TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT, ĐHQGHCM

KHOA HỆ THỐNG THÔNG TIN

ĐỒ ÁN CUỐI KỲ

PHÂN TÍCH DỮ LIỆU VỚI R/PYTHON

ĐỀ TÀI: ỨNG DỤNG LSTM VÀ GRU TRONG DỰ BÁO GIÁ CỔ PHIẾU NGẮN HẠN

Giảng viên hướng dẫn: ThS. Nguyễn Phát Đạt

Group 1:

1. **Trần Sĩ Đan - K214061258**
2. **Nguyễn Thiên Huy - K214060396**
3. **Trịnh Quốc Thịnh - K214060412**
4. **Dương Văn Nhựt Duy - K214060391**
5. **Giả Hoàng Nam Phương - K214061744**

Thành phố Hồ Chí Minh, tháng Năm, 2024

DANH SÁCH THÀNH VIÊN

STT	Họ và tên	MSSV	Nhiệm vụ	Đánh giá
1	Trần Sĩ Đan	K214061258	Nhóm trưởng	10/10
2	Nguyễn Thiên Huy	K214060396	Thành viên	10/10
3	Trịnh Quốc Thịnh	K214060412	Thành viên	10/10
4	Dương Văn Nhựt Duy	K214060391	Thành viên	10/10
5	Giả Hoàng Nam Phương	K214061744	Thành viên	10/10

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin cảm ơn trường Đại học Kinh tế - Luật - Đại học Quốc gia Thành phố Hồ Chí Minh đã đưa môn học *Phân tích dữ liệu với R/Python* vào chương trình giảng dạy.

Đặc biệt, để hoàn thành bài báo cáo này, chúng em xin được gửi lời cảm ơn sâu sắc đến thầy Nguyễn Phát Đạt - giảng viên hiện đang giảng dạy môn *Phân tích dữ liệu với R/Python* lớp chúng em. Những bài giảng cung cấp kiến thức đầy bổ ích và thú vị, cùng với sự nhiệt tình của thầy trong quá trình giảng dạy là nguồn cảm hứng to lớn đã hỗ trợ chúng em hoàn thành bài báo cáo này.

Kiến thức là vô hạn mà sự tiếp nhận kiến thức của bản thân mỗi người luôn tồn tại những hạn chế nhất định. Do đó, trong quá trình làm bài báo cáo, chúng em sẽ khó có thể tránh khỏi những thiếu sót. Nhóm chúng em rất mong nhận được sự góp ý từ thầy để nhóm có thể hoàn thiện hơn nữa. Điều này có ý nghĩa rất to lớn đối với nhóm chúng em.

Kính chúc thầy sức khỏe và gặt hái được nhiều thành công hơn nữa trong sự nghiệp giảng dạy.

Chúng em xin chân thành cảm ơn!

Nhóm thực hiện

Nhóm 1

CAM KẾT

Chúng em xin cam kết rằng kết quả dưới đây hoàn toàn là sự vận dụng hiểu biết của chúng em trên cơ sở kiến thức đã được giảng dạy từ bộ môn *Phân tích dữ liệu với R/Python* của thầy Nguyễn Phát Đạt, kết hợp với tham khảo các nguồn tài liệu tham khảo từ sách, báo, và các phương tiện truyền thông khác.

Thành phố Hồ Chí Minh, tháng Năm năm 2024

Nhóm 1

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
TFT	Time-Frequency Transform
LSTM	Long short-term memory
GRU	Gated Recurrent Unit
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN ĐỀ TÀI

1.1. Lý do lựa chọn đề tài

Thị trường chứng khoán là một lĩnh vực phức tạp và biến động, đòi hỏi sự hiểu biết sâu rộng và các công cụ dự đoán chính xác để tối ưu hóa lợi nhuận và giảm thiểu rủi ro. Trong bối cảnh này, ứng dụng học máy để dự đoán giá cổ phiếu đã trở thành một xu hướng phổ biến, nhờ vào khả năng xử lý dữ liệu lớn và nhận diện các mẫu phức tạp mà các phương pháp truyền thống khó có thể làm được.

Ứng dụng học máy trong dự đoán giá cổ phiếu đã được nghiên cứu rộng rãi. Tuy nhiên, các nghiên cứu như của Qiu, Wang và Zhou (2020) chỉ ra rằng, mặc dù các mô hình LSTM có thể đưa ra dự đoán chính xác trong ngắn hạn, nhưng việc dự đoán nhiều bước trong tương lai lại gặp nhiều khó khăn do lỗi dự đoán tích lũy qua từng bước, làm giảm độ chính xác tổng thể. Điều này cho thấy, dự báo dài hạn thường không thực tế và ít hiệu quả hơn so với dự báo ngắn hạn.

Thêm vào đó, các chỉ số kỹ thuật và yếu tố tiền tệ có ảnh hưởng đáng kể đến biến động giá cổ phiếu. Nghiên cứu của Yildirim, Toroslu và Fiore (2021) đã chứng minh rằng việc kết hợp các chỉ số kỹ thuật cùng với các yếu tố kinh tế vĩ mô như tỷ giá hối đoái có thể cải thiện độ chính xác của các mô hình dự đoán giá cổ phiếu. Nghiên cứu này nhấn mạnh rằng các chỉ số kỹ thuật và tỷ giá hối đoái có thể giúp nắm bắt các mẫu và xu hướng phức tạp, từ đó nâng cao khả năng dự đoán chính xác của mô hình.

Ngoài ra, các nghiên cứu cũng so sánh hiệu quả của LSTM và GRU trong việc dự báo giá cổ phiếu. Nghiên cứu của Gao, Wang và Zhou (2021) đã thực hiện các thí nghiệm so sánh hiệu suất của các mô hình LSTM và GRU dưới nhiều tham số khác nhau. Kết quả cho thấy cả hai mô hình đều có khả năng dự đoán giá cổ phiếu hiệu quả, và không có sự chênh lệch đáng kể về hiệu suất giữa hai mô hình. Hơn nữa, khi áp dụng các phương pháp giảm chiều như LASSO, cả hai mô hình đều cho kết quả dự đoán tốt hơn so với khi sử dụng PCA.

Dựa trên những phát hiện này, nhóm quyết định thực hiện đề tài "*Ứng dụng LSTM và GRU trong dự báo giá cổ phiếu ngắn hạn*". Việc kết hợp các biến chỉ báo kỹ thuật và tỷ giá hối đoái vào các mô hình LSTM và GRU không chỉ giúp nắm bắt được các yếu tố ảnh hưởng đến giá cổ phiếu mà còn nâng cao độ chính xác của các dự báo

ngắn hạn, hỗ trợ hiệu quả cho các nhà đầu tư trong việc đưa ra quyết định nhanh chóng và chính xác.

1.2. Mục tiêu của đề tài

Mục tiêu của nghiên cứu này là đề xuất và triển khai một phương pháp dựa trên LSTM và GRU để dự báo giá cổ phiếu trong tương lai gần. Trong bài báo này, nhóm xây dựng mô hình LSTM và GRU nhằm đưa ra dự báo hiệu quả và chính xác, đồng thời so sánh với các phương pháp khác đã có.

Trước khi bắt đầu lên ý tưởng cho mô hình và thực nghiệm, nhóm đã xác định các mục tiêu cụ thể bao gồm:

- Giới thiệu lý thuyết về dự báo chuỗi thời gian và mô hình LSTM cùng GRU, bao gồm các ý tưởng xây dựng của mô hình và cách chúng có thể áp dụng trong việc dự báo giá cổ phiếu trong ngắn hạn.

- Nghiên cứu các phương pháp và thuật toán LSTM, GRU trong lĩnh vực dự báo chuỗi thời gian, đặc biệt là trong việc xử lý dữ liệu chuỗi dài và phức tạp như giá cổ phiếu.

- Xây dựng và huấn luyện mô hình LSTM và GRU để dự báo giá cổ phiếu dựa trên lịch sử thị trường chứng khoán.

- Đánh giá hiệu suất của mô hình đề xuất thông qua các thử nghiệm, so sánh với các mô hình truyền thống và lựa chọn mô hình tốt nhất để đạt được dự báo chính xác và hiệu quả.

1.3. Đối tượng và phạm vi nghiên cứu

1.3.1. Đối tượng nghiên cứu

Giá cổ phiếu lịch sử của mã VNM, tức Công ty Cổ phần Sữa Việt Nam (tiếng Anh: Vietnam Dairy Products Joint Stock Company), thường được biết đến với thương hiệu Vinamilk.

1.3.2. Phạm vi nghiên cứu

(i) Về mặt lý thuyết:

- Khảo sát các phương pháp và thuật toán trong dự báo chuỗi thời gian, đặc biệt là trong lĩnh vực dự báo giá cổ phiếu ngắn hạn.
- Tìm hiểu cơ sở lý thuyết của mô hình LSTM và GRU trong việc dự đoán giá cổ phiếu.

- Nghiên cứu và thử nghiệm việc kết hợp mô hình LSTM và GRU để cải thiện hiệu suất dự đoán giá cổ phiếu ngắn hạn.
- Phát triển phương pháp và công cụ để đánh giá và so sánh hiệu suất của mô hình so với các phương pháp dự báo khác.

(ii) Về mặt thực nghiệm:

- Thực hiện huấn luyện các mô hình LSTM và GRU độc lập trên dữ liệu lịch sử của giá cổ phiếu ngắn hạn.
- Đánh giá và so sánh hiệu suất của các mô hình này nhằm xác định mô hình nào có khả năng dự đoán tốt hơn.
- Sử dụng các phép đo như MAE, RMSE, MAPE để đánh giá hiệu suất của các mô hình trên tập dữ liệu kiểm tra.
- Lựa chọn mô hình tốt nhất dựa trên kết quả đánh giá để áp dụng cho việc dự đoán giá cổ phiếu ngắn hạn trong thực tế.

1.4. Phương pháp nghiên cứu

Bài nghiên cứu dựa trên các phương pháp sau:

(i) Nghiên cứu lý thuyết:

- Phương pháp phân tích và tổng hợp lý thuyết: Tổng hợp và đưa ra luận điểm chính thu được từ quá trình phân tích, tổng hợp từ các lý thuyết nền tảng và các nghiên cứu trước.
- Phương pháp phân loại và hệ thống hóa lý thuyết: Dựa vào các thông tin thu thập được tiến hành hệ thống hóa và phân thành các mục vấn đề với hướng đi cụ thể, thống nhất, từ đó đưa ra kết luận cuối cùng.

(ii) Nghiên cứu thực nghiệm:

- Thu thập dữ liệu, thực hiện tiền xử lý và khám phá dữ liệu, đề xuất mô hình phân tích dự báo giá cổ phiếu ngắn hạn.

1.5. Kết cấu bài nghiên cứu

Bài nghiên cứu được chia thành cấu trúc 5 chương, chi tiết như sau:

Chương 1: Giới thiệu tổng quan đề tài

Trình bày tổng quát về bài nghiên cứu bao gồm mục tiêu chọn đề tài, mục tiêu nghiên cứu, đối tượng, phạm vi, phương pháp nghiên cứu và bố cục đề tài.

Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan

Bài báo cáo đưa ra các lý thuyết cơ bản, phương pháp được sử dụng trong bài báo cáo và các công trình nghiên cứu, đóng góp học thuật có liên quan.

Chương 3: Phương pháp nghiên cứu

Chương này tập trung vào việc trình bày phương pháp nghiên cứu được áp dụng trong bài nghiên cứu, bao gồm quy trình thực nghiệm, phương pháp lấy mẫu dữ liệu và các mô hình học sâu được sử dụng.

Chương 4: Kết quả thực nghiệm

Từ mô hình ở chương 3, tiến hành thực hiện mô tả, tiền xử lý dữ liệu, đánh giá hiệu suất của mô hình và tinh chỉnh các tham số (nếu cần thiết).

Chương 5: Kết luận và hướng phát triển

Từ hiệu quả nhận được, bài báo cáo rút ra những kết luận và đưa ra hướng đề xuất cho tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Cổ phiếu

2.1.1. Khái niệm

Trong lĩnh vực tài chính, cổ phiếu là một loại chứng quyền tài sản thể hiện sự sở hữu của nhà đầu tư trong một công ty. Chứng quyền này được biểu thị dưới dạng giấy tờ, cho phép chủ sở hữu tham gia vào quản trị và chia sẻ lợi nhuận của công ty. Cổ phiếu đương nhiên góp phần định giá của công ty và được giao dịch trên thị trường chứng khoán.

Cổ phiếu có một danh sách đặc điểm quan trọng mà nhà đầu tư nên hiểu rõ:

- Giá trị cổ phiếu: Giá trị của cổ phiếu phản ánh sự đánh giá của thị trường về tiềm năng tăng trưởng và lợi nhuận của công ty. Giá cổ phiếu phụ thuộc vào nhiều yếu tố, bao gồm kết quả kinh doanh, tình hình kinh tế và sự quan tâm của nhà đầu tư.

- Quyền lợi chủ sở hữu: Chủ sở hữu cổ phiếu được hưởng quyền lợi kinh tế và quyền tham gia vào quyết định quan trọng của công ty thông qua việc bỏ phiếu tại các cuộc họp cổ đông. Quyền lợi này bao gồm nhận cổ tức, quyền mua cổ phiếu thêm (nếu có), và quyền biểu quyết trong các quyết định quan trọng.

- Tính thanh khoản: Một yếu tố quan trọng để đánh giá tính thanh khoản của cổ phiếu là khả năng mua và bán cổ phiếu trên thị trường chứng khoán. Cổ phiếu với tính thanh khoản cao cho phép nhà đầu tư dễ dàng mua và bán cổ phiếu mà không gặp khó khăn đáng kể.

Hiện nay, đã có nhiều các mô hình phân tích được đưa ra nhằm dự đoán giá cổ phiếu. Tuy nhiên, các mô hình này có thể bị ảnh hưởng bởi các yếu tố như nhiễu dữ liệu và biến động không thể kiểm soát được trên thị trường khiến các nhà đầu tư gặp khó khăn trong việc đưa ra quyết định.

2.1.2. Đặc trưng

Các yếu tố đặc trưng của cổ phiếu ảnh hưởng đến biến động giá cổ phiếu đã được nghiên cứu và được chia thành hai nhóm chính là yếu tố vĩ mô và yếu tố nội tại doanh nghiệp. Tiêu biểu tại Việt Nam, Lộc, T. Đ. (2014) đã nghiên cứu về yếu tố vĩ mô tập trung vào việc đo lường ảnh hưởng của các yếu tố kinh tế, chính trị và xã hội đến giá

cổ phiếu. Trong khi đó, Chi, L. H. D. (2021) tập trung vào các yếu tố nội tại của doanh nghiệp, như hiệu suất tài chính, quản lý doanh nghiệp và các chỉ số tài chính khác.

Phạm vi của nghiên cứu này tập trung vào việc xác định sự ảnh hưởng của những yếu tố nội tại đến giá cổ phiếu của doanh nghiệp. Một số đặc trưng quan trọng có thể ảnh hưởng đến biến động giá cổ phiếu bao gồm:

- Tính thanh khoản: Mức độ thanh khoản của cổ phiếu có thể ảnh hưởng đến biến động giá cổ phiếu. Chi, L. H. D. (2021) đã chỉ ra rằng cổ phiếu có tính thanh khoản cao thường có biên độ giá thấp hơn, trong khi cổ phiếu có tính thanh khoản thấp có thể có biên độ giá cao hơn.

- Rủi ro: Nghiên cứu trên cũng cho thấy cổ phiếu có mức độ rủi ro cao thường có biên độ giá cao hơn, trong khi cổ phiếu có mức độ rủi ro thấp có thể có biên độ giá thấp hơn.

- Yếu tố cơ bản: Các yếu tố cơ bản của công ty có thể ảnh hưởng đến biến động giá cổ phiếu. Ví dụ, kết quả kinh doanh, tình hình tài chính, các sự kiện quan trọng như mua bán công ty, thay đổi lãnh đạo, hoặc thay đổi chiến lược kinh doanh có thể gây ra biến động giá cổ phiếu như được đề cập đến trong nghiên cứu của Lộc, T. Đ. (2014).

