đối thấp (1.51), cho thấy sự ổn định hơn so với GDP. Giá trị tối thiểu là 3.39% và tối đa là 9.66%, khoảng cách này không quá lớn, thể hiện luồng vốn FDI vào Việt Nam duy trì ở mức ổn định qua các năm. Lạm phát (INF), đo bằng chỉ số giá tiêu dùng, có giá trị trung bình là 5.72%, nhưng độ lệch chuẩn cao (5.45%) cho thấy sự biến động đáng kể qua thời gian. Đáng chú ý, có năm lạm phát rơi xuống mức âm (-1.71%), thể hiện tình trạng giảm phát, trong khi mức cao nhất đạt 23.12%, cho thấy đã từng có giai đoạn lạm phát bùng phát mạnh. Kiều hối cá nhân đóng góp trung bình 4.46% GDP với mức thấp nhất là 3.09% và cao nhất là 7.98%. Độ lệch chuẩn

là 1.26, phản ánh sự biến động nhẹ. Giá trị các phân vị (3.34%, 4.30%, 5.14%) cho thấy kiều hối duy trì vai trò khá ổn định như một nguồn thu ngoại hối của nền kinh tế. Di cư ròng cho thấy sự biến động rất mạnh, dao động từ -220,458 đến 571,113 người. Giá trị trung bình là 23,336 người nhưng độ lệch chuẩn rất lớn (215,904), phản ánh sự không đồng đều giữa các năm. Phân vị giữa (-4,979 người) cho thấy phần lớn các năm Việt Nam có xu hướng di cư ròng âm, tức là số người rời khỏi lớn hơn số người nhập cư. Tổng dân số trung bình của Việt Nam trong giai đoạn nghiên cứu là khoảng 88.35 triệu người, với mức dao động từ 76.29 triệu đến hơn 100 triệu người. Độ lệch chuẩn ở mức 7.83 triệu người, cho thấy dân số tăng dần theo thời gian. Các phân vị 25%, 50%, và 75% lần lượt là 81.08, 88.47 và 95.18 triệu người, phản ánh sự gia tăng dân số đều đặn. Độ mở thương mại trung bình đạt 140.89% GDP, cho thấy nền kinh tế Việt Nam có mức độ hội nhập quốc tế rất cao. Khoảng dao động từ 102.79% đến 186.68% thể hiện sự thay đổi rõ rệt theo từng năm, với độ lệch chuẩn là 22.97. Điều này phản ánh chính sách mở cửa và tăng trưởng mạnh trong xuất nhập

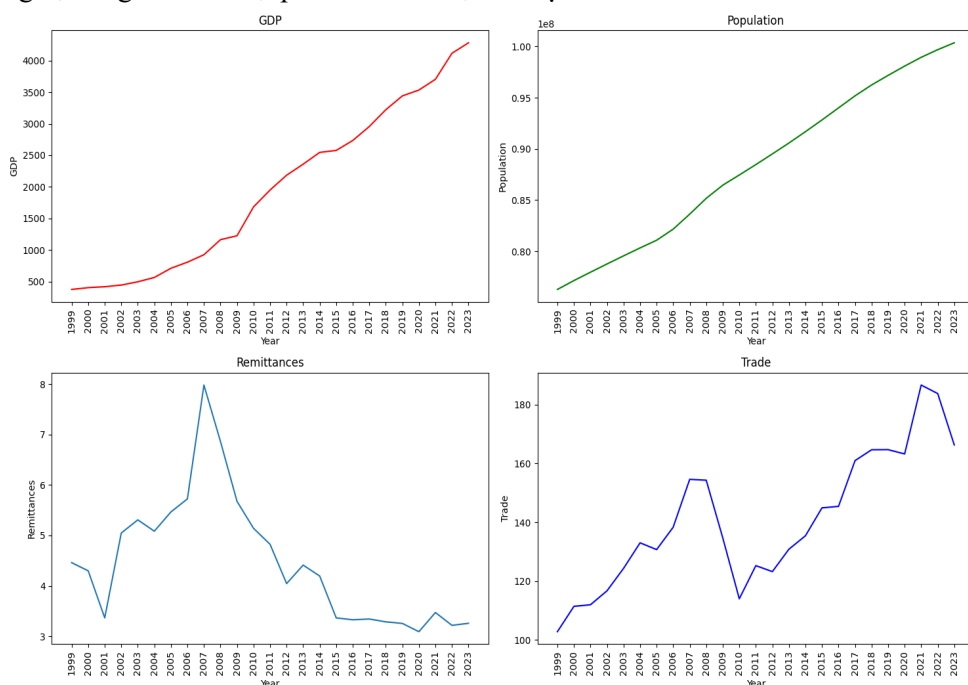
khẩu. Tỷ lệ chi tiêu chính phủ cho giáo dục trung bình đạt 16.23%, với mức thấp nhất là 14.47% và cao nhất là 18.78%. Biến số này có độ lệch chuẩn thấp (1.11), cho thấy mức chi tiêu tương đối ổn định qua thời gian. Các phân vị không có sự khác biệt quá lớn, phản ánh cam kết đều đặn của chính phủ đối với lĩnh vực giáo dục.

