

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN THẾ ANH

**PHÂN TÍCH XU HƯỚNG GIÁ CỔ PHIẾU
DỰA TRÊN CÁC KỸ THUẬT HỌC SÂU**

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS. TRẦN ĐÌNH QUẾ

HÀ NỘI-2025

Đề án được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS. TRẦN ĐÌNH QUẾ**

Phản biện 1: PGS.TS. Nguyễn Hà Nam

Phản biện 2: TS. Trần Đăng Công

Đề án đã được bảo vệ trước Hội đồng chấm đề án thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 11h ngày 19 tháng 07 năm 2025

Có thể tìm hiểu đề án tại:

Thư viện của Học viện Công nghệ Bưu chính Viễn Thông

MỞ ĐẦU

Thị trường chứng khoán là một trong những lĩnh vực lớn được các nhà đầu tư quan tâm, dự đoán xu hướng giá cổ phiếu trên thị trường chứng khoán luôn là chủ đề nóng, một bài toán không dễ đối với các nhà nghiên cứu ở cả lĩnh vực tài chính và kỹ thuật. Đã có nhiều bài báo nghiên cứu về chủ đề với dữ liệu cổ phiếu ở những thị trường với nền kinh tế lớn như Mỹ, Anh, Trung Quốc. Những năm gần đây, thị trường chứng khoán Việt Nam bùng nổ với số tài khoản mở mới và về cả khối lượng giao dịch. Đặc biệt trong bối cảnh nền kinh tế nước nhà đang không ngừng phát triển và triển vọng nâng hạng của thị trường, qua đó thu hút thêm nhiều sự quan tâm đối với thị trường chứng khoán và đặc biệt là cổ phiếu. Do đó, đề án mong muốn đóng góp một phần nhỏ vào hướng nghiên cứu này.

Đề án tập trung nghiên cứu về mối liên hệ giữa giá cổ phiếu trong tương lai với những chỉ báo kỹ thuật của cổ phiếu đó thời điểm hiện tại và quá khứ. Sử dụng kỹ thuật học sâu như RNN, LSTM. Qua đó, xác định mối liên quan, mức độ liên quan giữa giá cổ phiếu và các chỉ báo kỹ thuật và sự phù hợp của các mô hình học sâu đối với bài toán dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam.

Đề án gồm ba chương:

- **Chương 1: Bài toán phân tích giá cổ phiếu**

Trong chương này, đề án sẽ trình bày bài toán phân tích xu hướng giá cổ phiếu, giới thiệu các chỉ báo kỹ thuật và xem xét các nghiên cứu đi trước của bài toán phân tích xu hướng giá cổ phiếu. Các chương tiếp theo sẽ đi sâu vào các khía cạnh cơ sở lý thuyết, kỹ thuật và thực nghiệm của giải pháp, với mục đích đưa ra một cách tiếp cận hiệu quả cho vấn đề.

- **Chương 2: Giới thiệu các mô hình học sâu cho bài toán**

Chương này sẽ cung cấp một cái nhìn tổng quan toàn diện về các kỹ thuật học sâu được sử dụng trong đề án. Phần đầu sẽ giới thiệu tổng quan về RNN, mạng nơron nhân tạo chuyên xử lý cho dữ liệu dạng chuỗi thời gian. Sau đó, đề án sẽ tìm hiểu về kiến trúc của mạng nơron LSTM, là sự cải tiến từ mạng nơron RNN, giải quyết vấn đề biến mất và bùng nổ độ dốc (gradients). Chương tiếp theo sẽ đi sâu vào việc triển khai và đánh giá giải pháp này, bao gồm cả việc tối ưu mô hình để giải quyết những thách thức cụ thể trong quá trình nghiên cứu của đề án.