- Tác động của ngành công nghiệp: Tùy theo thị trường, lĩnh vực, các tin tức, chính sách, hoặc xu hướng trong ngành có thể tác động đến giá cổ phiếu của các công ty trong ngành đó.

- Tâm lý nhà đầu tư: Theo Chi, L. H. D. (2021), sự lạc quan hoặc bi quan của nhà đầu tư cũng có thể tác động đến quyết định mua bán cổ phiếu và góp phần tạo ra biến động trên thị trường.

Tuy nhiên, cần lưu ý rằng các yếu tố đặc trưng của cổ phiếu có thể tác động một cách phức tạp và không đồng nhất đến giá cổ phiếu trên thị trường. Các yếu tố này cần được nghiên cứu và đánh giá kỹ càng để hiểu rõ hơn về quan hệ giữa chúng và biến động giá cổ phiếu.

2.1.3. Ảnh hưởng của các chỉ số kỹ thuật đến giá cổ phiếu

Ảnh hưởng của các chỉ số kỹ thuật đến giá cổ phiếu trên thị trường chứng khoán là chủ đề được quan tâm trong nghiên cứu tài chính. Chỉ báo kỹ thuật là công cụ được các nhà giao dịch và nhà đầu tư sử dụng để phân tích dữ liệu lịch sử về giá và khối lượng nhằm dự đoán biến động giá trong tương lai và đưa ra quyết định đầu tư.

Một số nghiên cứu đã kiểm tra tính hiệu quả của các chỉ số kỹ thuật khác nhau trong việc dự đoán giá cổ phiếu. Ví dụ, một nghiên cứu của Tuyên (2024) tập trung vào việc đánh giá hiệu suất của mô hình LSTM-GRU phức tạp trong việc dự báo xu hướng giá cổ phiếu. Nghiên cứu sử dụng các chỉ báo kỹ thuật làm đầu vào cho mô hình và phân tích tác động của chúng đến biến động giá cổ phiếu.

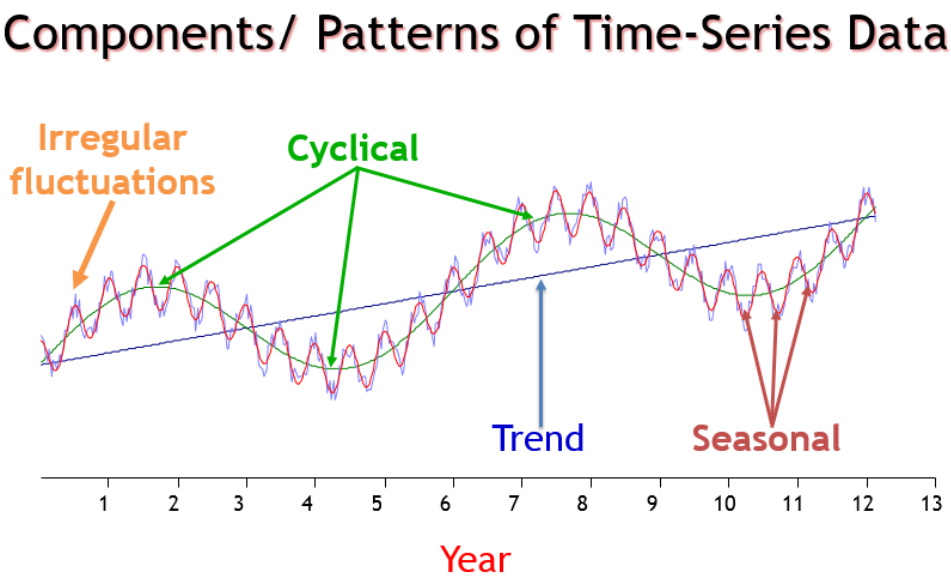
Một nghiên cứu khác của Nguyễn (2012) khám phá các yếu tố ảnh hưởng đến khả năng sinh lời của cổ phiếu niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh. Nghiên cứu xem các chỉ số kỹ thuật như VN-Index và giá đóng cửa của cổ phiếu là những yếu tố chính ảnh hưởng đến lợi nhuận cổ phiếu.

Phan (2022) nghiên cứu ảnh hưởng của các chỉ số tài chính đến quyết định của nhà đầu tư cá nhân trên thị trường chứng khoán. Nghiên cứu nhấn mạnh tầm quan trọng của phân tích kỹ thuật và nghiên cứu về các chỉ số kỹ thuật khác nhau để đưa ra quyết định đầu tư sáng suốt.

Nhìn chung, các nghiên cứu này cho thấy rằng các chỉ số kỹ thuật đóng một vai trò quan trọng trong việc hiểu và dự đoán biến động giá cổ phiếu trên thị trường chứng khoán. Tuy nhiên, điều quan trọng cần lưu ý là hiệu quả của các chỉ báo kỹ thuật có thể khác nhau tùy thuộc vào điều kiện thị trường cụ thể và cách giải thích các chỉ báo của các nhà đầu tư cá nhân.

2.2. Lý thuyết ứng dụng trong đề tài

2.2.1. Học sâu cho dự báo chuỗi thời gian



Hình 2.1. Time-series

Lim và các cộng sự (2021) đã trình bày một định nghĩa toàn diện về các mô hình dự báo chuỗi thời gian sử dụng kỹ thuật học sâu (deep learning). Theo đó, các mô hình này có mục tiêu dự đoán các giá trị tương lai của một biến mục tiêu yi,tcho một thực thể i tại thời điểm t. Mỗi thực thể biểu thị cho một tập hợp thông tin có tính tuần tự theo thời gian, chẳng hạn như các đo đạc từ các trạm thời tiết khác nhau trong lĩnh vực khí hậu học hoặc các chỉ số sinh tồn của bệnh nhân trong lĩnh vực y học. Các thực thể này có thể được quan sát đồng thời, tạo nên một hệ thống dữ liệu đa chiều và phức tạp, đòi hỏi các mô hình học sâu phải có khả năng xử lý và khai thác thông tin từ nhiều nguồn dữ liệu đa dạng và đồng thời. Kỹ thuật học sâu, với khả năng tự động học các đặc trưng phức tạp từ dữ liệu thô mà không cần phải thiết kế các đặc trưng thủ công, đã chứng tỏ tính hiệu quả vượt trội trong việc nắm bắt các mối quan hệ tiềm ẩn và phi tuyến tính trong dữ liệu chuỗi thời gian, từ đó cải thiện độ chính xác của dự báo. Sự phát triển này mở ra nhiều cơ hội mới cho việc áp dụng các mô hình dự báo chuỗi thời gian trong nhiều lĩnh vực khác nhau, từ dự báo thời tiết, tài chính cho đến chăm sóc sức khỏe và quản lý chuỗi cung ứng.

Trong trường hợp đơn giản nhất, các mô hình dự báo một bước thời gian tiếp theo có dạng (1):

$$\hat{y}_{i,t+1} = f(y_{i,t-k:t}, x_{i,t-k:t}, s_i) \quad (1)$$

Trong đó:

- $\hat{y}_{i,t+1}$ là dự báo của mô hình,
- $y_{i,t-k:t} = \{y_{i,t-k}, \ldots, y_{i,t}\}$ là các quan sát của mục tiêu và đầu vào bên ngoài,
- $x_{i,t-k:t} = \{x_{i,t-k}, \ldots, x_{i,t}\}$ tương ứng trong khoảng cửa sổ quan sát,
- k, s_i là siêu dữ liệu tĩnh liên quan đến thực thể (ví dụ: vị trí cảm biến),
- f(.) là hàm dự đoán được học bởi mô hình.

Ứng dụng học sâu đã chứng tỏ là một công cụ mạnh mẽ trong dự báo chuỗi thời gian. Với khả năng học và tự điều chỉnh từ dữ liệu lịch sử, các mô hình học sâu có thể phân tích các xu hướng phức tạp và các mối quan hệ giữa các biến số để đưa ra những dự đoán chính xác về tương lai. Đặc biệt, các mô hình này có khả năng tự động học từ dữ liệu mới, giúp cải thiện độ chính xác của dự báo theo thời gian và điều chỉnh dự đoán

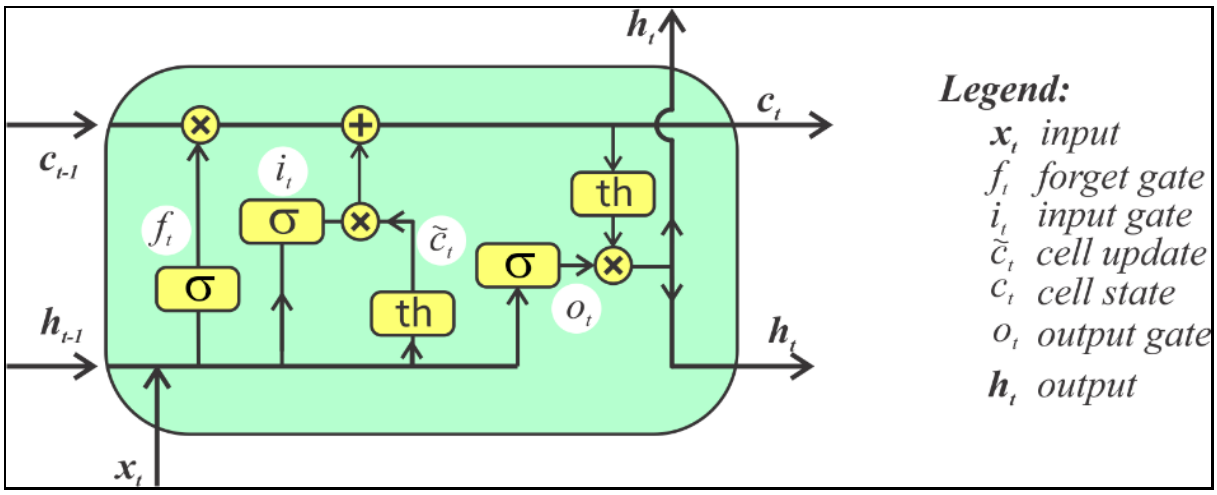
dựa trên các biến đổi không ngừng của dữ liệu. Chính nhờ khả năng này, học sâu trở nên đặc biệt linh hoạt và có thể thích ứng với các môi trường kinh doanh đang biến động, cung cấp một lợi thế quan trọng trong việc ra quyết định và lập kế hoạch chiến lược trong nhiều lĩnh vực khác nhau. Khả năng này không chỉ giúp tối ưu hóa hiệu suất dự báo mà còn mở ra tiềm năng cho những ứng dụng mới và sáng tạo, từ việc quản lý rủi ro tài chính đến tối ưu hóa chuỗi cung ứng và cải thiện chất lượng chăm sóc sức khỏe.

2.2.2. LSTM

2.2.2.1. Khái niệm

Trong một bối cảnh ngày càng phức tạp và biến động của thị trường tài chính, việc phát triển các mô hình dự báo hiệu quả là một thách thức quan trọng. LSTM, một dạng của mạng nơ-ron hồi tiếp, đã được chứng minh là một công cụ mạnh mẽ trong việc xử lý dữ liệu chuỗi thời gian, với khả năng học và ghi nhớ các mẫu dài hạn.

Nhờ cơ chế phản hồi tự động của lớp ẩn, mô hình RNN có lợi thế trong việc giải quyết các vấn đề phụ thuộc dài hạn, nhưng có những khó khăn trong ứng dụng thực tế. Để giải quyết vấn đề biến mất độ dốc của RNN, Sepp Hochreiter và Jurgen Schmidhuber (1997) đã đề xuất mô hình LSTM, và gần đây được cải tiến và phát triển bởi Alex Graves (2013).



Hình 2.2. Cấu trúc mô hình LSTM (Nguồn: Sách Modern Time Series Forecasting with Python)

Long Short-Term Memory (LSTM) là một kiến trúc mạng nơ-ron tái phát (recurrent neural network) được thiết kế để vượt qua các hạn chế của các mạng nơ-ron

khác trong việc hiệu quả hóa và sử dụng dữ liệu tuần tự. Khác với các mạng nơ-ron truyền thống, LSTM được thiết kế đặc biệt để giữ thông tin qua thời gian, giúp nó rất phù hợp với các bài toán liên quan đến phụ thuộc thời gian và tương tác trên dải dữ liệu. Đơn vị LSTM bao gồm một ô bộ nhớ lưu trữ thông tin được cập nhật bởi ba cổng đặc biệt: cổng đầu vào (input gate), cổng quên (forget gate) và cổng đầu ra (output gate).

2.2.2.2. Các thành phần

LSTM hoạt động trên một bộ cơ chế đặc biệt, cho phép nó lưu giữ và tận dụng thông tin tuần tự một cách hiệu quả. So với mô hình RNN, LSTM có thêm thành phần mới là ô nhớ, đóng vai trò là bộ nhớ dài hạn và được sử dụng cùng với bộ nhớ trạng thái ẩn của RNN. Trong LSTM, nhiều cổng có nhiệm vụ đọc, thêm và quên thông tin từ những ô nhớ này. Ô nhớ này hoạt động như một đường cao tốc có độ dốc, cho phép các cổng đi qua tương đối không bị cản trở thông qua mạng. Đây là cải tiến quan trọng giúp tránh biến mất độ dốc trong RNN.

Các thành phần cốt lõi của các đơn vị LSTM – cell state, input gate, forget gate, và output gate – góp phần chung vào khả năng xử lý và lưu giữ dữ liệu tuần tự của mạng trong khoảng thời gian kéo dài. Nhìn chung, các thành phần này cho phép LSTM giải quyết các thách thức trong việc nắm bắt các mô hình thời gian và phụ thuộc tầm xa, khiến nó trở thành tài sản không thể thiếu trong các nhiệm vụ AI đòi hỏi sự hiểu biết theo ngữ cảnh và phân tích tuần tự. Cụ thể theo sách *Modern Time Series Forecasting with Python* mô hình LSTM gồm các thành phần sau:

- Input gate: điều chỉnh luồng thông tin mới, quyết định lượng thông tin cần đọc từ đầu vào hiện tại và trạng thái ẩn trước đó.

$$I_t = \sigma(W_{xi} . x_i + W_{hi} . H_{t-1} + b_i) \quad (2)$$

- Forget gate: quyết định lượng thông tin cần quên trong bộ nhớ dài hạn.

$$F_t = \sigma(W_{xf} . x_i + W_{hf} . H_{t-1} + b_f) \quad (3)$$

- Output gate: quyết định nên sử dụng bao nhiêu Cell State hiện tại để tạo Hidden State hiện tại, đây là đầu ra của ô.

$$O_t = \sigma(W_{xo} . x_i + W_{ho} . H_{t-1} + b_o) \quad (4)$$

Ở đây, W_{xi}, W_{xf}, W_{xo}, W_{hi}, W_{hf} và W_{ho} là các thông số trọng lượng có thể học được; b_i, b_f và b_o là các tham số sai lệch có thể học được; H_{t-1} là trạng thái ẩn từ dấu thời gian trước đó.

- Cell state: là một bộ nhớ dài hạn mới, được ba cổng Input gate, Forget gate và Output gate dùng để cập nhật và quên đi bộ nhớ này. Nếu Cell state từ dấu thời gian trước đó là Ct-1 , thì ô LSTM tính toán candidate cell state, sử dụng một cổng khác, nhưng lần này có kích hoạt *tanh*:

$$C_l = tanh(W_{xc} . x_t + W_{hc} . H_{t-1} + b_c) \quad (5)$$

Ở đây, W_{xc}, W_{hc} là các thông số trọng lượng có thể học được; b_c là các tham số sai lệch có thể học được; H_{t-1} là trạng thái ẩn từ dấu thời gian trước đó.

