Tạp chí Kinh tế và Dự báo
Economy & Forecast Review Online
VIỆN CHIẾN LƯỢC PHÁT TRIỂN - BỘ KẾ HOẠCH VÀ ĐẦU TƯ

e-ISSN: 2734-9365

Dự báo chỉ số chứng khoán bằng học máy: Bằng chứng thực nghiệm từ thị trường chứng khoán Việt Nam

14:24 | 24/06/2024

EFR ***Nghiên cứu đánh giá hiệu quả của các mô hình học máy trong việc dự đoán biến động của chỉ số VNIndex. Từ đó, cung cấp công cụ hữu ích cho các nhà đầu tư và các nhà quản lý quỹ trong việc ra quyết định đầu tư trên thị trường chứng khoán Việt Nam.***

Đào Lê Kiều Oanh, Nguyễn Thị Minh Châu

Trường Đại học Ngân hàng TP. Hồ Chí Minh

Tóm tắt

Nghiên cứu đánh giá hiệu quả của các mô hình học máy trong việc dự đoán biến động của chỉ số VNIndex. Kết quả nghiên cứu cho thấy, phương pháp mạng tích chập thời gian (Temporal Convolutional Networks - TCN) và mạng bộ nhớ dài ngắn (Long Short - Term Memory - LSTM) có khả năng dự báo biến động chỉ số VNIndex với độ chính xác cao, trong đó LSTM thể hiện có hiệu quả dự báo tốt hơn. Phát hiện từ nghiên cứu này không chỉ góp phần vào lý thuyết dự báo tài chính mà còn cung cấp công cụ hữu ích cho các nhà đầu tư và các nhà quản lý quỹ trong việc ra quyết định đầu tư trên thị trường chứng khoán Việt Nam.

Từ khóa: *dự báo, chỉ số chứng khoán, học sâu*

Summary

This research evaluates the effectiveness of machine learning models in predicting VNIndex fluctuations. Research results show that the method of temporal convolutional networks (TCN) and long short–term memory networks (LSTM) are capable of predicting VNIndex fluctuations with high accuracy, in which LSTM shows better forecasting performance. Findings from this study not only contribute to financial forecasting theory but also provide useful tools for investors and fund managers in making investment decisions in the Vietnamese stock market.

Keywords: *forecasting, stock index, deep learning*

GIỚI THIỆU

Trong bối cảnh nền kinh tế toàn cầu ngày càng biến động và phức tạp, việc cung cấp dự báo chỉ số chứng khoán có tính chính xác cao trở thành một yếu tố quan trọng giúp các nhà đầu tư đưa ra quyết định đúng đắn (Chính và Hoàng, 2009).

Tại Việt Nam, chủ đề nghiên cứu về thị trường chứng khoán, đặc biệt là biến động chỉ số thị trường, luôn thu hút sự quan tâm của các nhà đầu tư trên thị trường. Việc dự báo biến động chỉ số chứng khoán trên thị trường Việt Nam vẫn đối mặt với nhiều thách thức do đặc thù của thị trường mới nổi, bao gồm: sự biến động cao và ảnh hưởng của các yếu tố kinh tế không ổn định. Sự biến động bất thường và đột ngột của thị trường chứng khoán Việt Nam mang lại không ít rủi ro cho các nhà đầu tư, cũng như sự phát triển bền vững của thị trường chứng khoán

Trên thực tế, có nhiều nghiên cứu, như: Nguyễn Hồ Diệu Uyên và Nguyễn Thị Thanh Huyền (2014), Dương Ngân Hà (2018)... đã kiểm định hiệu suất mô hình phức hợp LSTM-GRU thông qua dự báo biến động chỉ số chứng khoán, cho thấy đây là một phương pháp hứa hẹn trong việc dự báo chỉ số chứng khoán của Việt Nam. Tuy nhiên, trong phạm vi khảo lược chưa tìm thấy nghiên cứu so sánh phương pháp TCN, LSTM trong dự báo chỉ số chứng khoán, mặc dù đây được chứng minh là 2 phương pháp mang lại hiệu quả cao trong dự báo và được sử dụng ngày càng nhiều trong các nghiên cứu dự báo đối với dữ liệu thời gian. Vì vậy, bài viết được thực hiện nhằm mục đích lấp đầy khoảng trống nghiên cứu này. Bằng cách áp dụng các mô hình học máy tiên tiến vào dữ liệu của thị trường Việt Nam, nghiên cứu không chỉ đánh giá khả năng dự báo của các mô hình này để phục vụ cho việc dự báo ở các mô hình phức tạp hơn của các chủ thể ra quyết định trên thị trường chứng khoán.

CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Cơ sở lý thuyết

Dự báo biến động chỉ số chứng khoán được xây dựng dựa trên dự báo chỉ số chứng khoán tương lai. Các dự báo giá có giá trị đáng kể đối với nhà đầu tư, các nhà giao dịch và các cơ quan tài chính trong việc ra quyết định thông minh, quản lý rủi ro và cải thiện chiến lược đầu tư (Cavalcante và cộng sự, 2016; Tang và cộng sự, 2022). Tuy nhiên, việc dự đoán chính xác giá

của chuỗi thời gian tài chính thường gặp khó khăn do độ phức tạp và sự bất định của dữ liệu. Theo giả thuyết thị trường hiệu quả mà Fama (1970) đưa ra, giá tài sản phản ánh mọi thông tin có sẵn. Điều này ngụ ý rằng, việc phát triển các chiến lược giao dịch có lợi nhuận thường không kịp thích ứng với sự biến động của giá, dẫn đến những điều chỉnh giá trước khi chiến lược kịp được áp dụng.

Ngày càng nhiều các nghiên cứu sử dụng học máy trong dự báo biến động chỉ số giá tài sản tài chính nói chung và chỉ số thị trường chứng khoán nói riêng (Kumbure và cộng sự, 2022). Các công trình nghiên cứu tổng quan của Zhang và cộng sự (2024), Rouf và cộng sự (2021) đều cho thấy, đây là chủ đề hấp dẫn với ngày càng nhiều các phương pháp học máy được sử dụng nhằm có thể đưa ra những dự báo tốt hơn, phục vụ cho các chủ thể ra quyết định. Theo thống kê dựa trên phần mềm Perish chuyên dùng để khảo lược, tổng số lượng công trình là 150 được thực hiện với nhiều mẫu dữ liệu và phương pháp học máy khác nhau.

Xét riêng tại Việt Nam, số lượng nghiên cứu cùng chủ đề còn khá khiêm tốn, nhưng phương pháp nghiên cứu khá đa dạng. Nguyễn Hồ Diệu Uyên và Nguyễn Thị Thanh Huyền (2014) sử dụng phương pháp ARIMA để dự báo thay đổi của chỉ số chứng khoán trên sàn chứng khoán TP. Hồ Chí Minh. Dương Ngân Hà (2018) đã sử dụng phương pháp tự hồi quy vector Var để dự báo biến động chỉ số VNIndex qua khối lượng giao dịch ròng và giá trị giao dịch ròng của nhà đầu tư nước ngoài. Trương Thị Thùy Dương (2023) sử dụng phương pháp XGBoost để dự báo chiều biến động của chỉ số chứng khoán. Trần Đăng Tuyên (2024) đã kiểm định hiệu suất mô hình phức hợp LSTM-GRU thông qua dự báo biến động chỉ số chứng khoán, cho thấy đây là một phương pháp hứa hẹn trong việc dự báo chỉ số chứng khoán của Việt Nam. Dự báo giá trong chuỗi thời gian tài chính đa dạng, nhưng mỗi phương pháp lại có những giới hạn riêng. Phân tích cơ bản, dựa trên kiến thức chuyên ngành để lọc thông tin từ dữ liệu, đã ít được ưa chuộng, vì hiệu quả thấp (Rouf và cộng sự, 2021). Theo Cheng và cộng sự (2015), các phương pháp thống kê, như: ARIMA và GARCH, mặc dù dựa trên cấu trúc toán học và giả định thống kê để mô hình hóa, lại thường bỏ lỡ mẫu phi tuyến tính hoặc mối quan hệ phức tạp trong dữ liệu thực tế. Các kỹ thuật học máy, như: mạng nơ-ron nhân tạo (ANN) và hồi quy vector hỗ trợ (SVR) có khả năng tự động nhận diện mẫu ẩn, nhưng lại phụ thuộc vào kỹ thuật xử lý dữ liệu và bị hạn chế bởi độ phức tạp của mô hình khi có nhiều dữ liệu đào tạo (Jiang, 2021). Trong khi đó, kỹ thuật học sâu đã thể hiện sự ảnh hưởng mạnh mẽ, làm thay đổi cách tiếp cận trong lĩnh vực này. Các mô hình dựa trên học sâu có khả năng tự động học hỏi và thích ứng với mẫu phức tạp, ít phụ thuộc vào kiến thức chuyên môn và tận dụng lợi thế từ dữ liệu đào tạo dồi dào (Goodfellow và cộng sự, 2016, Hua và cộng sự, 2019).