- **Chương 3: Thực nghiệm và đánh giá**

Chương này đề án trình bày khung thử nghiệm cho giải pháp, từ thu thập dữ liệu cổ phiếu Việt Nam, tiền xử lý và chuẩn hóa dữ liệu sau đó là đưa vào huấn luyện các mô hình đã trình bày ở Chương 2. Sau đó là đánh giá, thử nghiệm so sánh các biến thể của mô hình. Kết quả cho thấy tính hiệu quả của việc ứng dụng các kỹ thuật học sâu trong

việc dự báo xu hướng giá cổ phiếu có sử dụng thêm cả các chỉ báo kỹ thuật. Những phát hiện này không chỉ khẳng định phương pháp đề xuất mà còn mở ra hướng nghiên cứu trong tương lai về bài toán dự báo xu hướng giá cổ phiếu.

CHƯƠNG 1: BÀI TOÁN PHÂN TÍCH GIÁ CỔ PHIẾU

1.1 Giới thiệu bài toán

Bài toán phân tích giá cổ phiếu nói chung có hai cách tiếp cận phổ biến đó là phân tích cơ bản và phân tích kỹ thuật. Phân tích cơ bản đòi hỏi rất nhiều hiểu biết về lĩnh vực tài chính và xã hội và đặc thù của từng doanh nghiệp. Trong khi đó, phân tích kỹ thuật thì rất trực quan và không phụ thuộc vào đặc thù của từng doanh nghiệp. Những nhà đầu tư dài hạn sẽ tập trung nhiều vào phân tích cơ bản để đưa ra quyết định còn phân tích kỹ thuật được những nhà đầu tư ngắn hạn ưu tiên sử dụng.

Phân tích kỹ thuật là việc sử dụng dữ liệu giao dịch trong quá khứ để dự đoán biến động giá trong tương lai. Dựa trên tâm lý thị trường, kinh tế học hành vi và phân tích định lượng, phân tích kỹ thuật hướng đến mục tiêu sử dụng hiệu suất trong quá khứ để dự đoán xu hướng của cổ phiếu trong tương lai. Hai hình thức phân tích kỹ thuật phổ biến nhất là các mẫu biểu đồ và các chỉ báo kỹ thuật mang tính chất thống kê.

Ở một khía cạnh khác, các kỹ thuật học sâu hoạt động theo nguyên tắc trích xuất các đặc trưng từ dữ liệu thô bằng cách sử dụng các nơron để xác định các khía cạnh khác nhau có liên quan đến dữ liệu đầu vào qua đó dự đoán đầu ra [18]. Qua đây, đặt ra bài toán về mối liên hệ giữa những chỉ báo kỹ thuật và khả năng ứng dụng vào việc hỗ trợ trích xuất đặc trưng cho các mô hình học sâu, qua đó nâng cao khả năng dự đoán giá cổ phiếu của mô hình. Do đó, đề án sẽ tập trung vào bài toán phân tích giá cổ phiếu dựa vào việc sử dụng các chỉ báo kỹ thuật và các mô hình học sâu để xác định giá đóng cửa của cổ phiếu.

Phát biểu bài toán:

Gọi:

- Giá cổ phiếu tại thời điểm t là p_t
- Giá cổ phiếu được dự đoán tại thời điểm $t+1$ là $\hat{p}_{t+1}$
- Ma trận của các chỉ báo kỹ thuật:

$$\mathbf{X}_t = \begin{bmatrix} TI_1(t) & TI_2(t) & \dots & TI_N(t) \\ TI_1(t-1) & TI_2(t-1) & \dots & TI_N(t-1) \\ \dots & \dots & \dots & \dots \\ TI_1(t-M) & TI_2(t-M) & \dots & TI_N(t-M) \end{bmatrix}$$

$\mathbf{X}_t$ là ma trận kích thước $M \times N$ với N đại diện cho số lượng chỉ báo kỹ thuật và $M+1$ là số ngày quan sát trong quá khứ.

Mục tiêu là tìm một hàm dự đoán $f(\cdot)$ sao cho:

$$\hat{p}_{t+1} = f(\mathbf{X}_t)$$

Với hàm mục tiêu là:

$$\min_f (\mathbf{p}_{t+1} - \hat{\mathbf{p}}_{t+1})^2$$

Đã có nhiều nghiên cứu về bài toán phân tích giá cổ phiếu sử dụng các kỹ thuật học máy khác nhau. Một số kỹ thuật tiêu biểu được sử dụng như mô hình thống kê ARIMA, mô hình học máy như SVM, RF và trong đó có cả các mô hình học sâu hiện đại như RNN, LSTM.