Dưới đây là công thức cập nhập của Cell state hoặc bộ nhớ dài hạn của ô:

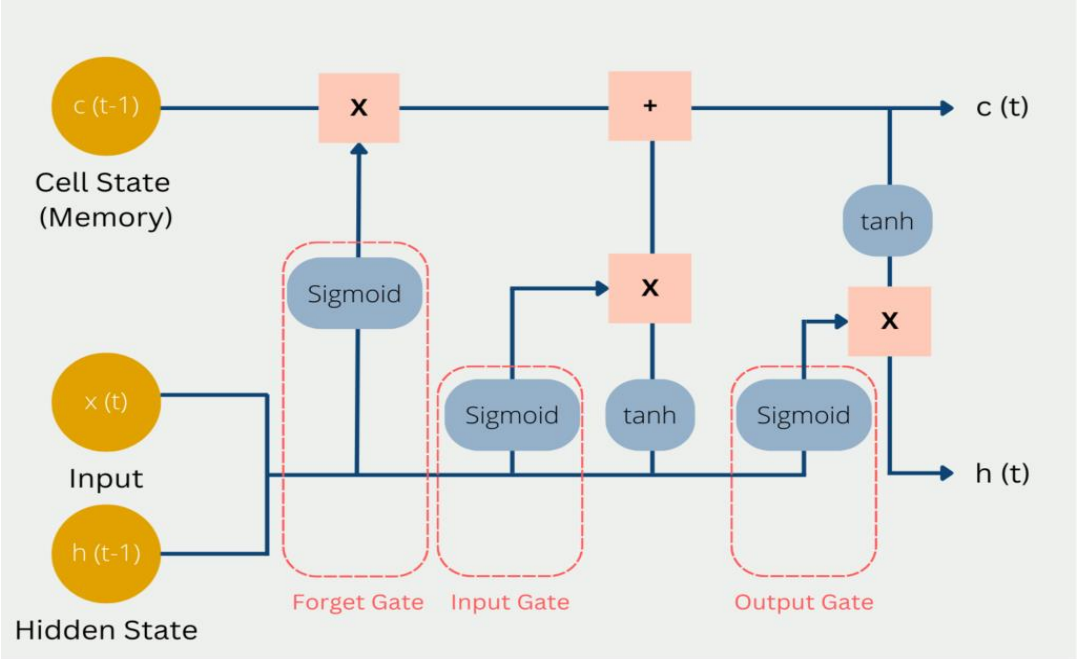
$$C_t = F_t \odot C_{t-1} + I_t \odot C_t) \quad (6)$$

Ở đây, $\odot$ là phép nhân theo phần tử.

Công thức của Hidden state hiện sau khi sử dụng Cell state hiện tại mới được tạo và cổng đầu ra để quyết định mức độ thông tin để truyền tới bộ dự đoán:

$$H_t = O_t \odot tanh(C_t) \quad (7)$$

2.2.2.3. Cách thức hoạt động



Hình 2.3. Mô hình LSTM

Tại thời điểm t:

x_t: dữ liệu đầu vào của ô LSTM

h_{t-1}: đầu ra của ô LSTM ở thời điểm trước đó

c_t: giá trị của ô nhớ

h_t: đầu ra của ô LSTM

Quá trình tính toán của một ô LSTM được chia thành các bước sau:

(1) Tính toán giá trị của ô nhớ tạm $\tilde{c}_t$

(2) Tính toán giá trị của cổng vào i_t.

(3) Tính toán giá trị của cổng quên f_t.

(4) Tính toán giá trị của ô nhớ thời điểm hiện tại c_t.

(5) Tính toán giá trị của cổng ra o_t.

(6) Cuối cùng, tính toán đầu ra của ô LSTM h_t.

Trong Forget gate, nó quyết định thông tin nào hiện tại và trước đó được giữ lại và thông tin nào sẽ bị loại bỏ. Điều này bao gồm Hidden state từ thẻ trước và đầu vào hiện tại. Các giá trị này được chuyển vào hàm sigmoid, hàm này chỉ có thể xuất các giá trị từ 0 đến 1. Giá trị 0 có nghĩa là thông tin trước đó có thể bị quên vì có thể có thông tin mới, quan trọng hơn. Số một có nghĩa là thông tin trước đó được bảo tồn. Kết quả từ việc này được nhân với Cell state hiện tại để kiến thức không còn cần thiết sẽ bị lãng quên vì nó được nhân với 0 và do đó bị loại bỏ.

Trong Input gate, nó quyết định giá trị của đầu vào hiện tại để giải quyết bài toán. Đối với điều này, đầu vào hiện tại được nhân với Hidden state và ma trận trọng số của lần chạy cuối cùng. Sau đó, tất cả thông tin xuất hiện quan trọng trong Input gate sẽ được thêm vào Cell state và tạo thành Cell state mới c(t). Cell state mới này hiện là trạng thái hiện tại của bộ nhớ dài hạn và sẽ được sử dụng trong lần chạy tiếp theo.
Trong Output gate, đầu ra của mô hình LSTM sau đó được tính toán ở Hidden state. Tùy thuộc vào ứng dụng, nó có thể là một từ bổ sung cho ý nghĩa của câu chẳng hạn. Để làm điều này, hàm sigmoid quyết định thông tin nào có thể đi qua cổng đầu ra và sau đó Cell state sẽ được nhân lên sau khi được kích hoạt bằng hàm *tanh*.

2.2.2.4. Ưu điểm và hạn chế

Ưu điểm:

- Khả năng duy trì sự phụ thuộc lâu dài: LSTM có khả năng giữ và sử dụng thông tin từ những khoảng thời gian xa trước đó, giúp mạng nắm bắt được các mẫu dữ liệu phức tạp và dài hạn.

- Tăng cường xử lý dữ liệu tuần tự trong các ứng dụng AI: Khả năng của LSTM trong việc xử lý dữ liệu tuần tự, như văn bản, âm thanh, hoặc chuỗi thời gian, làm cho nó trở thành công cụ quan trọng trong nhiều ứng dụng trí tuệ nhân tạo.

- Giảm thiểu các vấn đề về độ dốc biến mất và bùng nổ: LSTM được thiết kế với các cổng đặc biệt để kiểm soát luồng thông tin và trạng thái của mạng, giúp giảm thiểu vấn đề của việc biến mất hoặc bùng nổ độ dốc trong quá trình huấn luyện.

- Tạo điều kiện học tập theo ngữ cảnh và mô hình dự đoán: Các cơ chế trong LSTM cho phép nó học từ ngữ cảnh và dự đoán các sự kiện tiếp theo trong dữ liệu tuần tự, làm cho nó phù hợp cho các nhiệm vụ như dự đoán chuỗi thời gian hoặc tự động hoá ngôn ngữ tự nhiên.

Hạn chế:

- Độ phức tạp tính toán trong đào tạo và triển khai: LSTM yêu cầu nhiều tính toán trong quá trình huấn luyện và triển khai, đặc biệt là khi xử lý các dữ liệu lớn và phức tạp.

- Dễ bị trang bị quá mức trong một số trường hợp nhất định: Trong một số tình huống, LSTM có thể học quá nhiều từ dữ liệu đào tạo và trở nên quá tinh chỉnh cho dữ liệu mới, gây ra hiện tượng quá mức (overfitting).

- Yêu cầu điều chỉnh và tối ưu hóa tham số mở rộng: Để đạt được hiệu suất tốt nhất, LSTM cần phải được điều chỉnh và tối ưu hóa các tham số mở rộng một cách cẩn thận, điều này có thể tốn thời gian và công sức đáng kể.

2.2.2.5. Ứng dụng thực tế

Trong xử lý ngôn ngữ tự nhiên, LSTM được sử dụng để tạo ra các hệ thống AI có khả năng tạo ra phản hồi tự nhiên trong giao tiếp, phân tích cảm xúc trong văn bản và dịch máy chính xác.

Trong nhận dạng giọng nói, LSTM giúp hệ thống chính xác chuyển đổi âm thanh thành văn bản và hỗ trợ ra lệnh bằng giọng nói.

Trong dự báo tài chính, LSTM cho phép hệ thống AI dự đoán xu hướng và hành vi thị trường thông qua việc xử lý dữ liệu chuỗi thời gian, từ đó cung cấp thông tin quan trọng cho việc ra quyết định đầu tư và phân tích rủi ro.

2.2.3. GRU

GRU (Gated recurrent unit) là một kiến trúc mạng nơ-ron hồi quy (RNN) được giới thiệu bởi Cho và cộng sự (2014). Nó là một biến thể của mạng LSTM (Long Short-Term Memory) và được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên, dự đoán chuỗi thời gian và các tác vụ khác. So với mô hình LSTM, GRU loại bỏ thành phần bộ nhớ dài hạn và thay bằng cách sử dụng trạng thái ẩn (hidden state) để truyền tải thông tin. Mô hình này sử dụng một số lượng cổng ít hơn để điều chỉnh thông tin, làm mô hình đơn giản hơn và nhanh hơn trong việc tính toán. Các cổng trong GRU bao gồm cổng "quên" (Reset gate) và cổng "cập nhật" (Update gate), cho phép xác định xem thông tin nào sẽ được bỏ qua và thông tin mới nào sẽ được cập nhật vào trạng thái ẩn. Qua cách này, GRU tạo ra một cơ chế tự nén thông tin và điều chỉnh quá trình truyền tải thông tin, giúp cải thiện độ chính xác và hiệu suất của mô hình.

Hai cổng của GRU bao gồm:

- Cổng đặt lại: Cổng này quyết định bao nhiêu trạng thái ẩn trước đó sẽ được coi là trạng thái ẩn trong dấu thời gian hiện tại. Phương trình của cổng quên được mô tả dưới đây:

$$R_t = \sigma(W_{xr} . x_t + W_{hr} . H_{t-1} + b_r) \quad (8)$$

Theo Kostadinov, S. (2019) cổng đặt lại (reset gate) đóng vai trò quan trọng trong việc quyết định mức độ thông tin từ quá khứ cần được lãng quên. Công thức tính toán cho cổng đặt lại tương tự như công thức của cổng cập nhật, nhưng sự khác biệt nằm ở trọng số và cách sử dụng của từng cổng. Cổng đặt lại cho phép mô hình kiểm soát và loại bỏ thông tin không cần thiết từ các bước thời gian trước đó, giúp tối ưu hóa quá trình học tập và duy trì các thông tin có ý nghĩa.

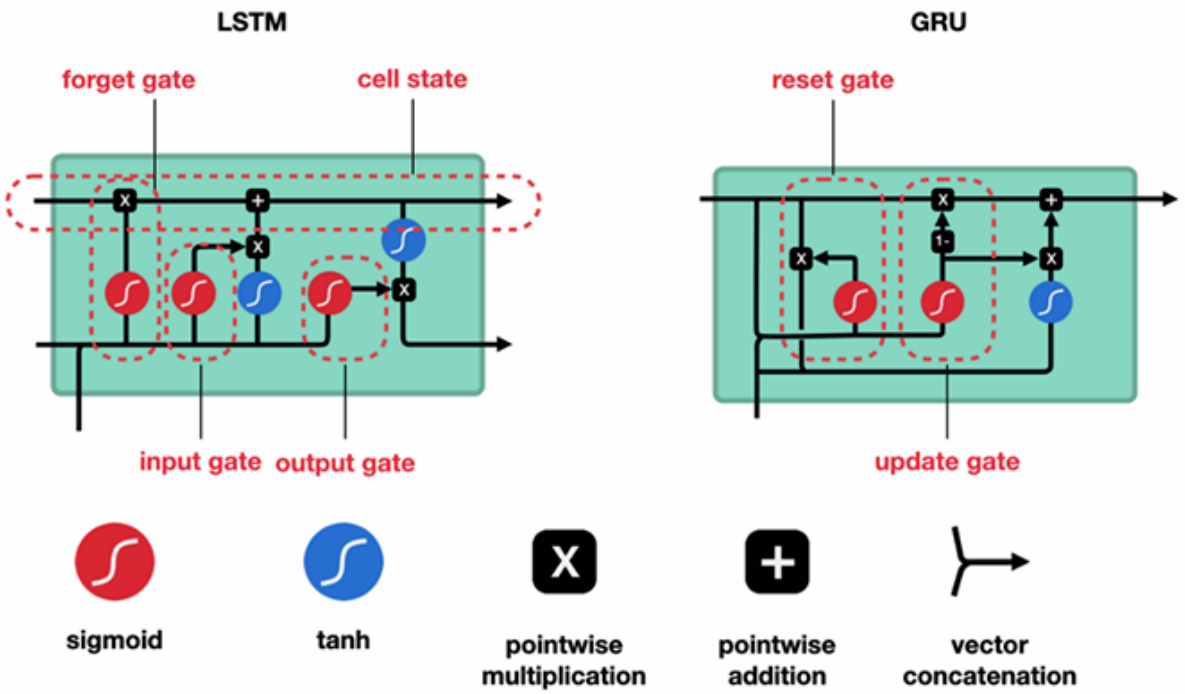
- Cổng cập nhật: Cổng cập nhật quyết định bao nhiêu trạng thái ẩn trước đó sẽ được chuyển tiếp và bao nhiêu trạng thái ẩn hiện tại sẽ được ghi vào. Phương trình được thể hiện như sau:

$$U_t = \sigma(W_{xu} . x_t + W_{hu} . H_{t-1} + b_u) \quad (9)$$

Trong đó:

- W_{xr}, W_{hr}, W_{xu}, W_{hu}: là các thông số trọng lượng có thể học được,
- b_r, b_u: là tham số sai lệch có thể học được,
- H_{t-1}: là trạng thái ẩn từ dấu thời gian trước đó.

Theo Kostadinov, S. (2019) cổng cập nhật (update gate) đóng vai trò quan trọng trong cơ chế hoạt động của các đơn vị GRU (Gated Recurrent Unit). Cổng cập nhật giúp mô hình xác định mức độ thông tin từ các bước thời gian trước đó cần được duy trì và truyền tải tới các bước thời gian sau. Điều này mang lại sức mạnh đáng kể cho mô hình, vì nó có khả năng quyết định sao chép toàn bộ thông tin từ quá khứ, từ đó loại bỏ nguy cơ vấn đề gradient biến mất (vanishing gradient). Khả năng này cho phép các đơn vị GRU duy trì và cập nhật thông tin một cách hiệu quả qua các chuỗi thời gian dài, cải thiện hiệu suất của mô hình trong việc học các phụ thuộc dài hạn. Chúng ta sẽ tìm hiểu chi tiết hơn về việc sử dụng cổng cập nhật sau. Hiện tại, cần lưu ý công thức tính U_t (giá trị của cổng cập nhật tại thời điểm t).



Hình 2.4. Sơ đồ cổng của LSTM so với GRU

Hình 2.4. đã trực quan hóa sự khác nhau giữa sơ đồ cổng của hai mô hình LSTM và GRU. Khi nói về dữ liệu chuỗi thời gian, giữa hai kiến trúc mạng nơ-ron hồi tiếp (RNN) phổ biến là Long Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU) có sự khác biệt nhất định trong việc xử lý dữ liệu chuỗi thời gian, cụ thể như sau:

LSTM và GRU đều có các cổng (gates) để kiểm soát việc truyền thông tin qua các bước thời gian, tuy nhiên, chúng có cách thức hoạt động khác nhau. LSTM sử dụng ba cổng: cổng quên (forget gate), cổng đầu vào (input gate) và cổng đầu ra (output gate), trong khi GRU chỉ sử dụng hai cổng: cổng cập nhật (update gate) và cổng đặt lại (reset gate). Sự khác biệt này ảnh hưởng đến cách mỗi mô hình xử lý thông tin và quản lý trạng thái bên trong.

Đặc điểm tiêu biểu của LSTM là khả năng duy trì trạng thái bên trong của mạng (cell state), nhờ vào cổng quên giúp mô hình quyết định nên giữ hoặc loại bỏ thông tin từ bước thời gian trước đó. Điều này giúp LSTM có khả năng học và ghi nhớ các phụ thuộc dài hạn trong dữ liệu chuỗi thời gian. Trong khi đó, GRU sử dụng cổng cập nhật để quyết định mức độ thông tin nào cần được truyền tiếp qua các bước thời gian, cùng với cổng đặt lại để kiểm soát việc lãng quên thông tin không cần thiết. Cấu trúc đơn giản hơn của GRU giúp giảm độ phức tạp của mô hình, làm cho việc huấn luyện và triển khai trở nên đơn giản hơn so với LSTM.

Tuy nhiên, sự đơn giản của GRU có thể làm giảm sức mạnh mô hình trong việc xử lý các chuỗi dữ liệu phức tạp, trong khi LSTM có thể đạt được hiệu suất tốt hơn nhờ vào khả năng duy trì trạng thái bên trong phong phú hơn. Điều này đặt ra một thách thức trong việc lựa chọn kiến trúc phù hợp nhất với bài toán cụ thể và yêu cầu của dữ liệu chuỗi thời gian.