Vì vậy, nghiên cứu được thực hiện nhằm bước đầu đánh giá hiệu quả của thuật toán học sâu TCN, LSTM trong dự báo chỉ số chứng khoán với trường hợp nghiên cứu tại thị trường chứng khoán Việt Nam dựa trên chuỗi thời gian đơn biến của chỉ số VNIndex.

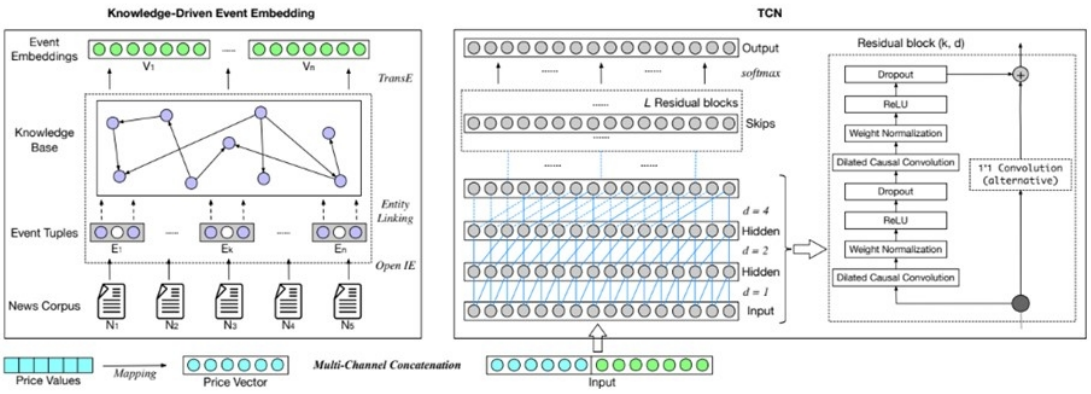
Phương pháp nghiên cứu

Dữ liệu sử dụng trong nghiên cứu là giá đóng cửa hàng ngày của chỉ số VNIndex từ năm 2001 đến năm 2023, được thu thập từ Vietstock. Chuỗi thời gian tài chính đơn biến phổ biến nhất được ngành tài chính phân tích là chuỗi giá đóng cửa hàng ngày (Zhang và cộng sự, 2024).

TCN là một trong những phương pháp hiện đại và tiên tiến trong việc xử lý dữ liệu chuỗi thời gian. TCN là một loại mạng tích chập có thiết kế cụ thể giúp chúng phù hợp để xử lý chuỗi thời gian. TCN đáp ứng 2 nguyên tắc chính: đầu ra của mạng có cùng độ dài với chuỗi đầu vào; và chúng ngăn chặn sự rò rỉ thông tin từ tương lai về quá khứ bằng cách sử dụng các tích chập nhân quả (Bai và cộng sự, 2018). Tích chập nhân quả khác với tích chập tiêu chuẩn ở thực tế là phép toán tích chập được thực hiện để thu được đầu ra tại thời điểm t không có tương lai các giá trị làm đầu vào. TCN sử dụng các lớp tích chập (convolutional layers) với độ sâu và số lượng filter khác nhau để nắm bắt các mẫu và xu hướng trong dữ liệu chuỗi thời gian (Liu và cộng sự, 2019). TCN có khả năng xử lý dữ liệu nhanh hơn, ổn định hơn trong việc huấn luyện và có khả năng học các mẫu dài hạn một cách hiệu quả. TCN ít bị ảnh hưởng bởi sự xáo trộn dữ liệu và có khả năng tổng quát hóa tốt hơn trong các bài toán chuỗi thời gian. TCN sử dụng sự tích chập nhân quả giãn nở để có thể nắm bắt được sự phụ thuộc lâu dài hơn và ngăn ngừa mất thông tin (Thill và cộng sự, 2020).