1.1.1 Giới thiệu về các yếu tố kỹ thuật

Chỉ báo kỹ thuật là các phép tính toán học được sử dụng để phân tích dữ liệu về giá, khối lượng giao dịch. Mục đích chung của việc sử dụng các chỉ báo này là để dự đoán xu hướng giá, xác định các điểm mua bán tiềm năng và đánh giá rủi ro trong giao dịch.

1.1.1.1 Nhóm chỉ báo về xu hướng

1.1.1.2 Nhóm chỉ báo về biến động

1.1.1.3 Nhóm chỉ báo về dao động

1.1.1.4 Nhóm chỉ báo về khối lượng

1.2 Các nghiên cứu liên quan

Những năm gần đây, sự phát triển của các mô hình học sâu đã giúp việc giải quyết các bài toán về dữ liệu dạng chuỗi thời gian hiệu quả hơn rõ rệt, đặc biệt là việc ứng dụng của mạng nơ-ron hồi quy RNN và các biến thể.

1.2.1 Các nghiên cứu trong nước liên quan

Mô hình học sâu kết hợp sử dụng LSTM-GRU

Trong nghiên cứu được công bố vào năm 2024, Trần Đăng Tuyên đã thực hiện sử dụng các biến thể của mô hình mạng nơ-ron hồi quy tiêu biểu như LSTM và GRU và cả

dạng phức hợp của chúng. Nghiên cứu cho thấy việc sử dụng các biến thể mô hình mạng nơron hồi quy như LSTM, GRU và các kết hợp của chúng cho hiệu quả tốt, hứa hẹn những nghiên cứu về ứng dụng của các mô hình mạng nơron hồi quy với những kết quả tốt hơn nữa trong tương lai [5].

1.2.2 Các nghiên cứu Quốc tế liên quan

Bài toán phân tích giá cổ phiếu thu hút được rất nhiều sự quan tâm của giới nghiên cứu Quốc tế, do đó các kỹ thuật hiện đại nhất đều được các nhà nghiên cứu thử nghiệm và áp dụng.

Trong nghiên cứu công bố năm 2024, Kui Fu và Yanbin Zhang đề xuất một mô hình dự đoán giá cổ phiếu kết hợp giữa dữ liệu tâm lý thị trường và dữ liệu giá từ nhiều nguồn. Tâm lý thị trường được mô hình BERT phân tích từ dữ liệu tin tức và những bình luận trực tuyến của các nhà đầu tư. Đồng thời, kết hợp chỉ báo kỹ thuật để xây dựng mô hình dự đoán thông qua biến thể của LSTM với cơ chế chú ý. Kết quả thực nghiệm cho thấy việc kết hợp tâm lý thị trường cải thiện độ chính xác của dự đoán ở mức độ lớn. Mô hình chứng minh tính mạnh mẽ và khái quát trong các tập dữ liệu thực nghiệm [10].

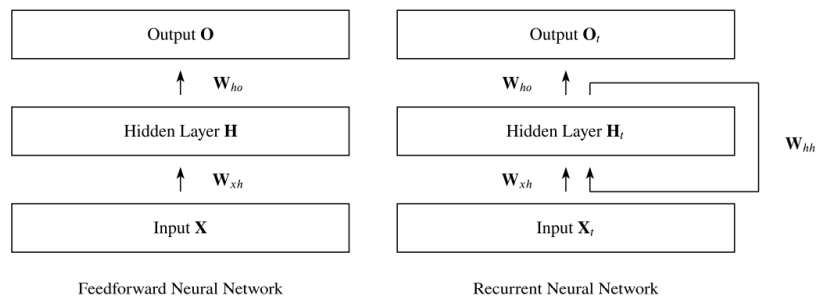
1.3 Kết luận chương

Chương này tạo tiền đề cho đề án bằng cách giới thiệu bài toán phân tích giá cổ phiếu, trong phần 1.2, đề án đi sâu vào bối cảnh nghiên cứu của bài toán ở cả trong và ngoài nước. Chương tiếp theo đề án sẽ trình bày khung cơ sở lý thuyết của các thành phần trong giải pháp được đề xuất.