2.3. Các nghiên cứu liên quan

Dự báo giá cổ phiếu là một trong những thách thức quan trọng trong lĩnh vực tài chính, đặc biệt là trong ngành đầu tư cổ phiếu, ảnh hưởng đến quy trình ra quyết định và các chiến lược cắt giảm rủi ro. Thành công của các phương pháp học máy trong dự báo chuỗi thời gian đã mở ra nhiều cơ hội để cải thiện độ chính xác và hiệu quả trong việc dự báo giá cổ phiếu tương lai. Martínez và các cộng sự (2019) đã thiết kế một phương pháp dựa trên mô hình k người hàng xóm gần nhất (k-NN) cho việc dự báo chuỗi thời gian. Kết quả cho thấy rằng không chỉ việc lựa chọn mô hình mà còn việc sử dụng các kỹ thuật tiền xử lý phù hợp để đạt được các dự báo chính xác. Ngoài ra, bài nghiên cứu của Sezer và các cộng sự (2020) đã khảo sát các phương pháp áp dụng các mô hình học sâu trong dự báo chuỗi thời gian tài chính. Khảo sát này cho thấy các phương pháp học sâu phổ biến nhất áp dụng cho lĩnh vực tài chính là Convolutional

Neural Networks (CNNs), Deep Belief Networks (DBNs), và Long-Short Term Memory (LSTM).

Các công trình về dự báo giá cổ phiếu sử dụng học sâu đã trình bày bằng chứng thuyết phục về hiệu suất siêu việt của học sâu trong việc giải quyết bài toán này. Zhou và các cộng sự (2020) cũng đã đưa ra một mô hình tiên tiến dựa trên transformer cho việc dự báo chuỗi thời gian dài, được gọi là "Informer". Các tác giả đã đề xuất sự phụ thuộc của các điểm thời gian trên một khoảng cách xa cần được mô hình hoá một cách hiệu quả thông qua phương pháp tự chú ý tinh vi (self-attention distilling), nhờ vậy Informer là một mô hình phù hợp hơn cho các nhiệm vụ yêu cầu dự báo mạnh mẽ trên một tầm nhìn xa. Kulachinskaya (2022) đã đề xuất một hệ thống trí tuệ nhân tạo dựa trên mạng nơ-ron và Word2vec, cho thấy biến động giá được tích hợp với một số nguồn thông tin khác để đưa vào các yếu tố phân tích cơ bản khác góp phần cải thiện hiệu quả dự đoán giá hợp đồng tương lai. Zheng và các cộng sự (2023) đã so sánh các mô hình Transformer với các mô hình thông thường và cho thấy kết quả rằng các mô hình dựa trên Transformer, trong đó có Informer, giúp giải quyết các điểm yếu của các mô hình thông thường và cho ra kết quả tốt hơn.

Trong các mô hình dự báo chuỗi thời gian dự báo học sâu, hai trong các mô hình phổ biến và hiệu quả nhất cho giải quyết bài toán này là Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU). Nguyen và Yoon (2019) đã đề xuất một phương pháp mới cho việc dự đoán biến động giá cổ phiếu ngắn hạn bằng cách sử dụng học chuyển giao. Họ tiền huấn luyện một mô hình cơ sở với các đơn vị LSTM sử dụng dữ liệu từ nhiều cổ phiếu khác nhau, sau đó điều chỉnh mô hình cơ sở bằng dữ liệu từ một cổ phiếu mục tiêu và các đặc điểm đầu vào khác nhau để cải thiện hiệu suất. Tashiro và cộng sự (2019) áp dụng các kiến trúc mạng nơ-ron tích chập vào các đặc trưng dựa trên đơn đặt hàng để dự đoán xu hướng giá trung bình. Kết quả cho thấy các bộ lọc làm mịn mà họ đề xuất thay vì các đặc trưng nhúng cải thiện độ chính xác. Họ cũng tiến hành mô phỏng đầu tư để chứng minh tính hiệu quả và tính thực tế của mô hình của họ. Trong nghiên cứu so sánh của Shahi và cộng sự (2020) về dự báo giá cổ phiếu, các mô hình LSTM và GRU đã được đánh giá, cho thấy hiệu suất được cải thiện khi tâm trạng tin tức tài chính được tích hợp cùng các đặc điểm cổ phiếu trong một kiến trúc học sâu hợp tác. Ngoài ra, Sen và cộng sự (2021) trình bày mười mô hình hồi quy học sâu cho dự báo

giá cổ phiếu chính xác, tập trung đặc biệt vào một công ty trong ngành ô tô ở Ấn Độ. Nghiên cứu của họ nhấn mạnh về độ chính xác và hiệu suất của các mô hình này trong việc dự đoán giá trong tương lai. Mehtab và cộng sự (2022) đề xuất một phương pháp mạnh mẽ sử dụng các mô hình học sâu dựa trên CNN và LSTM để dự báo giá cổ phiếu chính xác. Tập trung vào thời gian thực hiện và giá trị RMSE, nghiên cứu nhằm mục tiêu cải thiện độ chính xác của dự báo và đã thành công. Vuong và cộng sự (2022) giới thiệu một phương pháp kết hợp XGBoost cho việc lựa chọn đặc trưng và LSTM cho dự báo giá cổ phiếu. Kết quả thực nghiệm dựa trên tập dữ liệu Forex trong giai đoạn 2008-2018 cho thấy rằng phương pháp này vượt trội hơn phương pháp ARIMA trên các chỉ số đánh giá như MAE, MSE và MRSE.

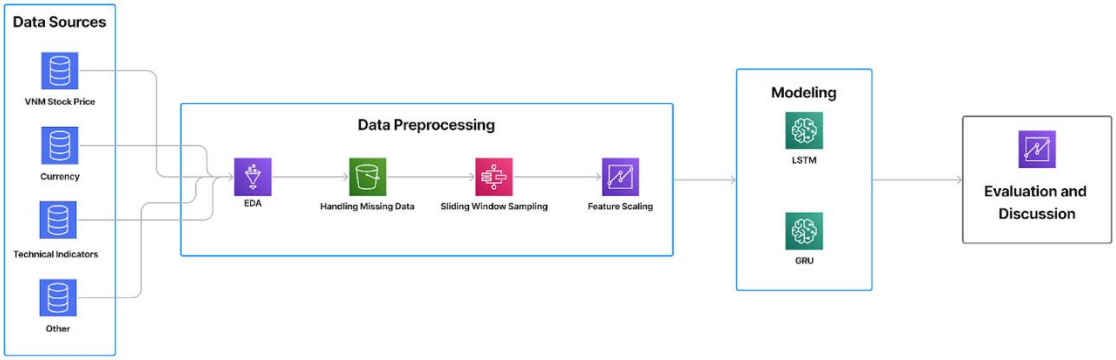
Mặc dù các nghiên cứu trên đều đưa ra kết luận rằng LSTM và GRU hoạt động hiệu quả cho bài toán dự báo chuỗi thời gian, rất ít các bài nghiên cứu đưa ra phân tích tổng quan khi sử dụng dự báo với cửa sổ trượt, một công cụ rất hiệu quả để liên tục dự báo và cập nhật các xu hướng ngắn hạn. Alberg và cộng sự (2018) giới thiệu hai thuật toán ARIMA dựa trên cửa sổ trượt không theo mùa và hai thuật toán ARIMA dựa trên cửa sổ trượt theo mùa để dự báo tải trọng ngắn hạn trong các công tơ thông minh. Để đánh giá phương pháp đề xuất, họ sử dụng luồng dữ liệu tiêu thụ thực tế theo giờ được ghi lại bởi sáu đồng hồ thông minh trong khoảng thời gian 16 tháng. Kết quả cho thấy kết quả tốt nhất (MAPE = 10,05%) thu được với mô hình ARIMA không theo mùa, đồng thời cho thấy tính ứng dụng cao của phương pháp sliding window trong dự báo chuỗi thời gian. Trong lĩnh vực tài chính, Kumar và cộng sự (2018) đã đánh giá hiệu quả của Long Short Term Memory (LSTM) trong việc áp dụng các chỉ số phân tích kỹ thuật như biến đầu vào để dự đoán giá cổ phiếu của mã AAPL trên sàn NASDAQ. Hiệu suất với ba lớp kích hoạt đầu ra phổ biến được kiểm tra cùng với bộ tối ưu Adam , được so sánh bằng cách sử dụng độ lệch bình phương trung bình căn (Root Mean Square Deviation - RMSE). Kết quả cho thấy mô hình có giá trị RMSE trung bình là 12.483 với kích hoạt đầu ra tuyến tính được chuẩn hóa trong khoảng (0,1) và 3.258 khi chuẩn hóa trong khoảng (-1,1), 21.769 với kích hoạt đầu ra sigmoid chuẩn hóa trong khoảng (0,1) và 21.738 với kích hoạt đầu ra tanh chuẩn hóa trong khoảng (-1,1). Các nghiên cứu này đã áp dụng cửa sổ trượt với LSTM và giải quyết vấn đề cập nhật các biến động liên tục của giá cổ phiếu, tuy nhiên chưa đưa ra sự so sánh với mô hình khác để kiểm

định tính hiệu quả tổng quát. Ngoài ra, một vấn đề khác là các mô hình này chưa sử dụng dữ liệu bao gồm đa dạng yếu tố, trong đó có một số yếu tố thường ảnh hưởng đến giá cổ phiếu bao gồm biến động chung của thị trường, các chỉ số kỹ thuật và ảnh hưởng từ thị trường tiền tệ.

Để giải quyết các hạn chế này, nhóm đề xuất sử dụng một hướng tiếp cận dự đoán giá cổ phiếu ngắn hạn sử dụng các mô hình LSTM và GRU trên dữ liệu chuỗi thời gian đa biến. Vượt qua những thách thức này, phương pháp dự báo ngắn hạn dựa trên cửa sổ trượt, đồng thời cân nhắc đa dạng yếu tố ảnh hưởng, mang đến lợi ích to lớn trong việc phân tích, dự báo giá cổ phiếu ngắn hạn, đóng vai trò như một công cụ giúp hỗ trợ ra quyết định và giảm thiểu rủi ro.

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Tổng quan mô hình nghiên cứu



Hình 3.1. Mô hình nghiên cứu

Nhằm đánh giá được tính hiệu quả của việc áp dụng dự báo chuỗi thời gian sử dụng các mô hình học sâu, nhóm đề xuất quy trình thực nghiệm gồm 4 bước được trình bày ở hình 3.1. Mô hình nghiên cứu này giúp thể hiện các bước cụ thể trong dự báo chuỗi thời gian từ thu thập, tiền xử lý và so sánh các mô hình, lựa chọn mô hình tối ưu nhất. Các bước chi tiết trong phương pháp nghiên cứu gồm:

i) Thu thập dữ liệu: Đây là bước đầu tiên và quan trọng đối với chất lượng của dữ liệu đầu vào cho bài toán dự đoán. Nhằm hình thành dữ liệu chuỗi thời gian đa biến, các loại dữ liệu sau đã được thu thập:

- Dữ liệu lịch sử giao dịch chứng khoán: dữ liệu giá các mã chứng khoán được thu thập sử dụng package VNSTOCK. Các mã được thu thập gồm VNM là dữ giá dự báo mục tiêu và VNINDEX nhằm thể hiện biến động chung của thị trường.
- Dữ liệu tiền tệ: dữ liệu điểm USDINDEX thể hiện biến động của đồng USD và dữ liệu tỷ giá giữa đồng VND và USD.
- Các chỉ báo kỹ thuật: được trích xuất từ giá ‘close’ của mã VNM.

ii) Tiền xử lý dữ liệu: Dữ liệu cần được chuẩn bị trước khi có thể được sử dụng để huấn luyện mô hình dự báo. Tiền xử lý dữ liệu bao gồm các bước sau:

- Phân tích khám phá dữ liệu: thực hiện phân tích để tìm hiểu bản chất của dữ liệu và tìm ra các điểm dữ liệu cần làm sạch để tối ưu cho mô hình dự đoán.

- Xử lý giá trị thiếu: các giá trị thiếu trong dữ liệu có thể được xử lý bằng cách điền vào giá trị kế tiếp (Forward Fill) hoặc trước đó (Backward Fill) trong dữ liệu chuỗi thời gian.

- Lấy mẫu dữ liệu: chia tập dữ liệu huấn luyện và kiểm thử theo phương pháp cửa sổ trượt.

- Biến đổi dữ liệu: dữ liệu có thể cần được chuyển đổi sang dạng phù hợp cho mô hình dự báo. Ví dụ, dữ liệu có thể cần được chuẩn hóa hoặc khử chuẩn hóa.

iii) Mô hình hóa: Sau khi có được dữ liệu đã tiền xử lý, bước tiếp theo bao gồm xây dựng các mô hình dự báo. Có hai mô hình dự báo chuỗi thời gian được sử dụng nhằm mục đích so sánh và chọn ra mô hình tốt nhất gồm LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit). LSTM và GRU đều là hai loại mô hình mạng nơ-ron tái phát cơ bản trong học sâu, thích hợp cho việc xử lý dữ liệu chuỗi thời gian. LSTM sử dụng các cổng đặc biệt để điều chỉnh thông tin bộ nhớ, trong khi GRU sử dụng cơ chế cổng để điều chỉnh việc truyền thông tin. Cả hai đều giải quyết vấn đề biến mất gradient và giúp mô hình ghi nhớ thông tin từ quá khứ để dự báo tương lai, với GRU thường có hiệu quả tính toán cao hơn và ít tham số hơn so với LSTM.

iv) Đánh giá: Mô hình dự báo cần được đánh giá để đảm bảo rằng nó có thể dự đoán chính xác giá trị tương lai của chuỗi dữ liệu. Quá trình đánh giá bao gồm các bước sau:

- Tính toán các số liệu hiệu suất: các số liệu hiệu suất như lỗi trung bình tuyệt đối (MAE), lỗi bình phương trung bình (MSE) và tỷ lệ lỗi trung bình (MAPE), được sử dụng để đánh giá hiệu suất của mô hình dự báo.

- Phân tích mô hình: phân tích các mô hình để tìm ra điểm mạnh, điểm yếu của chúng.

- So sánh các mô hình: các mô hình dự báo khác nhau có thể được so sánh để xác định mô hình tốt nhất.

3.2. Phương pháp lấy mẫu

3.2.1. Tách dữ liệu

Bảng 3.1. Tách dữ liệu chuỗi thời gian

Set	Start Date	End Date

Training Set	2014-01-01	2023-11-30
Validation Set	2023-12-01	2023-12-31
Test Set	2024-01-01	2024-01-31

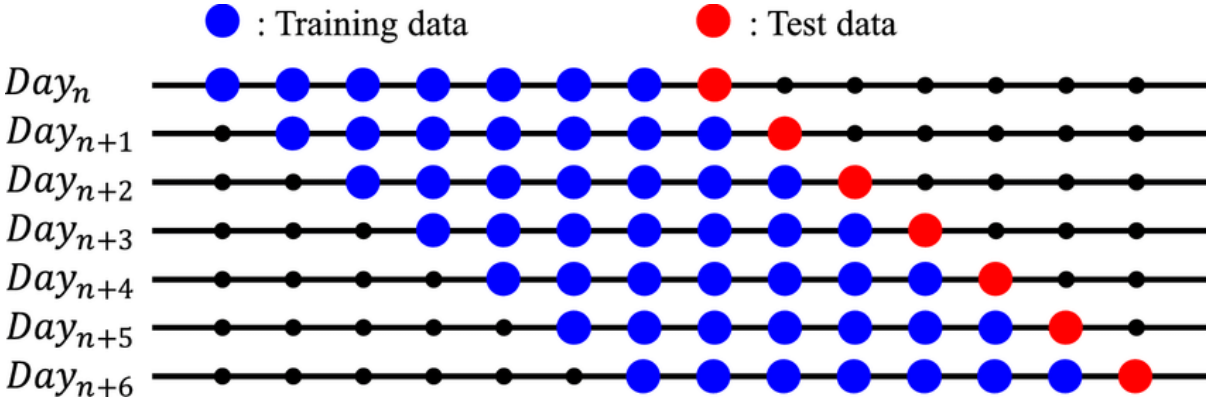
Tập dữ liệu được sử dụng cho thực nghiệm là dữ liệu giao dịch lịch sử theo ngày của mã VNM từ ngày 2014-01-01 đến ngày 2024-06-05. Dữ liệu chuỗi thời gian được thu thập bao gồm 3812 điểm thời gian, mỗi điểm thời gian đại diện cho một ngày. Ở *Bảng 3.1.*, bộ dữ liệu được chia làm 2 tập, điều này đóng vai trò quan trọng trong việc đánh giá hiệu suất của mô hình LSTM theo thời gian, mô phỏng một tình huống thực tế trong đó mô hình sẽ dự đoán giá cổ phiếu trong tương lai dựa trên dữ liệu lịch sử.