Trong khi đó, LSTM (Long-short term memory) là một loại mạng nơ-ron hồi quy sâu, được thiết kế đặc biệt để xử lý và dự báo chuỗi thời gian. Với cơ chế nhớ dài hạn và ngắn hạn, LSTM có khả năng lưu giữ thông tin qua nhiều bước thời gian, giúp dự báo những chuỗi có quan hệ phức tạp. Nghiên cứu của Chen và cộng sự (2024), thông qua dự báo biến động giá cổ phiếu trên thị trường chứng khoán Hồng Kông, đã củng cố thêm bằng chứng cho thấy, phương pháp TCN vượt trội hơn tất cả các mô hình khác được so sánh.

Hình 1: Minh họa kỹ thuật mạng tích chập thời gian



Nguồn: Deng và cộng sự (2019)

Để thực hiện nghiên cứu bằng phương pháp học máy, quy trình nghiên cứu được mô tả qua Hình 1. Theo đó, để huấn luyện một mô hình học sâu có thể dự đoán một số bước thời gian, giai đoạn tiền xử lý là cần thiết để chuyển đổi dữ liệu chuỗi thời gian ban đầu. Đầu tiên, dữ liệu được chuẩn hóa tối thiểu-tối đa (MinMax) để chia tỷ lệ các giá trị từ 0 đến 1, giúp cải thiện khả năng hội tụ của các mạng sâu, sau đó, chuyển đổi chuỗi thành các trường có thể được sử dụng để cung cấp dữ liệu mạng.

Đánh giá hiệu suất dự báo của mô hình, nghiên cứu sử dụng các chỉ số sai số phần trăm tuyệt đối có trọng số (WAPE), mức độ sai số trung bình (MAE), căn bậc hai của sai số bình phương trung bình (RMSE). WAPE được tính bằng tổng sai số tuyệt đối chia cho tổng giá trị thực tế, sau đó nhân với 100 để chuyển thành tỷ lệ phần trăm. WAPE giúp đánh giá sai số dự báo so với tổng giá trị thực tế, cho thấy mô hình dự báo chính xác đến mức nào theo tỷ lệ phần trăm. WAPE càng thấp, mô hình dự báo càng tốt. MAE cho biết mức độ sai số trung bình mà mô hình dự báo tạo ra, không tính đến hướng của sai số. MAE càng thấp, mô hình dự báo càng chính xác. RMSE là căn bậc hai của sai số bình phương trung bình, là một thước đo phổ biến để đánh giá độ chính xác của mô hình dự báo. RMSE nhấn mạnh hơn vào các sai số lớn vì các sai số này được bình phương trước khi tính trung bình. RMSE càng thấp, mô hình dự báo càng tốt.