CHƯƠNG 2: GIỚI THIỆU CÁC MÔ HÌNH HỌC SÂU CHO BÀI TOÁN

2.1 Giới thiệu mô hình RNN

2.1.1 Kiến trúc của mô hình RNN



Hình 2.1 Kiến trúc của mạng FFNN và RNN [20]

Trong đó công thức xác định giá trị của mô hình RNN:

$$H_t = \phi_h(X_t W_{xh} + H_{t-1} W_{hh} + b_h) \quad (1)$$

$$O_t = \phi_o(H_t W_{ho} + b_o) \quad (2) \quad [20]$$

Trong đó công thức xác định giá trị của mô hình FFNN:

$$H = \phi_h(XW_{xh} + b_h) \quad (3)$$

$$O = \phi_o(HW_{ho} + b_o) \quad (4) \quad [20]$$

Với:

X_t : là giá trị đầu vào tương ứng với thời gian t

H_t : là giá trị trạng thái ẩn tương ứng với thời gian t

O_t : là giá trị đầu ra ứng với thời gian t

W_{xh} : ma trận tham số của mô hình ứng với các giá trị đầu vào, dùng để tính toán giá trị cho trạng thái ẩn

W_{hh} : ma trận tham số của mô hình ứng với các giá trị trạng thái ẩn, dùng để tính toán giá trị cho trạng thái ẩn kế tiếp

W_{ho} : ma trận tham số của mô hình ứng với các giá trị trạng thái ẩn, dùng để tính toán giá trị đầu ra của mô hình

b_h : bias của trạng thái ẩn

b_o : bias của giá trị đầu ra

ϕ_h và ϕ_o : là các hàm kích hoạt, thường là sigmoid, tanh hay ReLU đối với mô

hình RNN

Vì H_t hồi quy bao gồm H_{t-1} và quá trình này xảy ra ở mọi bước thời gian nên RNN bao gồm dấu vết của tất cả các trạng thái ẩn trước H_{t-1} cũng như chính H_{t-1} . Nếu chúng ta so sánh ký hiệu đó cho RNN với ký hiệu tương tự cho FFNN thì chúng ta có thể thấy rõ sự khác biệt đã mô tả trước đó. Trong phương trình (3), chúng ta có thể thấy tính toán cho biến ẩn trong khi phương trình (4) hiển thị biến đầu ra [20].

Mô hình RNN có nhiều biến thể khác nhau, một số biến thể được biết đến rộng rãi như là mạng nơron hồi quy hai chiều (BRNN), mạng nơron hồi tiếp với nút có cổng (GRU) và mạng bộ nhớ dài ngắn hạn (LSTM).

2.1.2 Ứng dụng của mô hình RNN và các biến thể

Các biến thể của RNN được sử dụng trong việc dự đoán các dữ liệu dạng chuỗi thời gian. Một số bài toán xử lý văn bản như là nhận dạng giọng nói, sinh văn bản, dịch máy. Ngoài ra, các biến thể của RNN còn được ứng dụng trong lĩnh vực tài chính, cụ thể là dự đoán xu hướng giá của cổ phiếu như trong đề án nghiên cứu này.

2.1.3 Một số hạn chế của mô hình RNN

Hạn chế lớn nhất của mô hình RNN là vấn đề độ dốc bùng nổ hoặc biến mất.