- Tập Dữ Liệu Huấn Luyện: Gồm dữ liệu từ đầu năm 2014 đến ngày 02 tháng 09 năm 2023, cung cấp một khoảng thời gian đáng kể để mô hình học các mẫu và xu hướng cơ bản trong giá cổ phiếu.

- Tập Dữ Liệu Kiểm Tra: Từ ngày 03 tháng 9 năm 2023 đến ngày 06 tháng 5 năm 2024, chứa một tập dữ liệu không được mô hình nhìn thấy trước để đánh giá độ chính xác trong dự đoán và khả năng tổng quát hóa của mô hình với dữ liệu mới.

Trong bài nghiên cứu này, dữ liệu huấn luyện và kiểm thử sẽ được sử dụng để lấy mẫu theo phương pháp Rolling Forecast bằng cách chia dữ liệu thành nhiều cửa sổ chạy (sliding window), sau đó thực hiện dự đoán giá đóng cửa của ngày tiếp theo của cửa sổ.

3.2.2. Rolling Forecast



Hình 3.2. Phương pháp Sliding window cho chuỗi thời gian

Rolling Forecast đóng vai trò là nền móng trong phân tích dự đoán, cung cấp một khung công việc động để tích hợp thông tin mới vào các mô hình dự đoán theo thời

gian thực. Phương pháp này được đánh giá cao đặc biệt trong môi trường nhanh chóng nơi khả năng thích ứng với dữ liệu mới một cách nhanh chóng có thể ảnh hưởng đáng kể đến độ chính xác và hiệu quả của quyết định.

Theo Shijin Yuan và cộng sự (2021) phương pháp Rolling Forecast bao gồm việc sử dụng mô hình dự báo để đưa ra dự đoán cho một loạt thời điểm, sau đó cập nhật mô hình với dữ liệu thực tế cho từng thời điểm tiếp theo. Quá trình này được lặp lại trong một số thời điểm nhất định, với mô hình được cập nhật liên tục để kết hợp thông tin mới khi có sẵn. Trong bối cảnh được cung cấp, phương pháp Rolling Forecast đã được sử dụng trong nghiên cứu để dự báo cường độ bão dựa trên các mô hình LSTM. Nó liên quan đến việc thu thập các giá trị dự đoán tại mỗi thời điểm và sau đó cập nhật dự báo dựa trên phương pháp dự báo tổng hợp, kết hợp các dự đoán từ nhiều mô hình.

Bản chất của Rolling Forecast nằm trong việc điều chỉnh lặp lại của chúng, được thực hiện qua các khoảng thời gian ngắn gọn để đảm bảo cả hiệu quả tính toán và việc bắt kịp các xu hướng dữ liệu đương đại. Cốt lõi của phương pháp này là tham số bước cửa sổ, điều chỉnh tốc độ tiến triển của cửa sổ phân tích qua tập dữ liệu. Tham số này quan trọng trong việc xác định sự trùng lặp giữa các cửa sổ liên tiếp, ảnh hưởng đến độ tinh tế của phân tích và tải tính toán. Toán học, bước cửa sổ có thể được biểu diễn như sau:

$$\mathrm{WS} = \frac{TW}{N} \tag{10}$$

Trong đó:

- WS là bước cửa sổ,
- TW là tổng kích thước cửa sổ (ví dụ, 24 giờ cho dữ liệu hàng ngày),
- N là số bước tiến lên sau mỗi lần lặp.

Ví dụ, đặt N = 1 dẫn đến sự trùng lặp tối đa, cung cấp một phân tích chi tiết, mặc dù là công việc tính toán nặng nề. Ngược lại, tăng N giảm sự trùng lặp, tăng tốc độ tính toán nhưng có thể làm giảm độ sâu phân tích.

Việc chọn bước cửa sổ tối ưu là một quyết định chiến lược cần phải cân nhắc cẩn thận, phải cân đối giữa việc cần đánh giá mô hình một cách chi tiết và hạn chế về tài nguyên tính toán. Nếu chọn bước cửa sổ nhỏ hơn, đó sẽ tăng độ phân giải của việc đánh giá dự báo, giúp nhận biết được những biểu hiện tinh tế về hiệu suất của mô hình qua

thời gian. Tuy nhiên, điều này sẽ đồng nghĩa với việc tăng yêu cầu về tài nguyên tính toán. (Tashman, 2000) Ngược lại, nếu chọn bước cửa sổ lớn hơn, mặc dù sẽ tiết kiệm tài nguyên tính toán, nhưng có thể bỏ qua những mẫu dữ liệu tinh tế nhưng có ý nghĩa, có thể làm giảm chính xác của dự báo.

3.3. LSTM

Trong nghiên cứu này, chúng tôi áp dụng kiến trúc mạng LSTM với hai lớp LSTM để dự đoán giá chứng khoán. Mỗi lớp LSTM được cấu hình với số lượng đơn vị (units) và các tham số khác nhau, nhằm tối ưu hóa khả năng biểu diễn và dự đoán. Cụ thể, các thông số của các lớp LSTM được mô tả như sau:

Bảng 3.2. Các Layers trong mô hình LSTM

Layer (type)	**Output Shape**	**Param #**
lstm (LSTM)	(None, window_size, 128)	70,144
lstm_1 (LSTM)	(None, 64)	49,408
dense (Dense)	(None, 25)	1,625
dense_1 (Dense)	(None, 1)	26

Trong đó,

- *window_size*: kích thước của cửa sổ trượt sẽ không cố định là 30 hay 60.

- *lstm (LSTM):* Layer LSTM đầu tiên (lstm) có kiến trúc (None, 60, 128), với số lượng batch không xác định, 60 thời điểm trong chuỗi, và mỗi thời điểm được biểu diễn bằng một vector có kích thước 128. Layer này có tổng cộng 70,144 tham số, bao gồm trọng số và độ lệch, giúp mô hình học và hiểu mối quan hệ phức tạp trong dữ liệu chuỗi.

- *lstm_1 (LSTM):* Layer LSTM thứ hai (lstm_1) tạo ra đầu ra có kích thước (None, 64), với số lượng batch không xác định và mỗi batch được biểu diễn bằng một vector có kích thước 64. Với tổng cộng 49,408 tham số, layer này tiếp tục cung cấp thông tin quan trọng và giảm chiều dữ liệu cho lớp tiếp theo.

- *dense (Dense):* Layer Dense thứ ba (dense) chuyển đổi đầu ra từ lớp trước thành dạng (None, 25), với số lượng batch không xác định và mỗi batch được biểu diễn bằng một vector có kích thước 25. Có tổng cộng 1,625 tham số trong layer này, giúp mô hình tổng hợp thông tin từ các thời điểm trước đó thành biểu diễn rút ra và đưa ra dự đoán.

- *dense_1 (Dense):* Layer Dense cuối cùng (dense_1) tạo ra đầu ra có kích thước (None, 1), với số lượng batch không xác định và mỗi batch được biểu diễn bằng một vector có kích thước 1. Layer này chỉ có 26 tham số, nhằm dự đoán giá chứng khoán cho thời điểm tiếp theo dựa trên các biểu diễn đã học từ các lớp trước đó.

Sử dụng hai lớp LSTM trong một mô hình dự đoán giá chứng khoán không chỉ giúp nâng cao hiệu suất dự đoán mà còn mở ra nhiều tiềm năng khác trong lĩnh vực nghiên cứu và ứng dụng thị trường tài chính.

Theo Salman và cộng sự (2018), khả năng biểu diễn phức tạp của mô hình được tăng cường khi mỗi lớp LSTM có thể học và tái tạo các mẫu phức tạp từ dữ liệu đầu vào. Việc kết hợp nhiều lớp LSTM mở ra không gian biểu diễn rộng hơn, giúp mô hình nắm bắt được sự phức tạp của dữ liệu và dự đoán một cách chính xác hơn. Thứ hai, tính linh hoạt của mô hình được nâng cao, khi mỗi lớp LSTM có khả năng học các mặt khác nhau của dữ liệu đầu vào, giúp tạo điều kiện cho một mô hình linh hoạt, có khả năng học được các mối quan hệ phức tạp giữa các yếu tố đầu vào, từ đó cải thiện khả năng dự đoán. Cuối cùng, việc sử dụng nhiều lớp LSTM tạo ra một không gian biểu diễn phức tạp hơn, giúp mô hình hiểu sâu hơn về dữ liệu và dự đoán. Tuy nhiên, việc này có thể tăng độ phức tạp của mô hình và đòi hỏi thêm dữ liệu và tài nguyên tính toán, điều này cần được cân nhắc cẩn thận trong việc thiết kế và đánh giá mô hình.

Theo nghiên cứu của Milica Ćirić và cộng sự (2023) , việc sử dụng nhiều lớp LSTM (stacked LSTM) thường mang lại hiệu suất tốt hơn so với một lớp LSTM đơn. Việc sử dụng 1 lớp LSTM có thể gặp khó khăn khi học và biểu diễn các mẫu phức tạp hoặc khi dữ liệu có sự phụ thuộc đa chiều sâu. Bên cạnh đó khi so sánh với các mô hình nhiều lớp, một lớp LSTM đơn có thể không đạt được độ chính xác cao trong các bài toán phức tạp do không thể học các đặc trưng phức tạp và sâu hơn của dữ liệu. Khi sử dụng kết hợp đa lớp, mỗi lớp LSTM có khả năng học và tái tạo các mẫu phức tạp từ dữ liệu đầu vào, từ đó mở rộng không gian biểu diễn của mô hình. Điều này giúp mô hình nắm bắt được những đặc điểm phức tạp của dữ liệu, cải thiện độ chính xác của dự đoán.

Hơn nữa, các lớp LSTM khác nhau có thể học các khía cạnh khác nhau của dữ liệu, tạo ra một mô hình linh hoạt có khả năng nắm bắt các mối quan hệ phức tạp giữa các yếu tố đầu vào.

3.4. GRU

Trong nghiên cứu này, chúng tôi tập trung vào việc xây dựng một mô hình dự đoán giá chứng khoán sử dụng GRU và đánh giá hiệu suất của mô hình trên dữ liệu thực tế. Bảng dưới đây minh họa cấu trúc của mô hình GRU:

Bảng 3.3. Các Layers trong mô hình GRU

Layer (type)	**Output Shape**	**Param #**
gru (GRU)	(None, window_size, 128)	52,992
gru_1 (GRU)	(None, 64)	37,248
dense (Dense)	(None, 25)	1,625
dense_1 (Dense)	(None, 1)	26

Trong đó,

- *window_size*: kích thước của cửa sổ trượt sẽ không cố định là 30 hay 60.

- *gru (GRU):* Đây là lớp GRU đầu tiên trong mô hình, có kích thước đầu ra là (None, 60, 128), trong đó None là kích thước batch, 60 là số lượng thời điểm thời gian và 128 là kích thước của vector đầu ra từ mỗi thời điểm.

- *gru_1 (GRU):* Lớp GRU thứ hai trong mô hình có kích thước đầu ra là (None, 64), giảm kích thước không gian đầu ra từ lớp trước để giảm chiều của dữ liệu đầu vào cho lớp kế tiếp.

- *dense (Dense):* Lớp Dense với kích thước đầu ra (None, 25), áp dụng sau lớp GRU để thực hiện kết hợp thông tin từ các thời điểm thời gian thành một biểu diễn có kích thước nhỏ hơn.

- *dense_1 (Dense):* Lớp Dense cuối cùng với kích thước đầu ra (None, 1), được sử dụng để dự đoán giá chứng khoán tại thời điểm tiếp theo dựa trên biểu diễn được học từ các lớp trước đó.

Nhóm chọn mô hình GRU bởi vì nó mang lại những lợi ích trong việc dự đoán giá chứng khoán với tập dữ liệu trong ngắn hạn. Theo Joseph và M (2022), GRU có khả năng xử lý dữ liệu chuỗi một cách hiệu quả. GRU giúp cải thiện khả năng dự đoán bằng cách hiểu được mối quan hệ thời gian phức tạp hơn giữa các điểm dữ liệu. Bên cạnh đó, mô hình GRU cũng cung cấp hiệu suất tính toán cao hơn so với LSTM, giúp giảm thiểu tình trạng phức tạp tính toán, đặc biệt khi làm việc với dữ liệu lớn, giúp cải thiện tốc độ huấn luyện và tiên đoán của mô hình. Cuối cùng, GRU có khả năng học dài hạn từ dữ liệu chuỗi mà không gặp phải vấn đề triệt tiêu đạo hàm gradient. Đây là một ưu điểm quan trọng trong việc xây dựng mô hình dự đoán giá chứng khoán, vì thị trường tài chính thường biến động phức tạp và đòi hỏi mô hình có khả năng học và điều chỉnh liên tục từ dữ liệu mới. Theo ZEYUAN YU và cộng sự (2019) cũng đã chỉ ra rằng hiệu quả của việc sử dụng mô hình GRU đa lớp tốt hơn so với mô hình GRU đơn lớp. Nghiên cứu của họ cho thấy rằng mặc dù hiệu suất của các mô hình học sâu tương đối không nhạy cảm với sự kết hợp của số lớp và kích thước lớp, việc sử dụng nhiều lớp luôn mang lại hiệu quả tốt hơn so với chỉ một lớp. Điều này được minh chứng qua việc mô hình GRU ba lớp được áp dụng và đã cho kết quả dự báo chính xác hơn đáng kể so với mô hình GRU đơn lớp.

3.5. Các chỉ số đánh giá

Trong lĩnh vực dự báo chuỗi thời gian, việc đánh giá hiệu suất của mô hình là một phần quan trọng để đảm bảo rằng dự đoán được thực hiện có độ chính xác cao và đáng tin cậy. Các chỉ số đánh giá thông thường được sử dụng bao gồm Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), và Mean Absolute Percentage Error (MAPE).

Mean Square Error (MSE): là sự trung bình của bình phương của sai số giữa dự đoán của mô hình và giá trị thực tế. MSE đo lường độ lớn của sai số và càng nhỏ, mô hình càng chính xác. Tuy nhiên, do bình phương sai số, MSE có thể bị ảnh hưởng bởi các giá trị ngoại lệ.

$$\text{MSE} = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (11)$$

Root Mean Square Error (RMSE): RMSE là căn bậc hai của MSE và được sử dụng để đánh giá sai số trung bình giữa dự đoán và giá trị thực tế, với cùng đơn vị đo. RMSE cũng giúp loại bỏ ảnh hưởng của việc bình phương sai số.

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (12)$$

Mean Absolute Error (MAE): MAE là trung bình của giá trị tuyệt đối của sai số giữa dự đoán và giá trị thực tế. MAE đo lường sự chênh lệch trung bình giữa dự đoán và thực tế và thường được sử dụng khi không muốn quan tâm đến hướng của sai số.

$$\text{MAE} = \frac{1}{n}\Sigma_{i=1}^{n}|y_i - \widehat{y_i}| \quad (13)$$

Mean Absolute Percentage Error (MAPE): MAPE tính tỷ lệ phần trăm trung bình của sai số tuyệt đối so với giá trị thực tế. MAPE cung cấp một đánh giá về độ lớn của sai số so với giá trị thực tế và thường được sử dụng khi cần đánh giá mức độ sai lệch của mô hình theo phần trăm.

$$\text{MAPE} = \frac{1}{n}\Sigma_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \times 100\% \quad (14)$$

Trong đó:

- y_i là giá trị thực tế tại thời điểm,
- $\widehat{y_i}$ là giá trị được dự đoán tại thời điểm,
- n là số lượng điểm dữ liệu.