KẾT QUẢ NGHIÊN CỨU

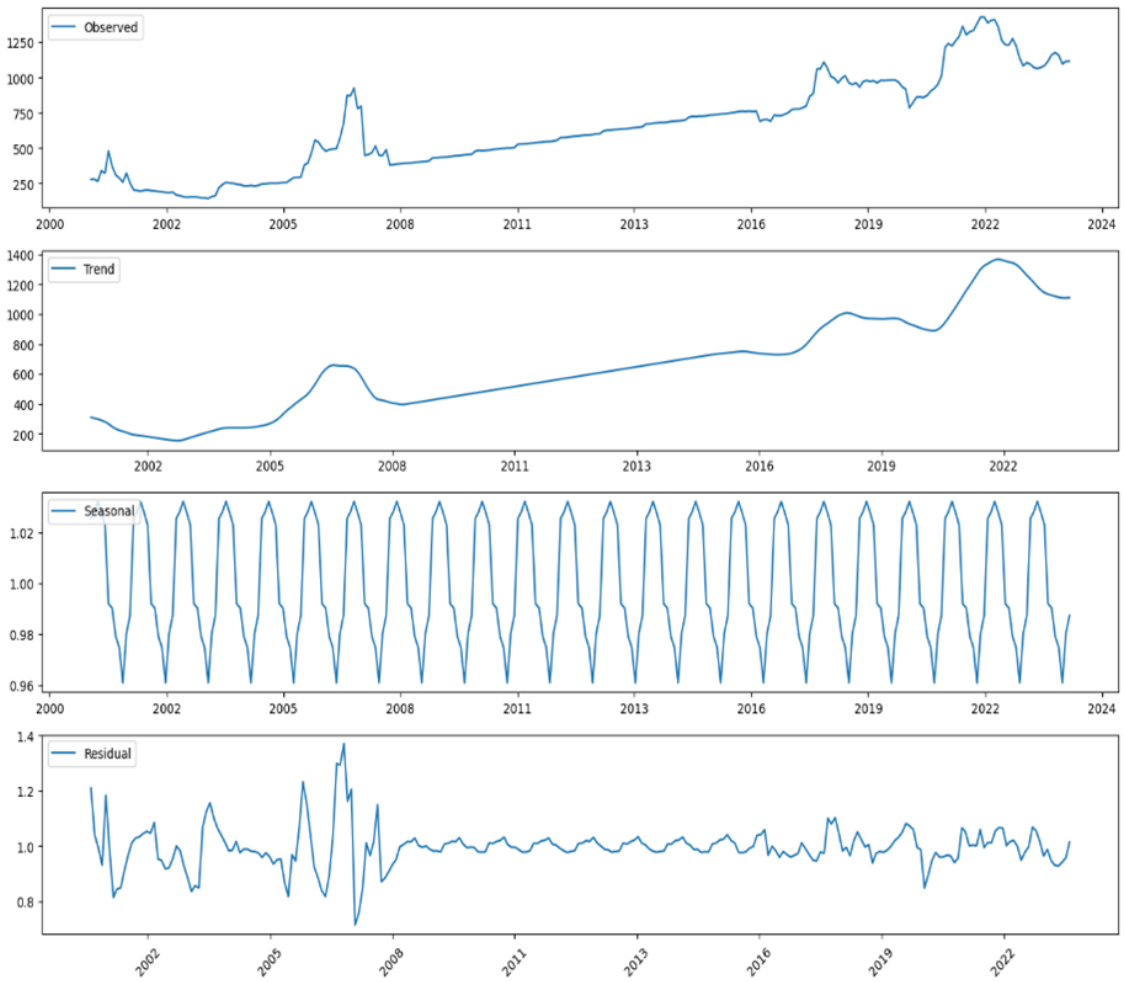
Dữ liệu gốc được thu thập trong giai đoạn nghiên cứu được thể hiện qua Hình 2 cho thấy, chỉ số VNIndex có biến động trong giai đoạn nghiên cứu.

Hình 2: Giá đóng cửa trung bình theo tháng của VNIndex trong giai đoạn 2001-2023

Nguồn: Tổng hợp của nhóm tác giả

Nghiên cứu thực hiện phân rã để thấy được các thành phần của dữ liệu chuỗi thời gian. Kết quả Hình 3 cho thấy, dữ liệu có tính xu hướng, mùa vụ - điển hình của chuỗi dữ liệu thời gian trong các nghiên cứu định lượng.

Hình 3: Phân tích thành phần dữ liệu chuỗi thời gian VNIndex



Nguồn: Tổng hợp của nhóm tác giả

Sau khi đánh giá dữ liệu, nghiên cứu thực hiện tiền xử lý dữ liệu và chuẩn hóa dữ liệu theo tối thiểu tối đa. Bộ dữ liệu chuẩn hóa được chia thành hai tập dữ liệu, bao gồm: Tập huấn luyện (80%) và Tập kiểm tra (20%). Các mô hình học sâu được xây dựng và huấn luyện trên tập dữ liệu này. Nghiên cứu sử dụng thuật toán TCN với 2 mô hình. Trong đó, thuật toán TCN gốc có các thông số gồm 2 lớp Conv1D với 64 filter, lớp Dense với 50 nút. Mô hình tối ưu hóa được cải thiện với việc tăng thêm các lớp kernel và sử dụng Dropout để cải thiện hiệu suất mô hình. Mô hình LSTM gốc có thông số cơ bản gồm: một lớp LSTM với 50 đơn vị, một lớp GRU với 50 đơn vị. Mô hình tối ưu hóa tăng gấp đôi đơn vị trong cả hai lớp của mô hình.