Hình 2 minh họa phân tích bản đồ nhiệt của các đặc trưng trong tập dữ liệu. Bản đồ nhiệt cung cấp cái nhìn trực quan về mức độ tương quan giữa các biến, hỗ trợ trong việc hình dung dữ liệu trên toàn bộ tập hợp đặc trưng. Phân tích này đánh giá mối quan hệ giữa các biến đầu vào với biến mục tiêu (GDP), từ đó phát hiện các quy luật và xu hướng tiềm ẩn trong dữ liệu. Sự thay đổi về cường độ trong bản đồ nhiệt phản ánh mức độ

và hướng của tương quan, giúp làm sáng tỏ các mối liên kết giữa các yếu tố. Đáng chú ý, đường chéo trên bản đồ thể hiện tương quan dương mạnh (được mã màu xanh lá), trong khi tương quan âm (được mã màu đỏ) thể hiện giá trị thấp



Hình 2. Ma trận hệ số tương quan



Hình 3. Biểu đồ biến động của một số biến theo thời gian

Dựa trên ma trận tương quan, biến GDP bình quân đầu người có mối tương quan dương rất mạnh với dân số (hệ số tương quan 0.99), cho thấy rằng sự gia tăng dân số đi đôi với sự tăng trưởng kinh tế tại Việt Nam trong giai đoạn nghiên cứu. Ngoài ra, độ mở thương mại cũng có mối tương quan dương mạnh với GDP (0.80), phản ánh vai trò quan trọng của hoạt động xuất nhập khẩu trong việc thúc đẩy tăng trưởng. Ngược lại, kiều hối lại có mối tương quan âm khá rõ (-0.66), cho thấy dòng tiền từ nước ngoài gửi về có thể gắn với việc người

lao động di cư ra nước ngoài do trong nước thiếu cơ hội, từ đó làm giảm GDP bình quân đầu người. Chi tiêu chính phủ cho giáo dục có mối tương quan âm nhẹ (-0.30), có thể do tác động của chính sách giáo dục cần thời gian dài để thể hiện trong tăng trưởng kinh tế. Các biến như lạm phát (-0.19), đầu tư trực tiếp nước ngoài (-0.075), và di cư ròng (0.024) đều có tương quan rất yếu với GDP, cho thấy ảnh hưởng không rõ ràng hoặc không đáng kể trong mẫu nghiên cứu này.

Biểu đồ GDP cho thấy xu **hướng tăng liên tục và rõ rệt** qua các năm từ 1999 đến 2023. Mức tăng mạnh đặc biệt được ghi nhận sau năm 2008, với tốc độ tăng trưởng nhanh hơn trong giai đoạn 2010–2023. Điều này phản ánh sự chuyển mình mạnh mẽ của nền kinh tế Việt Nam, với hiệu quả cải cách kinh tế, thu hút đầu tư nước ngoài và mở rộng thị trường xuất khẩu. Xu hướng đi lên bền vững này cũng cho thấy GDP có thể là một biến mục tiêu phù hợp khi đánh giá tác động của các yếu tố kinh tế vĩ mô.