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1. Thu thập dữ liệu

Dữ liệu được thu thập để làm đầu vào cho mô hình dự đoán được chia làm ba loại và được thu thập thông qua các nguồn khác nhau. Các tập dữ liệu này có bước thời gian theo ngày và kéo dài từ ngày 1/1/2014 đến ngày 6/5/2024. Các bảng sau mô tả chi tiết nội dung và tính chất của dữ liệu được sử dụng:

Bảng 4.1. Giá chứng khoán theo thời gian

Cột	Kiểu dữ liệu	Mô tả
time	object	Thời gian
Open	float64	Giá mở cửa
High	float64	Giá cao nhất
Low	float64	Giá thấp nhất
Close	float64	Giá đóng cửa
Adj Close	float64	Giá đóng cửa được điều chỉnh
Volume	float64	Khối lượng giao dịch

Bảng dữ liệu về giá chứng khoán theo thời gian bao gồm tổng cộng 7 đặc trưng. Nhìn chung, các đặc trưng này liên quan đến giá cổ phiếu và khối lượng giao dịch, cung cấp thông tin về giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa và khối lượng giao dịch trong mỗi phiên giao dịch. Điều này đóng góp cho mô hình dự báo chuỗi thời gian bằng cách cung cấp một bộ dữ liệu phong phú và đa dạng, giúp mô hình học được các mẫu và xu hướng trong dữ liệu và tạo ra dự báo chính xác hơn về giá cổ phiếu trong tương lai.

Bảng 4.2 .Tỷ giá USD và VND theo ngày

Cột	Kiểu dữ liệu	Mô tả

time	object	Thời gian
Price	object	Giá
Open	object	Giá mở cửa
High	object	Giá cao nhất
Low	object	Giá thấp nhất
Vol.	object	Khối lượng giao dịch
Change %	object	Phần trăm thay đổi

Dữ liệu về tỉ giá USD và VND theo ngày bao gồm tổng cộng 6 đặc trưng. Các đặc trưng này liên quan đến tỷ giá hối đoái USD/VND và phản ánh các thông tin quan trọng như giá, khối lượng giao dịch và phần trăm thay đổi trong mỗi ngày giao dịch. Điều này đóng góp cho mô hình dự báo chuỗi thời gian bằng cách cung cấp thông tin chi tiết về biến động của tỷ giá hối đoái trong quá khứ, giúp mô hình hiểu và dự đoán giá chứng khoán dựa trên các xu hướng tiềm ẩn và biến động của thị trường ngoại hối USD/VND.

Bảng 4.3. Chỉ số USD INDEX

Cột	**Kiểu dữ liệu**	**Mô tả**
time	object	Thời gian
Open	float64	Giá mở cửa
High	float64	Giá cao nhất
Low	float64	Giá thấp nhất

Close	float64	Giá đóng cửa
Adj Close	float64	Giá đóng cửa được điều chỉnh
Volume	float64	Khối lượng giao dịch

Bảng dữ liệu về chỉ số USD INDEX bao gồm tổng cộng 7 đặc trưng, tập trung vào các thông tin quan trọng liên quan đến chỉ số USD INDEX, là một chỉ số quan trọng trong ngành tài chính. Mô hình dự báo chuỗi thời gian dựa vào xu hướng và biến động của chỉ số này để dự báo giá chứng khoán trên cơ sở thị trường tiền tệ quốc tế có ảnh hưởng đến thị trường chứng khoán.

Bảng 4.4. Chỉ báo kỹ thuật

Cột	**Kiểu dữ liệu**	**Mô tả**
time	datetime64[ns]	Thời gian
open	int64	Giá mở cửa
high	int64	Giá cao nhất
low	int64	Giá thấp nhất
close	int64	Giá đóng cửa
volume	int64	Khối lượng giao dịch
ticker	object	Mã chứng khoán
MA_5	float64	Trung bình di động 5 ngày
RSI_14	float64	Chỉ số sức mạnh tương đối 14 ngày

%K	float64	Giá trị %K của Stochastic Oscillator
%D	float64	Giá trị %D của Stochastic Oscillator
MACD	float64	Phân kỳ hội tụ trung bình động (MACD)
MACD_Signal	float64	Đường tín hiệu MACD
Upper_Band	float64	Giới hạn trên của dải Bollinger
Middle_Band	float64	Đường trung bình của dải Bollinger
Lower_Band	float64	Giới hạn dưới của dải Bollinger
ATR_14	float64	Chỉ số phạm vi thực trung bình 14 ngày (ATR)
OBV	float64	Khối lượng cân bằng theo giá (OBV)
usdindex_close	float64	Giá đóng cửa của chỉ số USD Index
usd_vnd_price	object	Giá USD/VND

Bảng dữ liệu này bao gồm tổng cộng 20 đặc trưng, cung cấp một bộ dữ liệu đa dạng và chi tiết về các chỉ số kỹ thuật và giá cổ phiếu. Dữ liệu này giúp hiểu rõ hơn về tình hình thị trường và là cơ sở quan trọng nhất để dự đoán chính xác hơn về biến động giá cổ phiếu trong tương lai.

Hình 4.1. Trực quan hóa biến động giá đóng cửa của mã VNM

4.2. Tiền xử lý dữ liệu

4.2.1. Phân tích khám phá dữ liệu

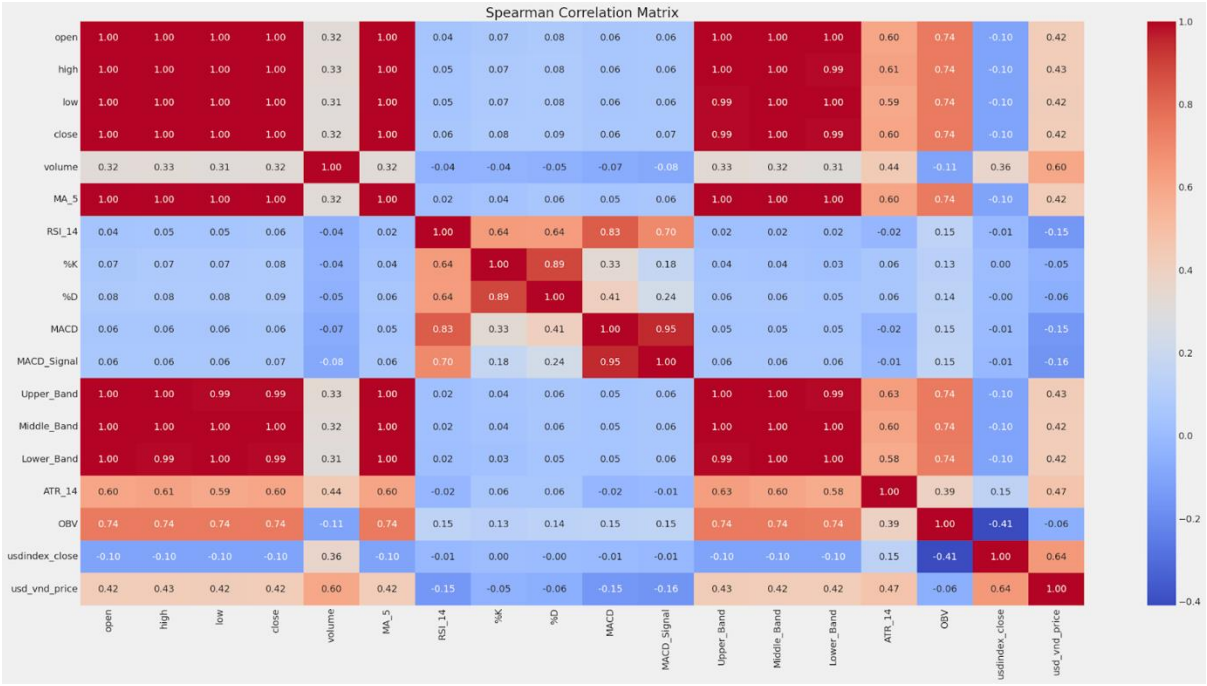
Sau khi thu thập đầy đủ dữ liệu, các bảng dữ liệu được gộp thành một bảng duy nhất thông qua cột ‘time’. Dữ liệu đã gộp thể hiện đầy đủ sự biến động của các yếu tố giá cổ phiếu và tiền tệ theo thời gian từ ngày 1/1/2014 đến ngày 6/5/2024. Để đảm bảo chất lượng và tính toàn vẹn của dữ liệu, nhóm tiến hành xác định tỉ lệ các điểm dữ liệu bị thiếu.

Bảng 4.5. Tỷ lệ giá trị bị thiếu

Tên Cột	Tỷ Lệ Phần Trăm
time	0.00%
open	0.00%
high	0.00%
low	0.00%
close	0.00%
volume	0.00%
ticker	0.00%
MA_5	0.15%
RSI_14	0.54%

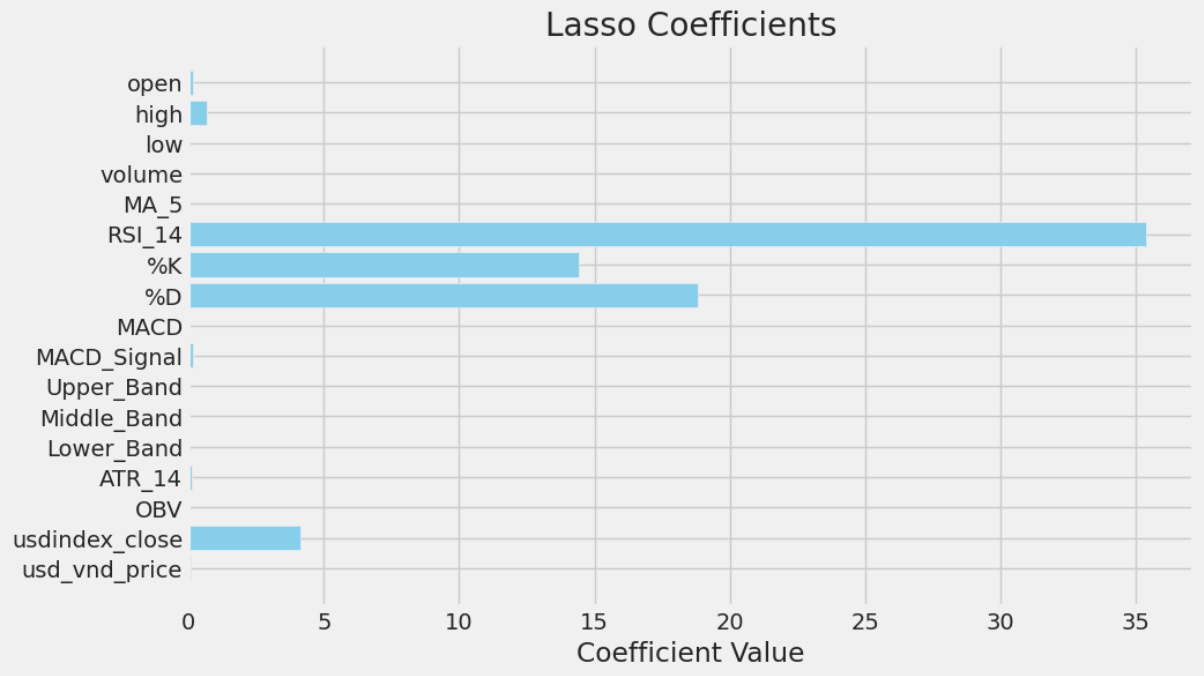
%K	0.31%
%D	0.31%
MACD	1.28%
MACD_Signal	1.28%
Upper_Band	0.15%
Middle_Band	0.15%
Lower_Band	0.15%
ATR_14	0.54%
OBV	0.00%
usdindex_close	0.31%
usd_vnd_price	0.00%

Kết quả cho thấy tất cả các đặc trưng có tỉ lệ thiếu đều dưới 1.5%. Các cột như "time", "open", "high", "low", "close", "volume", và "ticker" đều có tỷ lệ phần trăm là 0%, tức là không có dữ liệu nào bị thiếu trong các cột này, đây là những thông tin cơ bản và quan trọng đối với mô hình. Các chỉ số kỹ thuật như "MA_5", "%K", "%D", "Upper_Band", "Middle_Band", "Lower_Band", "ATR_14", "OBV", và "usd_vnd_price" cũng không có dữ liệu bị thiếu. Do đó, việc xử lý bằng phương pháp thay thế các giá trị này sẽ không gây nhiều ảnh hưởng đến ý nghĩa ban đầu của dữ liệu.



Hình 4.2. Phương pháp Spearman correlation

Phương pháp Spearman correlation là một phương pháp thống kê để đo lường mức độ tương quan giữa hai biến ngẫu nhiên. Nó được sử dụng để xác định mối quan hệ đồng biến hoặc nghịch biến giữa các biến mà không cần phải giả định về phân phối của chúng. Thay vì sử dụng giá trị số, phương pháp này dựa trên việc sắp xếp thứ tự của dữ liệu, sau đó tính toán hệ số tương quan dựa trên sự tương đồng giữa các vị trí thứ tự của các điểm dữ liệu. Để hiểu rõ hơn mối quan hệ giữa các đặc trưng, nhóm phân tích ở hai khía cạnh là mức độ tương quan của các biến với biến mục tiêu ‘close’ và mức độ tương quan giữa các biến với nhau. Kết quả cho thấy các biến ‘open’, ‘high’, ‘low’, ‘upper_band’, ‘middle_band’, ‘lower_band’ và ‘OBV’ có mức tương quan cao với biến mục tiêu khi có điểm tương quan cao hơn 0.75. Ngoài ra, các biến thể hiện giá như ‘open’, ‘high’, ‘low’ có mức độ tương quan cao với các chỉ số kỹ thuật. Điều này có thể giải thích thông qua việc các chỉ số kỹ thuật được trích xuất từ dữ liệu lịch sử của giá ‘close’, do đó cũng tương quan với các biến ‘open’, ‘high’, ‘low’.



Hình 4.3. Phương pháp Lasso Coefficients

Kết quả hệ số tương quan (Coefficients) thu được từ việc thực hiện hồi quy Lasso trên tập dữ liệu với "close" là biến mục tiêu cho thấy các biến có ảnh hưởng đến kết quả hồi quy là "RSI_14", "%K", "%D", "MACD_Signal" và "usdindex_close". Do đó, các biến này sẽ được sử dụng làm đầu vào cho mô hình dự báo chuỗi thời gian được đề xuất. Ngoài ra, kết quả này còn cho thấy một phát hiện là biến "usdindex_close", đại diện cho yếu tố thị trường tiền tệ, có ảnh hưởng đáng kể đáng giá cổ phiếu và có thể được sử dụng làm đầu vào cho mô hình dự báo giá cổ phiếu.

4.2.2. Làm sạch dữ liệu

Trong quá trình tiền xử lý dữ liệu, các giá trị thiếu trong một số trường dữ liệu như MA_5, RSI_14, %K, %D, MACD, MACD_Signal, Upper_Band, Middle_Band, Lower_Band, ATR_14 và usdindex_close đã được thay thế bằng cách sử dụng phương pháp điền trước (Forward Fill) và điền sau (Backward Fill). Kết quả là các giá trị thiếu trong các trường dữ liệu đã được thay thế bằng giá trị của điểm thời gian đứng trước hoặc sau nó. Điều này giúp làm đầy các giá trị thiếu và duy trì tính liên tục của dữ liệu, giúp đảm bảo tính chính xác và đáng tin cậy trong quá trình phân tích và xây dựng mô hình.

Tiếp theo, Min-Max Scaler được sử dụng để chuẩn hóa dữ liệu trong biến dataset. Phương pháp này giữ nguyên các tỷ lệ tương đối giữa các điểm dữ liệu, đồng thời giảm

ảnh hưởng của nhiễu và giúp mô hình hoạt động tốt hơn trên dữ liệu thực tế. Cụ thể, Min-Max Scaler sẽ chuyển đổi các giá trị trong dataset sao cho chúng nằm trong khoảng từ 0 đến 1, bằng cách sử dụng công thức:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (15)$$

Trong đó:

- x: là giá trị ban đầu của mỗi điểm dữ liệu,
- x_{scaled}: là giá trị sau khi được chuẩn hóa,
- x_{min}: là giá trị nhỏ nhất của mỗi đặc trưng trong dataset,
- x_{max}: là giá trị lớn nhất của mỗi đặc trưng trong dataset.