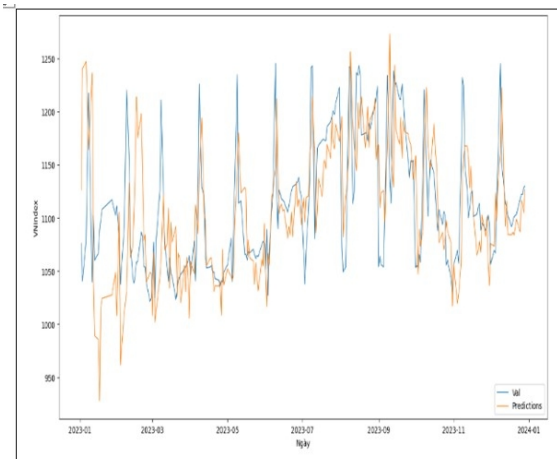
Kết quả đánh giá mô hình dựa trên 3 chỉ số MAE, RMSE và WAPE của các mô hình (Bảng) cho thấy, tính hiệu quả trong dự báo của hai thuật toán này trong dự báo chỉ số VNIndex. Trong đó, các chỉ số ở thuật toán LSTM thấp hơn so với TCN, chứng tỏ LSTM ít có sai số hơn, đồng nghĩa hiệu quả tốt hơn trong nhiệm vụ dự báo. Phát hiện này của bài viết trái ngược với kết quả nghiên cứu của Chen và cộng sự (2024), Zhang và cộng sự (2022) khi các nghiên cứu này cho thấy, TCN ưu việt hơn so với LSTM.

Bảng: Kết quả chỉ số đánh giá mô hình

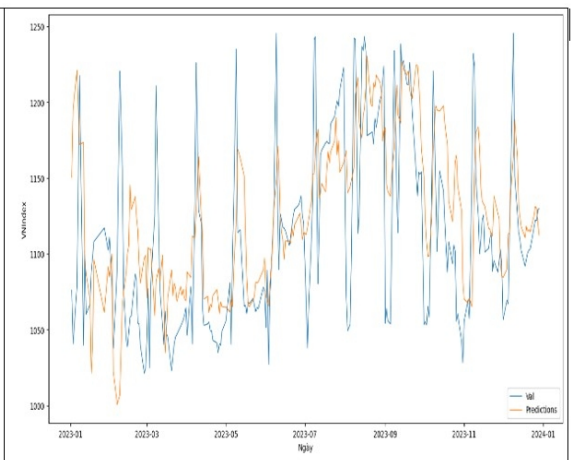
Phương pháp	Mô hình	MAE	RMSE	WAPE
TCN	Cơ bản	65.70	90.28	0.058
	Cải thiện	56.48	70.49	0.050
LSTM-GRU	Cơ bản	65.19	86.27	0.057
	Cải thiện	47.05	71.52	0.041