Biểu đồ dân số thể hiện một xu hướng **tăng đều đặn và ổn định**, từ dưới 80 triệu người năm 1999 lên trên 100 triệu người vào năm 2023. Không có sự biến động đột ngột nào trong dữ liệu này, cho thấy quá trình tăng dân số tự nhiên của Việt Nam diễn ra liên tục. Điều này phù hợp với tương quan rất mạnh giữa dân số và GDP như đã phân tích trước đó – dân số càng tăng, GDP bình quân đầu người cũng tăng theo.

Biểu đồ kiều hối có **dạng hình chuông**, với mức tăng mạnh từ năm 2002 và đạt đỉnh vào khoảng năm 2007–2008, sau đó **giảm dần và ổn định** ở mức thấp hơn từ năm 2014 đến 2023. Giai đoạn tăng mạnh có thể liên quan đến sự bùng nổ xuất khẩu lao động, đặc biệt sang các thị trường Mỹ, Nhật Bản, Hàn Quốc. Tuy nhiên, sự suy giảm sau đó có thể do biến động kinh tế toàn cầu (khủng hoảng tài chính 2008), chính sách nhập cư của các nước phát triển, hoặc sự cải thiện điều kiện trong nước khiến người dân ít gửi tiền về hơn.

Biểu đồ thể hiện mức độ thương mại (tính theo % GDP) có xu hướng **tăng dần theo thời gian**, nhưng với **những đợt biến động rõ rệt**. Các mốc dao động lớn xảy ra vào khoảng năm 2007–2009 (thời kỳ khủng hoảng tài chính toàn cầu), 2012 (có thể do điều chỉnh chính sách thương mại), và 2020–2021 (ảnh hưởng của đại dịch COVID-19). Tuy nhiên, xu hướng tổng thể là tăng – phản ánh sự hội nhập kinh tế quốc tế ngày càng sâu rộng của Việt Nam, qua các hiệp định thương mại tự do (FTA) và vai trò của Việt Nam trong chuỗi cung ứng toàn cầu.

4.3. Chỉ số hiệu suất

Để thể hiện độ mạnh mẽ của mô hình lai RNN-LSTM đề xuất, bốn chỉ số phân tích sự không chắc chắn được giới thiệu gồm: (i) Sai

số căn bậc hai trung bình (RMSE), (ii) Sai số tuyệt đối trung bình (MAE), (iii) Sai số phần trăm tuyệt đối trung bình (MAPE) và (iv) hệ số xác định (R^2) được thể hiện trong các công thức sau:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{y}_t - y_t)^2} \quad (18)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (19)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (20)$$

$$R^2 = 1 - \frac{\sum_{t=1}^N (\hat{y}_t - y_t)^2}{\sum_{t=1}^N (\bar{y}_t - \bar{y})^2} \quad (21)$$

Trong đó N là tổng số điểm dữ liệu, y_t là giá trị thực tế và $\bar{y}_t$ biểu thị giá trị dự đoán.

4.4. Thảo luận kết quả

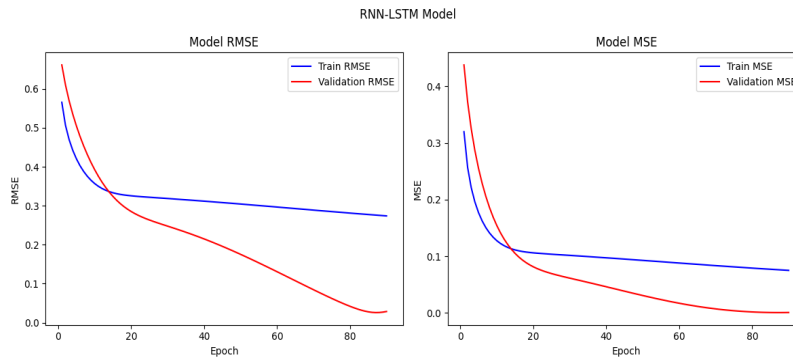
Trong phần này, nghiên cứu tiến hành điều tra khả năng dự đoán của mô hình được thiết lập, thông qua các thông số của các siêu tham số (RMSE), độ hội tụ và độ mạnh mẽ của mô hình lai (RNN-LSTM). Dữ liệu được sử dụng trong huấn luyện CNN, RNN, LSTM và mô hình lai RNN-LSTM đã được trình bày chi tiết trong các phần trước. Nghiên cứu thực hiện huấn luyện mô hình lai RNN-LSTM và các mô hình độc lập bằng cách sử dụng tập huấn luyện.