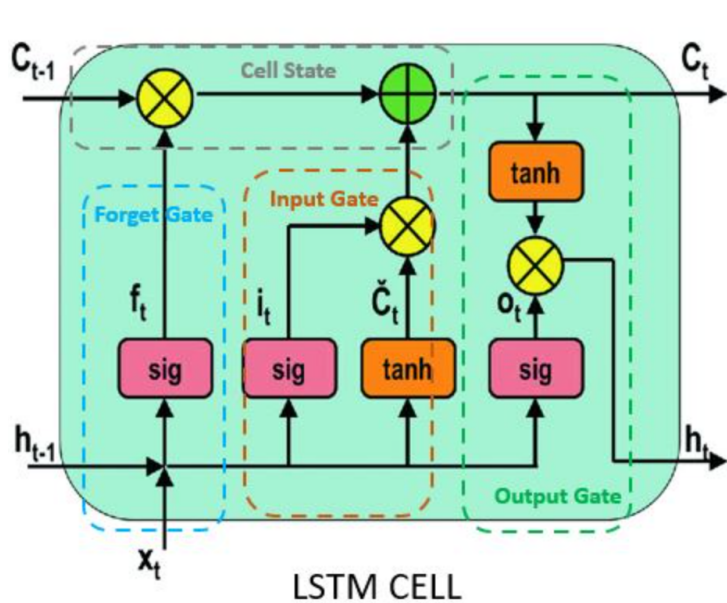
Do kiến trúc đặc thù, dữ liệu phải xử lý tuần tự theo từng bước thời gian, không thể xử lý song song đòi hỏi sức mạnh tính toán và thời gian huấn luyện mô hình RNN và các biến thể.

Ngoài ra, cũng giống như những mô hình học sâu khác, với nhiều cổng, ô trạng thái và trạng thái ẩn khiến việc nắm bắt trực quan cách thông tin được xử lý qua mạng là một vấn đề lớn, dẫn đến sự khó khăn trong việc lý giải mô hình.

2.2 Giới thiệu mô hình LSTM

2.2.1 Kiến trúc của LSTM

Mạng bộ nhớ dài ngắn hạn LSTM, một mạng nơ ron hồi quy đặc biệt, được thiết kế để khắc phục những hạn chế của RNN truyền thống trong việc nắm bắt các phụ thuộc dài hạn trong dữ liệu chuỗi. Kiến trúc của LSTM được thiết kế để điều chỉnh luồng thông tin, cho phép nó ghi nhớ và quên thông tin trong nhiều bước thời gian một cách hiệu quả.



Hình 2.2 Kiến trúc của LSTM [23]

Các thành phần của LSTM:

Cổng quên (Forget Gate): Cổng quên quyết định thông tin nào cần chú ý và thông

tin nào có thể bỏ qua. Thông tin từ đầu vào hiện tại $X(t)$ và trạng thái ẩn $h(t-1)$ được truyền qua hàm sigmoid. Sigmoid tạo ra các giá trị giữa 0 và 1. Nó kết luận xem phần đầu ra cũ có cần thiết hay không (bằng cách đưa đầu ra gần hơn với 1).

Cổng đầu vào (Input Gate): Xác định mức độ thông tin mới sẽ được cập nhật vào ô trạng thái. Thông tin từ đầu vào hiện tại $X(t)$ và trạng thái ẩn $h(t-1)$ được truyền qua hàm sigmoid và hàm tanh, sau đó chúng được nhân với nhau để tạo nên vector thông tin cho ô trạng thái.

Ô trạng thái (Cell State): đây là thành phần quan trọng nhất của LSTM. Thay vì chỉ lưu trữ thông tin đầu ra như RNN, LSTM lưu trữ thêm thông tin trạng thái, cho phép thông tin được truyền đi tương đối không thay đổi nếu cần thiết, đảm bảo rằng mạng có thể duy trì sự phụ thuộc lâu dài. Trạng thái ô được cập nhật bằng các cổng ở mỗi bước, thêm hoặc bớt thông tin khi cần.

Cổng đầu ra (Output Gate): Cổng đầu ra xác định giá trị của trạng thái ẩn tiếp theo. Trạng thái này chứa thông tin về các đầu vào trước đó. Sử dụng hàm sigmoid để lọc thông tin đầu vào hiện tại $X(t)$ và trạng thái ẩn $h(t-1)$, sử dụng hàm tanh để lọc thông tin của ô trạng thái. Cả hai đầu ra này đều được nhân với nhau để cho ra trạng thái ẩn.