Nguồn: Tổng hợp từ kết quả nghiên cứu

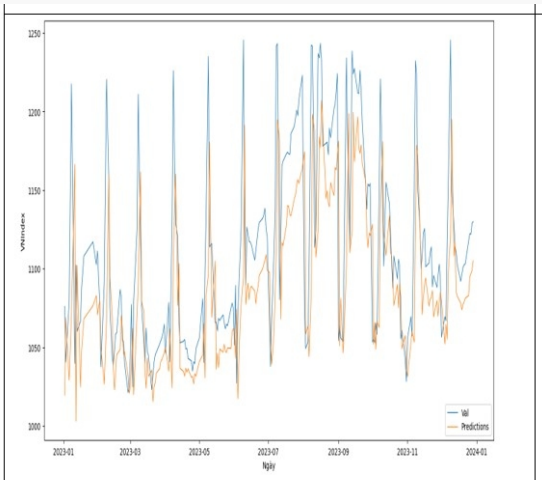
Minh họa khả năng dự báo của các mô hình được thể hiện lần lượt từ Hình 4 đến Hình 7. Các chuỗi dữ liệu dự báo trong cả 4 mô hình đều có độ tương đồng về diễn biến tăng giảm hay biến động của chỉ số VNIndex. Kết quả này cho thấy, khả năng của học sâu với 2 phương pháp LSTM và TCN trong việc dự báo biến động chỉ số chứng khoán chỉ với chuỗi dữ liệu đơn biến.



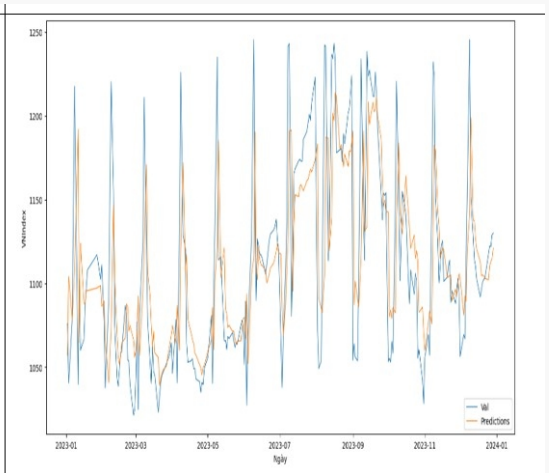
Hình 4: Kết quả dự báo của mô hình TCN cơ bản



Hình 5: Kết quả dự báo của mô hình TCN cải thiện



Hình 6: Kết quả dự báo của mô hình LSTM cơ bản



Hình 7: Kết quả dự báo của mô hình LSTM cải thiện

KẾT LUẬN

Nghiên cứu sử dụng 2 phương pháp đang được quan trong dự báo chuỗi thời gian là TCN và LSTM. Cả 2 thuật toán đều được huấn luyện và kiểm tra trên cùng một tập dữ liệu VNIndex. Các chỉ số đánh giá hiệu suất dự báo đều cho thấy, các chỉ số phản ánh sai số của LSTM thấp hơn đáng kể so với thuật toán còn lại, tức LSTM dự báo chính xác hơn. Kết luận này phù hợp với các nghiên cứu trước đó và mở ra hướng nghiên cứu trong tương lai về tối ưu hóa mô hình học sâu trong dự báo chỉ số chứng khoán nói riêng và dự báo chuỗi thời gian nói chung.

Bên cạnh kết quả đạt được, bài viết vẫn còn một số hạn chế, bao gồm: mới chỉ tập trung vào chuỗi dữ liệu đơn biến, trong khi đó, lý thuyết và các nghiên cứu thực nghiệm cho thấy dự báo với mô hình đa biến có tính chính xác cao hơn, đặc biệt khi được bổ sung các chỉ số phản ánh tâm lý nhà đầu tư, các chỉ báo kỹ thuật trên thị trường. Ngoài ra, công trình chủ yếu so sánh giữa hai phương pháp học sâu cơ bản là TCN và LSTM mà chưa mở rộng ra so sánh với các phương pháp khác trong dự báo. Dựa trên những hạn chế nêu trên, các công trình nghiên cứu sau có thể được mở rộng nhằm đánh giá hiệu quả của các phương pháp học sâu trong dự báo chỉ số VNIndex, cung cấp thêm công cụ hỗ trợ các nhà đầu tư ra quyết định./.

TÀI LIỆU THAM KHẢO

1. Bai, S.; Kolter, J.Z.; Koltun, V (2018), An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, ***arXiv***, arXiv:1803.01271.

2. Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., and Oliveira, A. L. I. (2016), Computational Intelligence and Financial Markets: A Survey and Future Directions, ***Expert Systems with Applications,*** 55, 194-211.

3. Chen, S., Guo, L., and Ge, L. (2024), Increasing the Hong Kong Stock Market Predictability: A Temporal Convolutional Network Approach, ***Computational Economics***, 1-26, DOI:10.1007/s10614-024-10547-y.