Để giảm thiểu RMSE và tổn thất trong huấn luyện và kiểm tra, nghiên cứu đã tiến hành thử nghiệm nhiều cấu hình của các siêu tham số chính (số lượng lớp ẩn, số nơ ron trong mỗi lớp ẩn, tỷ lệ học, kích thước batch, số lần epochs) bằng cách sử dụng tìm kiếm ngẫu nhiên cho mô hình lai RNN-LSTM, các kết quả thử nghiệm được trình bày trong Bảng 5.

Bảng 5. Các kết quả siêu tham số

Siêu tham số	Giá trị
Số lớp ẩn	3
Số nơ ron trong mỗi lớp ẩn	20–128
Kích thước Batch	16 - 64
Số lần epochs	20–100

Để khám phá hiệu quả không gian tìm kiếm rộng lớn, tìm kiếm ngẫu nhiên được sử dụng trong nghiên cứu này như một phương pháp để xác định các thiết lập siêu tham số tối ưu.



Hình 4. RMSE và các tổn thất trong huấn luyện trước khi tối ưu hóa tham số

Sự kết hợp tốt nhất của các siêu tham số, dẫn đến lỗi thấp nhất cho mô hình lai RNN-LSTM cho thấy rằng tỷ lệ học 0.0001 mang lại hiệu suất RMSE tốt nhất, đảm bảo hội tụ ổn định và đường cong học mượt hơn, trong khi các tỷ lệ học cao hơn (ví dụ: 0.001 trở lên) dẫn đến việc học ban đầu nhanh hơn nhưng có nhiều dao động hơn trong RMSE, cho thấy khả năng quá khớp. Đối với kích thước Batch cho mô hình lai RNN-LSTM, thiết lập 40 mang lại giá trị RMSE thấp nhất trên các đơn vị, cho thấy khả năng tổng quát tốt hơn do tính biến đổi nhiều hơn trong các cập nhật gradient, trong khi kích thước lớn hơn (ví dụ: 64) cho thấy thời gian huấn luyện nhanh hơn nhưng dễ gặp vấn đề hội tụ, thường dẫn đến điểm RMSE cao hơn do chưa khớp. Tương tự như vậy, giá trị tối ưu cho các siêu tham số khác được trình bày trong bảng 6 bên dưới.

Bảng 6. Bộ siêu tham số tối ưu

Siêu tham số	Giá trị
Số lớp ẩn	2
Số nơ rôn trong mỗi lớp ẩn	128
Kích thước Batch	40
Số lần epochs	90
Tỷ lệ học	0.0001
Optimizer	Adam
Time steps	3
Activation function	Relu
Loss function	Mean squared error