2.2.2 Một số hạn chế của LSTM

Tuy giải quyết được vấn đề độ dốc biến mất của mô hình RNN, LSTM vẫn còn đó vấn đề về sức mạnh tính toán và thời gian huấn luyện mô hình và sự khó khăn trong việc lý giải mô hình.

Mô hình LSTM rất linh hoạt và có nhiều tham số, giúp chúng dễ dàng ghi nhớ dữ liệu huấn luyện. Điều này có thể dẫn đến việc quá khớp [16].

Sự ổn định của các mô hình LSTM cũng là một vấn đề, đặc biệt là trong các kiến trúc nhiều lớp [16].

Mô hình LSTM bị ảnh hưởng bởi nhiều siêu tham số, chỉ một thay đổi nhỏ ở siêu tham số này cũng có thể ảnh hưởng lớn đến độ tốt của mô hình [16].

2.3 Kết luận chương

Tóm lại, chương này đã cung cấp một cái nhìn tổng quan toàn diện về các mô hình làm nền tảng cho đề án này. Bắt đầu bằng việc khám phá kiến trúc mô hình RNN, kiến trúc này đóng vai trò nền tảng. Tiếp theo đề án tìm hiểu sâu về kiến trúc của mô hình LSTM, một trong những biến thể phổ biến của RNN. Chương tiếp theo sẽ đi sâu vào việc triển khai và đánh giá việc ứng dụng mô hình, bao gồm cả cách tinh chỉnh các siêu tham số để giải quyết những thách thức cụ thể trong bài toán phân tích giá cổ phiếu.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1 Thiết kế giải pháp

Để giải quyết bài toán phân tích giá cổ phiếu, giải pháp được đưa ra trong đề án sẽ chia thành 5 phần chính:

Thu thập dữ liệu: Dữ liệu được thu thập từ hai nguồn. Đầu tiên là từ API của <https://www.vndirect.com.vn/>, dữ liệu lịch sử giao dịch của từng mã cổ phiếu niêm yết trên thị trường chứng khoán Việt Nam. Nguồn thứ hai là từ trang <https://www.investing.com/>, dữ liệu về chỉ số VNIndex được tải trực tiếp và kết quả là một tệp dữ liệu dạng CSV duy nhất.

Trích xuất các chỉ báo kỹ thuật: Phần này đi vào việc lấy dữ liệu thô được lưu trữ trong cơ sở dữ liệu, trích xuất các chỉ báo kỹ thuật rồi sau đó xuất dữ liệu đó ra dạng tệp CSV, thuận lợi cho việc xử lý dữ liệu và xây dựng mô hình ở các bước sau.

Xác định độ quan trọng của các chỉ báo: Sau khi đã có tệp CSV chứa thông tin giao dịch cùng với các chỉ báo kỹ thuật, phần này đi vào việc xác định độ quan trọng của các chỉ báo với kích thước cửa sổ là 20. Bước này sử dụng hai phương pháp để xếp hạng đặc trưng là RF và MI.

Xây dựng mô hình dự báo giá cổ phiếu: Bước này thực hiện việc xây dựng mô hình học sâu sử dụng thư viện TensorFlow, scikit-learn, scikeras với 3 bộ dữ liệu đầu vào tương ứng với số lượng chỉ báo kỹ thuật sử dụng là một, bốn và mười chỉ báo. Bước thời gian được sử dụng trong đề án là 20, nói cách khác là sẽ dùng dữ liệu lịch sử của 20 ngày trước đó để dự đoán. Các mô hình được thử nghiệm là mô hình RNN có một lớp RNN, mô hình RNN có hai lớp RNN, mô hình LSTM có một lớp LSTM và mô hình LSTM có hai lớp LSTM. Các tham số được tối ưu bằng phương pháp tìm kiếm lưới nhằm đi tìm mô hình với bộ tham số tối ưu giúp việc đánh giá khách quan các mô hình.