Kết quả cuối cùng là một mảng chứa dữ liệu đã được chuẩn hóa với các giá trị nằm trong khoảng từ 0 đến 1. Điều này giúp đồng nhất hóa phạm vi của các biến, làm cho dữ liệu dễ dàng đồng nhất và phù hợp cho việc sử dụng trong các mô hình máy học, đặc biệt là các mô hình dự đoán.

4.3. Xây dựng mô hình dự báo

Sau khi tiền xử lý dữ liệu, thu được các tập dữ liệu huấn luyện và kiểm thử, nhóm tiến hành xây dựng mô hình dự báo chuỗi thời gian. Hai mô hình được sử dụng là LSTM và GRU nhằm mục đích so sánh và đánh giá toàn diện cho hướng tiếp cận dự báo ngắn hạn này. Kiến trúc dùng để xây dựng mô hình được trình bày ở phần *3.3.* cho LSTM và *3.4.* cho GRU. Nhằm mục đích so sánh hiệu quả khi xây dựng mô hình với tất cả đặc trưng và mô hình với chỉ những đặc trưng được chọn, nhóm đã xây dựng mô hình trên nhiều không gian đặc trưng khác nhau. Các đặc trưng được chọn (selected features) là các đặc trưng có mức độ tương quan cao với biến mục tiêu gồm các đặc trưng sau:

Bảng 4.6. Selected features

Tên cột	**Ý nghĩa**
open	Giá mở cửa.

high	Giá cao nhất.
close	Giá đóng cửa.
low	Giá thấp nhất.
upper_band	Đường biên trên.
middle_band	Đường giữa.
lower_band	Đường biên dưới.
OBV	On-Balance Volume (OBV) là một chỉ báo kỹ thuật được sử dụng để đo lượng mua vào hoặc bán ra của một cổ phiếu, có thể được sử dụng để dự đoán hướng giá cổ phiếu.

Ngoài ra, mô hình dự đoán sử dụng dữ liệu với kích thước cửa sổ, tương đương độ dài chuỗi thời gian cũng được so sánh. Tham số được sử dụng để huấn luyện cả hai mô hình LSTM và GRU được sử dụng như sau:

Bảng 4.7. Các tham số huấn luyện trong LSTM và GRU

Tham số	**Giá trị**
Batch_size	64
Epochs	100
Optimizer	Adam

Loss	mean_squared_error

Bảng 4.8. Kết quả đánh giá các mô hình

Mô hình	**RMSE**	**MAE**	**MAPE**
LSTM (all features, window size 60)	1287.241	981.640	1.408
LSTM (selected features, window size 60)	1009.716	751.876	1.090
GRU (selected features, window size 60)	999.616	747.897	1.082
LSTM (selected features, window size 30)	4643.303	4078.389	5.975
GRU (selected features, window size 30)	1234.882	998.375	1.453

Ba mô hình LSTM và GRU đã được đánh giá dựa trên ba chỉ số chính: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), và Mean Absolute Percentage Error (MAPE). Mô hình LSTM sử dụng tất cả các đặc trưng (LSTM - all features) có RMSE cao nhất và MAE và MAPE lớn hơn so với các mô hình khác, cho thấy việc sử dụng quá nhiều đặc trưng có thể dẫn đến overfitting. Trong khi đó, mô hình LSTM và GRU với 8 đặc trưng được lựa chọn (LSTM - selected features) đều cho thấy sự cải thiện đáng kể về RMSE, MAE, và MAPE so với mô hình sử dụng tất cả các đặc trưng. Kết quả này cho thấy rằng việc lựa chọn các đặc trưng quan trọng giúp cải thiện hiệu suất dự đoán của cả hai mô hình. Từ kết quả này, có thể kết luận rằng việc sử dụng các đặc trưng tương quan nhiều với biến mục tiêu giúp cải thiện hiệu suất của mô hình dự báo chuỗi thời gian và giảm thiểu overfitting, kể cả khi sự tương quan này chưa thể hiện được mối quan hệ theo thời gian. Điều này cung cấp một cơ sở quan trọng cho việc áp dụng các mô hình học máy trong dự báo thị trường tài chính và các lĩnh vực khác có liên quan.

Mặc dù có một số chênh lệch nhỏ giữa mô hình LSTM và GRU với các đặc trưng được lựa chọn, nhưng cả hai đều cho thấy khả năng dự đoán tốt và hiệu suất gần như nhau trên tập dữ liệu này. Tuy nhiên, mô hình GRU có vẻ có hiệu suất nhỉnh hơn so với mô hình LSTM trong trường hợp này, nhưng chênh lệch này không đáng kể. Ngoài ra, khi giảm cửa sổ từ 60 xuống 30, chúng ta thấy mô hình LSTM sử dụng các đặc trưng đã lựa chọn cho kết quả dự báo kém hơn đáng kể, với RMSE, MAE và MAPE tăng lên đáng kể so với khi cửa sổ là 60. Trong khi đó, mô hình GRU vẫn giữ được hiệu suất tốt hơn, với giảm nhẹ về RMSE, MAE và MAPE so với khi cửa sổ là 60. Điều này có thể cho thấy rằng mô hình GRU có khả năng xử lý tốt hơn với các cửa sổ nhỏ hơn, trong khi mô hình LSTM có thể cần một cửa sổ lớn hơn để có kết quả tốt.

Tóm lại, việc sử dụng các đặc trưng được lựa chọn và cân nhắc kích thước của cửa sổ có thể ảnh hưởng đến hiệu suất của mô hình dự báo, và việc lựa chọn mô hình phù hợp với bài toán cụ thể là rất quan trọng. Với kết quả tốt nhất trong tất cả bối cảnh thực nghiệm, nhóm chọn GRU là mô hình tối ưu nhất cho bài toán dự báo chuỗi thời gian ngắn hạn được đặt ra.

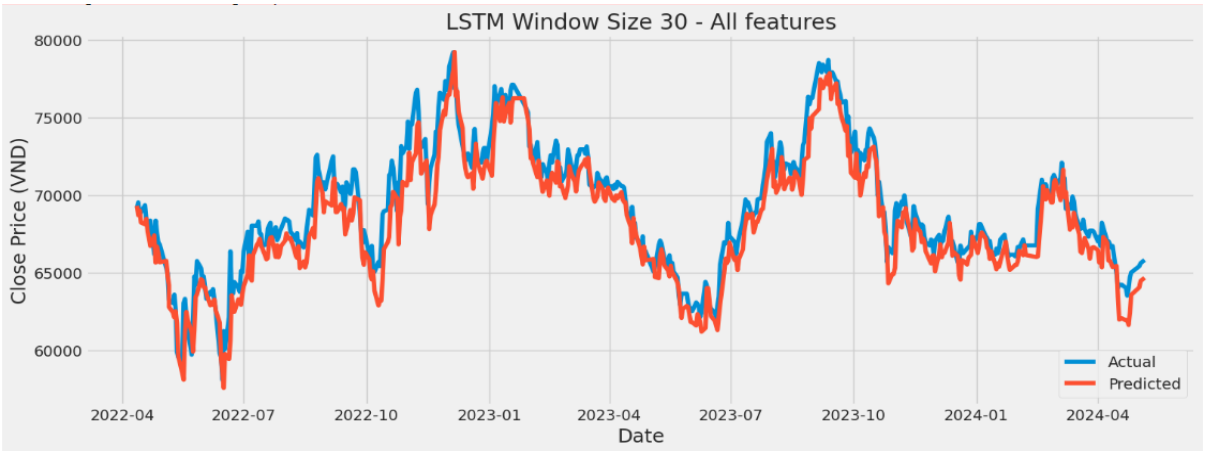
Bảng 4.9. Đánh giá kết quả dự đoán của mô hình GRU theo từng cửa sổ kích thước 60

Cửa sổ	**RMSE**	**MAE**	**MAPE**
1	484.504	484.504	0.701
2	508.160	508.160	0.731
3	701.083	701.083	1.020
4	175.182	175.182	0.254
…	…	…	…
513	1089.012	1089.012	1.683

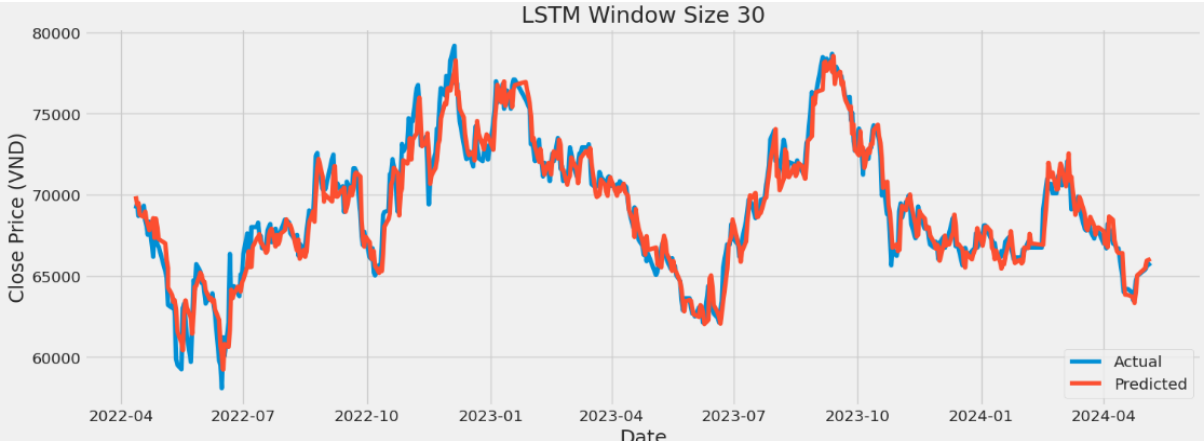
514	524.729	524.729	0.807
515	534.853	534.853	0.818
516	474.399	474.399	0.723
517	460.981	460.981	0.701

Có thể nhận thấy rằng kết quả dự đoán của mô hình GRU theo từng cửa sổ kích thước 60 có sự biến động đáng kể. Trong các cửa sổ đầu tiên, đặc biệt là cửa sổ 3, mô hình GRU cho thấy mức độ lỗi cao, với RMSE và MAE đều ở mức cao và MAPE lên đến hơn 1. Điều này cho thấy mô hình gặp khó khăn trong việc dự đoán đúng trong một số khoảng thời gian, có thể do sự biến động lớn hoặc sự phức tạp của dữ liệu. Tuy nhiên, trong một số thời điểm, mô hình GRU cho thấy hiệu quả đáng kể. Cụ thể, các giá trị RMSE, MAE, và MAPE thấp đáng kể, ví dụ là ở cửa sổ 4 và cửa sổ 517, nơi mà mô hình có hiệu suất dự đoán tốt, với RMSE và MAE thấp và MAPE dưới 0.8%. Tuy nhiên, cần lưu ý rằng có một số cửa sổ, ví dụ như cửa sổ 3 và cửa sổ 513 vẫn cho thấy mức độ lỗi cao, có thể cần được xem xét kỹ lưỡng để hiểu rõ hơn về nguyên nhân gây ra sự biến động trong kết quả dự đoán của mô hình GRU.

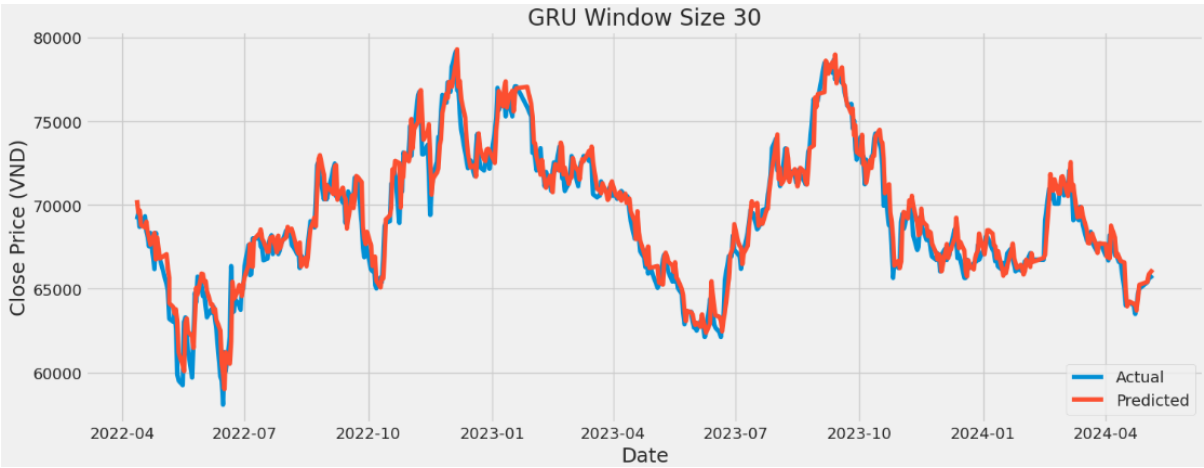
4.4. Thảo luận



Hình 4.4. Trực quan hóa kết quả dự báo của mô hình LSTM sử dụng tất cả đặc trưng



Hình 4.5. Trực quan hóa kết quả dự báo của mô hình LSTM sử dụng các đặc trưng quan trọng



Hình 4.6. Trực quan hóa kết quả dự báo của mô hình GRU sử dụng các đặc trưng quan trọng

Trực quan hóa kết quả dự báo cho thấy sự chênh lệch rõ ràng trong độ chính xác của mô hình sử dụng tất cả đặc trưng và mô hình chỉ sử dụng các đặc trưng quan trọng. Cụ thể, mô hình LSTM sử dụng tất cả đặc trưng có độ lệch giữa giá trị dự đoán và giá trị thực tế khá cao. Trong khi đó, hai mô hình sử dụng các đặc trưng quan trọng thể hiện sự biến động rất chính xác so với giá thực tế, với giá trị RMSE và MAE thấp hơn đáng kể và tỷ lệ phần trăm sai cũng được giảm xuống. Điều này cho thấy việc lựa chọn đặc trưng quan trọng có thể cải thiện đáng kể hiệu suất của mô hình dự báo chuỗi thời gian.

Bảng 4.10. Thống kê mô tả độ lệch trên các cửa sổ trượt

	RMSE	**MAE**	**MAPE (%)**

Mean	629.51	629.51	0.91
Median	570.96	570.96	0.82
Max	2597.73	2597.73	3.76
Min	39.35	39.35	0.06
Std	528.85	528.85	0.76

Ngoài ra, việc đánh giá các chỉ số thống kê của độ lệch mô hình trên các cửa sổ khác nhau giúp nhận định được tính ổn định của mô hình theo thời gian và xác định các thời điểm dự đoán không hiệu quả nhằm thực hiện phân tích chuyên sâu hơn. Việc này giúp tìm ra nguyên nhân ảnh hưởng đến kết quả dự đoán trong các thời điểm cụ thể. Trong đó, cửa sổ thứ 465 có chỉ số MAPE lớn nhất là 3.76, cửa sổ thứ 473 có chỉ số MAPE nhỏ nhất là 0.06. Độ lệch chuẩn của các chỉ số đánh giá cho các cửa sổ trong tập dữ liệu thể hiện một mức độ phân tán tương đối nhỏ, đặc biệt khi so sánh với giá trị của cổ phiếu, có thể đạt đến hàng chục hoặc thậm chí hàng trăm nghìn đồng. Với các giá trị độ lệch chuẩn chỉ trong khoảng vài trăm đồng, điều này cho thấy sự ổn định và đồng đều trong chất lượng của các mô hình dự báo trên các sliding window. Sự ít biến động này cũng có thể cho thấy rằng các mô hình có thể đang cung cấp dự báo ổn định và nhất quán, ít bị ảnh hưởng bởi các biến thể lớn trong dữ liệu đầu vào. Điều này có thể cung cấp niềm tin cho việc sử dụng các mô hình dự báo ngắn hạn và liên tục trong việc đánh giá và dự báo giá cổ phiếu trong tương lai.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trên cơ sở những nghiên cứu về dự báo giá cổ phiếu dựa trên dữ liệu chuỗi thời gian có sử dụng các phương pháp học sâu mà nhóm đã thực hiện khảo lược như Shahi và cộng sự (2020), Sen và cộng sự (2021), Vuong và cộng sự (2022), Kumar và cộng sự (2018), nhóm nhận thấy việc phân tích dự đoán giá chứng khoán sử dụng đa bước sẽ dẫn đến độ chính xác không cao trong ngữ cảnh thị trường biến động liên tục như thị trường cổ phiếu. Ngoài ra, nhiều nghiên cứu chỉ sử dụng một hoặc một ít biến mà chưa cân nhắc đến sự ảnh hưởng của nhiều yếu tố cũng khiến cho kết quả dự đoán có sai số tương đối lớn so với thực tế.