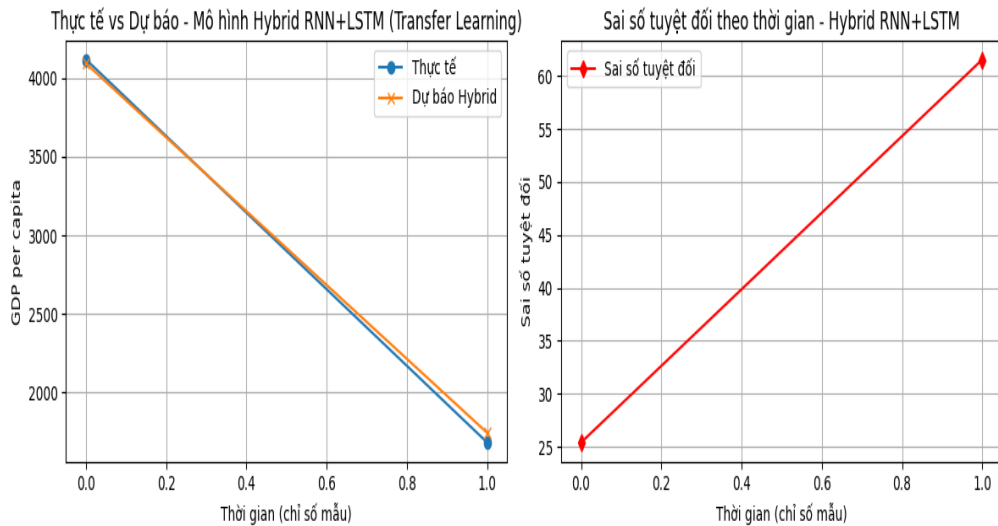
Sau khi chọn được bộ siêu tham số tối ưu, nghiên cứu sử dụng các tham số này để huấn luyện và thực hiện kiểm tra mức độ dự báo của mô hình lai RNN-LSTM trên tập kiểm tra. Bên cạnh đó, các mô hình riêng lẻ CNN, RNN, LSTM cũng được huấn luyện và thực

hiện kiểm tra mức độ dự báo chính xác. Kết quả kiểm tra mức độ dự báo chính xác của các mô hình được trình bày trong bảng 7 bên dưới.

Bảng 7. Đánh giá mức độ dự báo của các mô hình

	model	RMSE	MAE	MAPE (%)	R2
0	Hybrid RNN+LSTM (TL)	47.080388	43.473418	2.136822	0.998502
1	CNN	736.408141	553.217803	14.618790	0.633619
2	RNN	162.838812	119.977264	3.088071	0.982085
3	LSTM	852.660959	729.516937	22.781399	0.508811

Bảng kết quả dự báo GDP bình quân đầu người cho thấy sự khác biệt rõ rệt về hiệu suất giữa các mô hình học sâu. Mô hình lai RNN+LSTM đạt hiệu quả vượt trội nhất với RMSE chỉ 47.08 và MAE là 43.47, chứng tỏ sai số dự báo rất nhỏ. Đồng thời, mô hình này có MAPE chỉ 2.14%, tức là sai lệch phần trăm trung bình so với thực tế rất thấp. Giá trị $R^2 = 0.9985$ cho thấy mô hình giải thích gần như toàn bộ phương sai của dữ liệu, tức là có độ chính xác cực cao. Trong khi đó, mô hình RNN đơn thuần có hiệu suất cũng khá tốt với RMSE là 162.83 và R^2 đạt 0.9821, tuy nhiên vẫn kém xa mô hình lai về sai số. Mô hình CNN có độ chính xác thấp hơn, với RMSE lên đến 736.41 và R^2 chỉ đạt 0.6336, cho thấy không phù hợp lắm với chuỗi thời gian. Mô hình LSTM thuần túy là yếu nhất trong số các mô hình, với sai số cao nhất (RMSE = 852.66, MAE = 729.52) và R^2 thấp nhất (0.5088). Kết quả này cho thấy việc kết hợp RNN và LSTM, đặc biệt khi ứng dụng học chuyển giao, mang lại hiệu quả dự báo tốt nhất. Do đó, mô hình lai RNN+LSTM là mô hình nên được ưu tiên sử dụng để dự báo GDP bình quân đầu người trong tương lai.