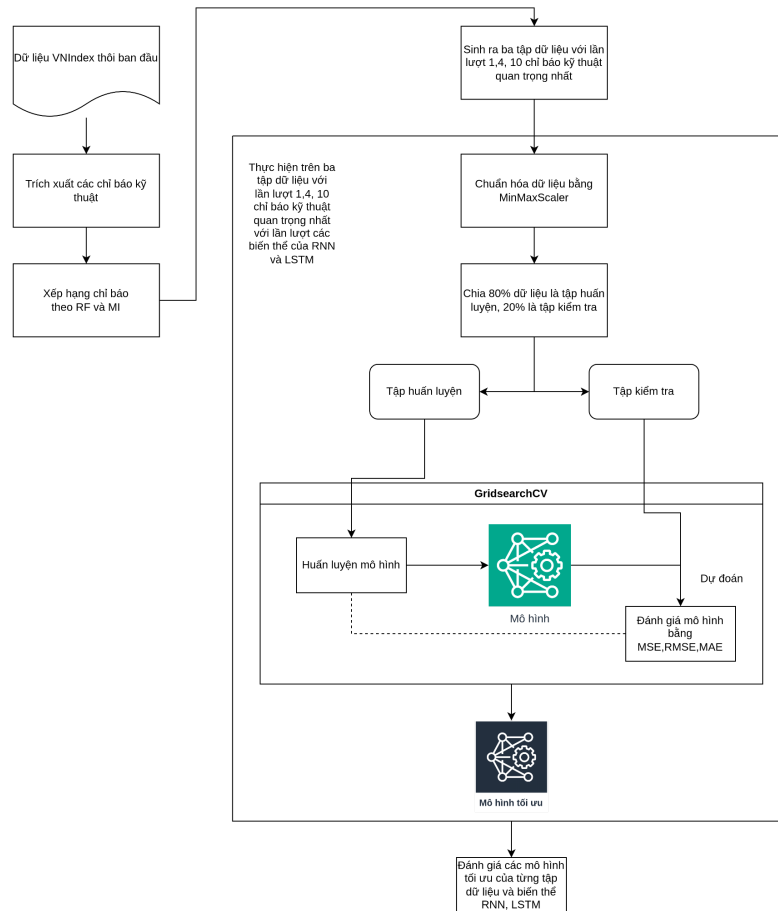
Đánh giá: Sau khi có được kết quả dự đoán giá cổ phiếu của các mô hình, việc đánh giá sẽ được dựa trên ba độ đo là MSE, MAE, RMSE, qua đó phân tích, so sánh về độ tốt của các mô hình, đánh giá hiệu quả và hạn chế của mỗi mô hình.



Hình 3.1 Chỉ số VN Index

Bảng 3.1 Thông tin chung của chỉ số VNIndex

coun t	mea n	std	min	25%	50%	75%	max	min date	max date
3493	781.27	307.28	336.73	506.53	684.2	1005.97	1528.57	2010-01-04	2023-12-29



Hình 3.2 Lưu đồ quy trình xử lý của phương pháp đề xuất

Hình 3.2 trên thể hiện một cách trực quan quy trình các bước thực hiện trong đề án, với việc tập trung đi biến thể mô hình của RNN và LSTM tối ưu với từng tập dữ liệu. Phần tiếp theo sẽ đi vào cụ thể các bước của quy trình này.

3.1.1 Tính toán các chỉ báo kỹ thuật

Như đã giới thiệu ở chương 1, các chỉ báo kỹ thuật được phân loại thành 4 nhóm: nhóm chỉ báo về xu hướng, nhóm chỉ báo về biến động, nhóm chỉ báo về động lượng và nhóm chỉ báo về khối lượng. Hiện nay đã có thư viện hỗ trợ tính toán các chỉ báo này trong ngôn ngữ lập trình python. Do đó, việc tính toán sẽ sử dụng thư viện hỗ trợ. Phần này sẽ giới thiệu một số chỉ báo kỹ thuật phổ biến và cách tính toán chúng.