4. Cheng, C., Sa-Ngasoongsong, A., Beyca, O., Le, T., Yang, H., Kong, Z., and Bukkapatnam, S. T. S. (2015), Time series forecasting for nonlinear and non-stationary processes: a review and comparative study, ***IIE Transactions***, 47(10), 1053-1071.

5. Chính, P. M., & Hoàng, V. Q. (2009), ***Kinh tế Việt Nam: Thăng trầm và đột phá,*** Nxb Chính trị Quốc gia.

6. Dương Ngân Hà (2018), Dự báo biến động của chỉ số VN-Index thông qua khối lượng giao dịch ròng và giá trị giao dịch ròng của nhà đầu tư nước ngoài, ***Tạp chí Khoa học& Đào tạo Ngân hàng,*** số 195, 18-25.

7. Fama, E. F. (1970), Efficient capital markets, ***Journal of finance***, 25(2), 383-417.

8. Goodfellow, I., Bengio, Y., and Courville, A. (2016), ***Deep learning: MIT press***, ISBN: 0262035618.

9. Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., and Zhang, H. (2019), Deep learning with long short-term memory for time series prediction. ***IEEE Communications Magazine***, 57(6), 114-119.

10. Kumbure, M. M., Lohrmann, C., Luukka, P., and Porras, J. (2022), Machine learning techniques and data for stock market forecasting: A literature review, ***Expert Systems with Applications***, 197.

11. Liu, Y., Dong, H., Wang, X., and Han, S. (2019), ***Time series prediction based on temporal convolutional network,*** In 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS***)*** (300-305), IEEE.

12. Jiang, W. (2021), Applications of deep learning in stock market prediction: recent progress, ***Expert Systems with Applications***, 184.

13. Nguyễn Hồ Diệu Uyên và Nguyễn Thị Thanh Huyền (2014), Ứng dụng mô hình ARIMA trong dự báo chỉ số VN-Index, ***Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng,*** số 12(85), 90-95

14. Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., and Kim, H. C. (2021), Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions, ***Electronics***, 10(21).

15. Tang, Q., Shi, R., Fan, T., Ma, Y., and Huang, J. (2021), Prediction of Financial Time Series Based on LSTM Using Wavelet Transform and Singular Spectrum Analysis, ***Mathematical Problems in Engineering,*** 1-13, DOI:10.1155/2021/9942410.

16. Thill, M., Konen, W., and Bäck, T. (2020), ***Time series encodings with temporal convolutional networks,*** In International Conference on Bioinspired Methods and Their Applications (161-173), Cham: Springer International Publishing.

17. Trần Đăng Tuyên (2024), Đánh giá hiệu suất mô hình phức hợp LSTM-GRU: nghiên cứu điển hình về dự báo chỉ số đo lường xu hướng biến động giá cổ phiếu trên Sàn Giao dịch chứng khoán TP. Hồ Chí Minh, ***Tạp chí Khoa học Đại học cần Thơ***, ***60***(1).

18. Trương Thị Thùy Dương (2023), Dự báo chiều biến động của chỉ số chứng khoán bằng thuật toán tăng cường, ***Tạp chí Khoa học và Đào tạo ngân hàng, số 252, 40-46.***

19. Zhang, C., Sjarif, N. N. A., and Ibrahim, R. (2024), Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022, ***Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery***, 14(1).

20. Zhang, C. X., Li, J., Huang, X. F., Zhang, J. S., and Huang, H. C. (2022), Forecasting stock volatility and value-at-risk based on temporal convolutional networks, ***Expert Systems with Applications***, 207.

21. Wan, R., Tian, C., Zhang, W., Deng, W., and Yang, F. (2022), A multivariate temporal convolutional attention network for time-series forecasting, ***Electronics***, 11(10).

Ngày nhận bài: 06/6/2024; Ngày phản biện: 14/6/2024; Ngày duyệt đăng: 24/6/2024

URL: https://kinhtevadubao.vn/du-bao-chi-so-chung-khoan-bang-hoc-may-bang-chung-thuc-nghiem-tu-thi-truong-chung-khoan-viet-nam-29030.html