Ở khía cạnh tích cực, những nghiên cứu trên đã cho thấy được sự hiệu quả của các phương pháp học sâu đối với bài toán dự báo giá cổ phiếu. Thêm vào đó, Bhandari và cộng sự (2022) cũng chỉ ra rằng việc sử dụng nhiều lớp khi chạy mô hình LSTM sẽ tốt hơn là chỉ một lớp.

Xuất phát từ những vấn đề và các hướng tiếp cận đã được chứng minh độ hiệu quả trên, nhóm đã thực hiện đề tài *"Ứng dụng LSTM và GRU trong dự báo giá cổ phiếu ngắn hạn"* tập trung vào việc dự đoán giá cổ phiếu ngắn hạn bằng cách thử nghiệm hai mô hình học sâu LSTM và GRU. Trong đề tài này, nhóm đã ứng dụng phương pháp "dự báo trượt" kết hợp với nhiều biến có liên quan như các yếu tố về tiền tệ hoặc chỉ số kỹ thuật của cổ phiếu. Qua thực nghiệm, mô hình đã mang lại kết quả dự đoán tốt hơn cho giá cổ phiếu ngắn hạn.

Từ việc sử dụng mô hình LSTM và GRU kết hợp dự báo trượt, nhóm nhận thấy cả hai mô hình đều có hiệu suất tốt trong việc dự đoán giá cổ phiếu. Khi áp dụng mô hình LSTM và GRU với tất cả các đặc trưng, kết quả đánh giá cho thấy mô hình LSTM và GRU đạt được RMSE cao nhất tương ứng là 2597.73 và 1234.88. Tuy nhiên, khi áp dụng các đặc trưng quan trọng bao gồm giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa và các chỉ báo kỹ thuật, kết quả cải thiện đáng kể với mô hình GRU, khi nó đạt được RMSE là 999.62, thấp hơn mô hình LSTM với RMSE là 1287.24. Điều này cho thấy khả năng ưu việt của mô hình GRU trong việc dự đoán giá cổ phiếu.

Được thực hiện trên tập dữ liệu phức tạp và biến động, mô hình LSTM và GRU đã chứng minh khả năng xử lý dữ liệu chuỗi thời gian trong lĩnh vực dự báo giá cổ

phiếu. Điều này được chứng minh qua sự giảm thiểu lỗi, với giá trị MAPE thấp nhất chỉ là 1.08%. Kết quả này cho thấy mô hình có khả năng dự đoán chính xác và đáng tin cậy về giá cổ phiếu trong ngắn hạn.

Những kết quả đánh giá và so sánh hai mô hình LSTM và GRU đã chứng minh khả năng dự đoán tốt và hiệu suất gần như nhau của hai mô hình trên tập dữ liệu thực nghiệm. Tuy nhiên, về tính ổn định khi giảm số lượng cửa sổ trượt, GRU lại chiếm ưu thế lớn hơn so với LSTM. Kết quả này mang lại một giá trị lớn trong việc dự báo giá cổ phiếu ngắn hạn và giúp nhà đầu tư và nhà phân tích tài chính đưa ra quyết định thông minh và giảm thiểu rủi ro trong giao dịch cổ phiếu. Thêm vào đó, thông qua quá trình lựa chọn đặc trưng, nghiên cứu cũng đã cho thấy được các chỉ số kỹ thuật của cổ phiếu có ảnh hưởng nhiều nhất đến với sự biến động giá cổ phiếu trên thị trường.

5.2. Hạn chế

Trong quá trình thực hiện nghiên cứu dự đoán giá cổ phiếu ngắn hạn bằng mô hình LSTM và GRU, nhóm nhận thấy mô hình có một số hạn chế nhất định cần được lưu ý để đảm bảo tính chính xác và độ tin cậy của kết quả.

Phần đầu tiên, hạn chế về tính hiệu quả của mô hình cần được xem xét thêm. Mặc dù hai mô hình LSTM và GRU đã được chứng minh là hiệu quả trong việc dự đoán chuỗi thời gian và giá cổ phiếu thông qua thực nghiệm do nhóm tiến hành, tuy nhiên, nó không phù hợp để áp dụng trong mọi tình huống thị trường. Hiệu suất của mô hình có thể bị ảnh hưởng trong các điều kiện thị trường biến động mạnh hoặc không tuân thủ các tiêu chuẩn cụ thể của thị trường.

Hạn chế tiếp theo đề cập đến việc nghiên cứu chưa thực nghiệm trên đa dạng kích thước cửa sổ khác nhau. Khi bỏ qua vấn đề này, nghiên cứu có thể bỏ sót những trường hợp đặc biệt mà tại đó kết quả của mô hình sẽ có sự khác biệt so với những kích thước cửa sổ khác. Từ đó, tính ổn định của mô hình cũng khó được đánh giá khách quan.

Kế đến, các hạn chế liên quan đến dữ liệu cần được lưu ý hơn. Chất lượng và tính sẵn có của dữ liệu đầu vào đã tác động đáng kể đến hiệu suất của mô hình. Sự đủ hoặc thiếu thông tin, tính đại diện và tính chính xác của dữ liệu có thể gây ảnh hưởng tiêu cực đến quá trình học và dự đoán của mô hình.

Một vấn đề khác, đề tài chưa thử nghiệm dự đoán nhiều điểm thời gian trong tương lai (ví dụ: giá cổ phiếu trong 5 ngày sắp tới). Nếu thực hiện được điều này, nghiên cứu sẽ càng chứng minh được độ chính xác của mô hình khi dự đoán giá cổ phiếu trong ngắn hạn.

Cuối cùng, hạn chế về khả năng dự đoán trong tương lai cũng cần được nhấn mạnh. Mô hình do nhóm đề xuất trong phạm vi đề tài này chỉ dự đoán dựa trên quá khứ. Ngoại trừ các yếu tố đã được lưu trữ trong dữ liệu quá khứ, mô hình không đoán trước được các yếu tố có khả năng gây chi phối trong tương lai, chẳng hạn như tin tức thị trường, những thay đổi trong chính sách của doanh nghiệp, nhà nước hoặc sự kiện không lường trước. Do đó, việc kết hợp các yếu tố này vào mô hình có thể tăng tính chính xác và độ tin cậy của dự đoán.

Tổng kết lại, trong quá trình thực hiện nghiên cứu về dự đoán giá cổ phiếu ngắn hạn bằng mô hình LSTM và GRU có sử dụng cửa sổ trượt, nhóm đã nhận thấy những hạn chế quan trọng cần được lưu ý ảnh hưởng đến khả năng dự đoán giá cổ phiếu trong ngắn hạn của mô hình, đòi hỏi sự cân nhắc kỹ lưỡng và phân tích chi tiết để đảm bảo tính chính xác, tính ổn định và độ tin cậy của kết quả.

5.3. Hướng phát triển

Thông qua quá trình thực nghiệm của đề tài, nhóm đã đưa ra được những kết quả tích cực thể hiện sự hiệu quả của mô hình LSTM, GRU cũng như xác định được những hạn chế gặp phải trong nghiên cứu. Ở bước tiếp theo, nhóm đề xuất các hướng phát triển tiềm năng cho đề tài này trong tương lai.

Một hướng phát triển quan trọng là mở rộng dữ liệu và tích hợp các yếu tố mới. Hiện nay, nhóm đã sử dụng các đặc trưng cơ bản như giá cổ phiếu và khối lượng giao dịch trong dữ liệu chuỗi thời gian. Tuy nhiên, để tăng tính chính xác và độ tin cậy của dự đoán, sự tích hợp thêm các chỉ số tài chính khác như doanh thu, lợi nhuận, dòng tiền hay các yếu tố kỹ thuật như biểu đồ giá và các chỉ báo kỹ thuật trong đánh giá giá cổ phiếu cần được xem xét để tăng hiệu quả mô hình. Sự bổ sung này sẽ cung cấp một cái nhìn toàn diện hơn về thông tin thị trường và cung cấp cơ sở để xây dựng mô hình dự đoán.

Ngoài ra, nghiên cứu về các mô hình và phương pháp học máy tiên tiến khác cũng là một hướng phát triển tiềm năng. Hiện nay, rất nhiều mô hình khác như mô hình

Transformer, mạng nơ-ron đa tầng, mạng học sâu đa tầng, và mạng nơ-ron tích chập có thể được khám phá và ứng dụng vào quá trình dự đoán. Việc nghiên cứu và so sánh các mô hình này sẽ giúp xác định phương pháp phù hợp nhất cho dự đoán giá cổ phiếu trong ngắn hạn.

Ở khía cạnh ứng dụng nghiên cứu vào quy trình vận hành thực tế, quản lý rủi ro cũng đóng vai trò quan trọng trong quá trình dự đoán giá cổ phiếu và là một hướng phát triển khả thi. Những nghiên cứu trong tương lai có thể xem xét thêm về các phương pháp đánh giá và quản lý rủi ro. Ví dụ, có thể tìm hiểu về phương pháp tạo ra các ngưỡng rủi ro hoặc xây dựng mô hình tự động để hỗ trợ quyết định đầu tư dựa trên các kết quả dự đoán.

Cuối cùng, trong quá trình chọn lọc đặc trưng, nghiên cứu đã cho thấy chỉ số về tiền tệ như USD_index có tác động đáng kể và là một trong những đặc trưng quan trọng của mô hình. Dựa vào kết quả này, dữ liệu của các đặc trưng về tiền tệ có thể được xem xét tích hợp thêm vào mô hình trong tương lai nhằm đưa ra được đánh giá một cách khách quan hơn và đầy đủ hơn. Việc tích hợp các yếu tố này sẽ cung cấp cái nhìn đa chiều và phong phú hơn về ảnh hưởng của tiền tệ đến giá cổ phiếu và giúp cải thiện tính chính xác của dự đoán.

Qua quá trình thực hiện nghiên cứu, nhóm đã xây dựng được mô hình phân tích dự đoán giá cổ phiếu trong ngắn hạn hiệu quả và nhận thấy được những hạn chế còn tồn đọng trong đề tài. Vì vậy, những đề xuất cho các hướng phát triển tiềm năng trong tương lai đã được đưa ra dựa trên thực trạng dữ liệu và kỹ thuật hiện có. Những hướng phát triển này hứa hẹn đóng góp vào việc nâng cao tính chính xác và tin cậy của dự đoán giá cổ phiếu trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Nguyen, T. T., & Yoon, S. (2019). A novel approach to short-term stock price movement prediction using transfer learning. Applied Sciences, 9(22), 4745.

[2] Tashiro, D., Matsushima, H., Izumi, K., & Sakaji, H. (2019). Encoding of high-frequency order information and prediction of short-term stock price by deep learning. Quantitative Finance, 19(9), 1499-1506.

[3] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A, 379(2194), 20200209.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[5] Graves, A., Generating sequences with recurrent neural networks, arXiv preprint. (2013)arXiv:1308.0850.

[6] Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691.

[7] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 11106-11115).

[8] Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. Artificial Intelligence Review, 52(3), 2019-2037.

[9] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied soft computing, 90, 106181.

[10] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).

[11] Lomakin, N., Kulachinskaya, A., Maramygin, M., & Chernaya, E. (2022). Improving Accuracy and Reducing Financial Risk When Forecasting Time Series of

SIU0 Future Contracts Employing Neural Network with Word2vec Vector News. In Complex Systems: Spanning Control and Computational Cybernetics: Applications: Dedicated to Professor Georgi M. Dimirovski on his Anniversary (pp. 281-298). Cham: Springer International Publishing.

[12] Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023, June). Are transformers effective for time series forecasting?. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 9, pp. 11121-11128).

[13] He, K., Zheng, L., Yang, Q., Wu, C., Yu, Y., & Zou, Y. (2023). Crude oil price prediction using temporal fusion transformer model. *Procedia Computer Science*, *221*, 927-932.

[14] Shahi, T. B., Shrestha, A., Neupane, A., & Guo, W. (2020). Stock price forecasting with deep learning: A comparative study. Mathematics, 8(9), 1441.

[15] Sen, J., & Mehtab, S. (2021, June). Accurate stock price forecasting using robust and optimized deep learning models. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-9). IEEE.

[16] Mehtab, S., & Sen, J. (2022). Analysis and forecasting of financial time series using CNN and LSTM-based deep learning models. In Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2021 (pp. 405-423). Springer Singapore.

[17] Vuong, P. H., Dat, T. T., Mai, T. K., & Uyen, P. H. (2022). Stock-price forecasting based on XGBoost and LSTM. Computer Systems Science & Engineering, 40(1).

[18] Alberg, D., & Last, M. (2018). Short-term load forecasting in smart meters with sliding window-based ARIMA algorithms. Vietnam Journal of Computer Science, 5, 241-249.

[19] Li, L., Dai, S., & Cao, Z. (2019, August). Deep long short-term memory (lstm) network with sliding-window approach in urban thermal analysis. In 2019 IEEE/CIC International Conference on Communications Workshops in China (ICCC Workshops) (pp. 222-227). IEEE.

[20] Tashman, L. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review International Journal of Forecasting, 16(4), 437-450.

[21] Salman, A. G., Heryadi, Y., Abdurahman, E., & Suparta, W. (2018). Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. Procedia Computer Science, 135, 89-98.
[22] vnstock. (2024, April 2). PyPI. https://pypi.org/project/vnstock/.
[23] Giá USD VND hôm nay | Giá đô hôm nay - Investing.com. (n.d.). Investing.com Việt Nam. https://vn.investing.com/currencies/usd-vnd.
[24] Joseph, M. (2022). Modern Time Series Forecasting with Python: Explore industry-ready time series forecasting using modern machine learning and deep learning. Packt Publishing Ltd.
[25] Kumar, S., & Ningombam, D. (2018, December). Short-term forecasting of stock prices using long short term memory. In 2018 International Conference on Information Technology (ICIT) (pp. 182-186). IEEE.
[26] Qiu, J., Wang, B., & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. PloS one, 15(1), e0227222.
[27] Yıldırım, D. C., Toroslu, I. H., & Fiore, U. (2021). Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. Financial Innovation, 7, 1-36.
[28] Gao, Y., Wang, R., & Zhou, E. (2021). Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021, 1-8.
[29] Lộc, T. Đ. (2014). CÁC NHÂN TỐ ẢNH HƯỞNG ĐẾN SỰ THAY ĐỔI GIÁ CỦA CỔ PHIẾU: CÁC BẰNG CHỨNG TỪ SỞ GIAO DỊCH CHỨNG KHOÁN THÀNH PHỐ HỒ CHÍ MINH. Tạp chí Khoa học Đại học cần Thơ, (33), 72-78.
[30] NGUYỄN THỊ, H. O. A. (2020). CÁC YẾU TỐ ĐẶC ĐIỂM CỦA DOANH NGHIỆP ẢNH HƯỞNG ĐẾN LỰA CHỌN CÔNG TY KIỂM TOÁN CỦA CÁC DOANH NGHIỆP Ở VIỆT NAM.
[31] Tuyên, T. Đ. (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. Tạp chí Khoa học Đại học Cần Thơ.
[32] Nguyễn, P. Đ. (2012). Các yếu tố ảnh hưởng đến tỷ suất sinh lời của cổ phiếu niêm yết trên sàn chứng khoán TP. HCM. Công nghệ ngân hàng.

[33] Phan, L. N. T. (2022). Ảnh hưởng của các nhóm chỉ tiêu phản ánh tình hình tài chính doanh nghiệp đến quyết định của nhà đầu tư cá nhân trên sàn chứng khoán. Truy cập ngày 15/05/2024 tại: http://thuvienlamdong.org.vn:81/bitstream/DL_134679/54708/1/CVv146S42022325.pdf.