Hình 5. Kết quả dự báo GDP bình quân đầu người trên tập kiểm tra

Mặc dù các kết quả nghiên cứu cho thấy rằng mô hình RNN-LSTM lai đã đem lại tối ưu hóa tính toán hơn trong quá trình huấn luyện so với ba mô hình còn lại. Tuy nhiên, mô hình RNN-LSTM lai đề xuất có một số hạn chế. Thứ nhất, hiệu suất dự đoán của mô hình phụ thuộc mạnh vào chất lượng, độ liên quan và khả năng sẵn có của dữ liệu huấn luyện. Trong nghiên cứu này, mô hình được huấn luyện và xác thực cụ thể bằng cách sử dụng các tập dữ liệu về các chỉ báo kinh tế từ Việt Nam. Bối cảnh địa lý và kinh tế này có thể hạn chế tính khả thi của mô hình đối với các khu vực khác có cấu trúc kinh tế khác nhau, yêu cầu điều chỉnh dữ liệu đáng kể và huấn luyện lại để thích ứng hiệu quả với các môi trường kinh tế mới. Hơn nữa, mặc dù đã nỗ lực giảm thiểu lỗi thông qua tinh chỉnh siêu tham số, quá trình tối ưu hóa mô hình vẫn tồn tại kém tài nguyên. Việc tìm kiếm để chọn lựa siêu tham số là hữu ích nhưng tốn kém về mặt tính toán, vì các mô hình học sâu như RNN và LSTM thường cần nhiều tinh chỉnh để đạt được kết quả tối ưu. Việc tinh chỉnh thêm các siêu tham số trên nhiều tham số như lớp, nơ-ron, tỷ lệ dropout vẫn yêu cầu sức mạnh xử lý và thời gian đáng kể. Một hạn chế khác nằm ở những thách thức vốn có của RNN và LSTM, bao gồm khả năng biến mất gradient, điều này làm cho việc huấn luyện các mô hình sâu trên các chuỗi dài trở nên phức tạp.

5. Kết luận và khuyến nghị

GDP bình quân đầu người là yếu tố thiết yếu để đánh giá tình hình chung của nền kinh tế

địa phương, khu vực và toàn cầu. Nghiên cứu này đề xuất một mô hình mới để dự báo GDP bình quân đầu người dựa trên các chỉ số kinh tế vĩ mô Việt Nam, được gọi là Mạng nơ-ron hồi quy-Bộ nhớ dài hạn ngắn hạn (RNN-LSTM). Thành phần LSTM được sử dụng để giải quyết vấn đề độ dốc biến mất mà RNN tích chập gặp phải. Bên cạnh đó, các mô hình độc lập như CNN, RNN và LSTM đã được chọn làm mô hình so sánh để đánh giá độ chính xác và tính ổn định của mô hình RNN-LSTM. Nghiên cứu sử dụng dữ liệu chuỗi thời gian, có nguồn gốc từ cơ sở dữ liệu của Ngân hàng Thế giới cho quốc gia Việt Nam. Dữ liệu đã được tiền xử lý và chuẩn hóa trước khi phân chia tập huấn luyện và tập kiểm tra.

Dựa trên kết quả nghiên cứu, tác giả có thể rút ra những kết luận sau: Mô hình lai RNN-LSTM được đề xuất thể hiện khả năng khái quát hóa mạnh mẽ, đưa ra dự đoán ổn định về GDP bình quân đầu người dựa trên các biến số vĩ mô nền kinh tế. Hiệu suất của mô hình này vượt trội hơn đáng kể so với các mô hình CNN, RNN và LSTM. So với ba mô hình khác, mô hình lai thể hiện độ chính xác và độ ổn định dự đoán vượt trội, đạt giá trị R2 cao nhất là 91,29%. Tuy nhiên, nghiên cứu thừa nhận những hạn chế tiềm ẩn do chất lượng dữ liệu nguồn gây ra. Mặc dù có nguồn từ các tổ chức đáng tin cậy, nhưng các tập dữ liệu kinh tế vốn bao gồm một thành phần ngẫu nhiên và nhiều tiềm ẩn. Để giải quyết vấn đề này, các bước xử lý trước và các kỹ thuật đánh giá

manh mẽ được sử dụng trong nghiên cứu này nhằm đảm bảo độ tin cậy của các dự đoán.

Mặc dù nghiên cứu chứng minh hiệu quả của mô hình lai RNN-LSTM trong việc dự đoán tăng trưởng GDP với độ chính xác cao, nhưng điều cần thiết là phải nhận ra rằng việc đánh giá mô hình không chỉ giới hạn ở độ chính xác. Nghiên cứu trong tương lai cũng nên ưu tiên tính khả giải thích và tính mạnh mẽ của các mô hình học máy, vì những yếu tố này đóng vai trò quan trọng trong việc đảm

Tài liệu tham khảo

- [1] T. De Cola and M. Mongelli, "Adaptive Time Window Linear Regression for Outage Prediction in Q/V Band Satellite Systems," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 808–811, Oct. 2018, doi: 10.1109/LWC.2018.2826543.
- [2] M. Y. Shams, Z. Tarek, E.-S. M. El-kenawy, M. M. Eid, and A. M. Elshewey, "Predicting Gross Domestic Product (GDP) using a PC-LSTM-RNN model in urban profiling areas," *Comput. Urban Sci.*, vol. 4, no. 1, p. 3, Jan. 2024, doi: 10.1007/s43762-024-00116-2.
- [3] X. Wu, Z. Zhang, H. Chang, and Q. Huang, "A Data-Driven Gross Domestic Product Forecasting Model Based on Multi-Indicator Assessment," *IEEE Access*, vol. 9, pp. 99495–99503, 2021, doi: 10.1109/ACCESS.2021.3062671.
- [4] G. Cicceri, G. Inserra, and M. Limosani, "A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study," *Mathematics*, vol. 8, no. 2, Art. no. 2, Feb. 2020, doi: 10.3390/math8020241.
- [5] Q. Gu, Q. Dai, H. Yu, and R. Ye, "Integrating Multi-Source Transfer Learning, Active Learning and Metric Learning paradigms for Time Series Prediction," *Appl. Soft Comput.*, vol. 109, p. 107583, Sep. 2021, doi: 10.1016/j.asoc.2021.107583.
- [6] H. Han *et al.*, "Time series forecasting model for non-stationary series pattern extraction using deep learning and GARCH modeling," *J. Cloud Comput.*, vol. 13, no. 1, p. 2, Jan. 2024, doi: 10.1186/s13677-023-00576-7.
- [7] Li J., Pan S., and Huang L., "A Machine Learning Based Method for Customer Behavior Prediction," *Teh. Vjesn. - Tech. Gaz.*, vol. 26, no. 6, Nov. 2019, doi: 10.17559/TV-20190603165825.
- [8] S. H. Faru, A. Waititu, and L. Nderu, "A Hybrid Neural Network Model Based on Transfer Learning for Forecasting Forex Market," *J. Data Anal. Inf. Process.*, vol. 11, no. 2, Art. no. 2, Mar. 2023, doi: 10.4236/jdaip.2023.112007.
- [9] M. S. Islam, E. Hossain, A. Rahman, M. S. Hossain, and K. Andersson, "A Review on Recent Advancements in FOREX Currency Prediction," *Algorithms*, vol. 13, no. 8, Art. no. 8, Aug. 2020, doi: 10.3390/a13080186.
- [10] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: a gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, doi: 10.1016/j.jclinepi.2006.01.014.
- [11] S. Van Buuren, Brand, J. P.L., Groothuis-Oudshoorn, C. G.M., and D. B. and Rubin, "Fully conditional specification in multivariate imputation," *J. Stat. Comput. Simul.*, vol. 76, no. 12, pp. 1049–1064, Dec. 2006, doi: 10.1080/10629360600810434.
- [12] P. Giudici, E. Raffinetti, and G. Toscani, "Measuring multidimensional inequality: a new proposal based on the Fourier transform," Jan. 25, 2024, *arXiv: arXiv:2401.14012*. doi: 10.48550/arXiv.2401.14012.

Ngày nhận bài: 24/4/2025

Ngày hoàn thành sửa bài: 17/5/2025

Ngày chấp nhận đăng: 04/6/2025