3.1.2 Xây dựng mô hình đánh giá độ quan trọng của các chỉ báo

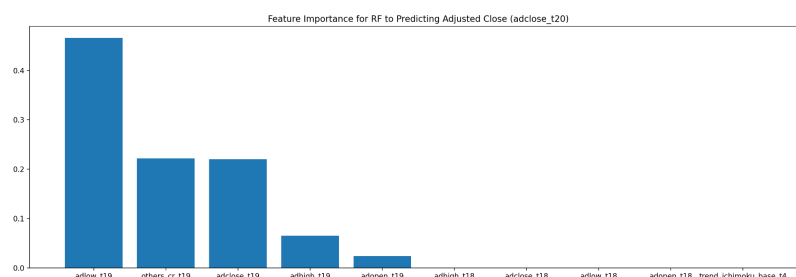
Sau khi đã trích xuất các chỉ báo kỹ thuật với số lượng tổng cộng là 90 chỉ báo. Phần này đi vào thực hiện việc xác định độ quan trọng của các chỉ báo qua đó giảm số chiều dữ liệu, tập trung vào những chiều dữ liệu quan trọng.

Bước này thực hiện việc sử dụng dữ liệu các chỉ báo kỹ thuật của hai mươi ngày liền trước (kí hiệu từ t , t_1 , ..., t_{19}) để xác định độ liên quan với giá đóng cửa của ngày t_{20} . Qua đó đánh giá độ quan trọng của các chỉ báo.

3.1.2.1 Sử dụng thuật toán RF

Trong nghiên cứu của mình năm 2023 về những kỹ thuật trích chọn và tổng hợp đặc trưng cho bài toán giá cổ phiếu của Htun và cộng sự, đã nghiên cứu tổng cộng 208 bài nghiên cứu trong vòng 12 năm (từ 2011-2022) đã kết luận rằng kỹ thuật học máy được sử dụng nhiều nhất là RF và SVM [11].

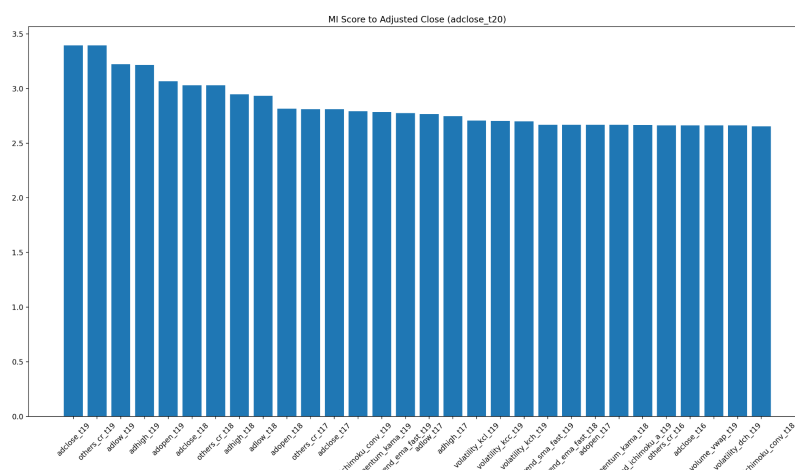
Do đó, đề án sử dụng kỹ thuật học máy RF để đánh giá độ liên quan của các chỉ báo kỹ thuật đối với giá trị đóng cửa của cổ phiếu nhằm lọc bỏ những chỉ báo kỹ thuật gây nhiễu.



Hình 3.7 Độ quan trọng của các chỉ báo theo RF

3.1.2.1 Sử dụng điểm MI

Không giống như thuật toán RF, MI xác định mối quan hệ đơn biến, đánh giá từng đặc trưng một cách độc lập, qua đó tránh việc “cho điểm” tập trung vào một vài chỉ báo quan trọng, khiến các chỉ báo khác được “cho điểm” thấp, khó đánh giá mức độ quan trọng.



Hình 3.8 Độ quan trọng của các chỉ báo theo MI

Qua quá trình sử dụng hai phương pháp, cả hai cùng có chung tập những chỉ báo quan trọng nhất, nhưng RF chỉ có 5 chỉ báo có điểm, các chỉ báo sau đó điểm gần như bằng không.

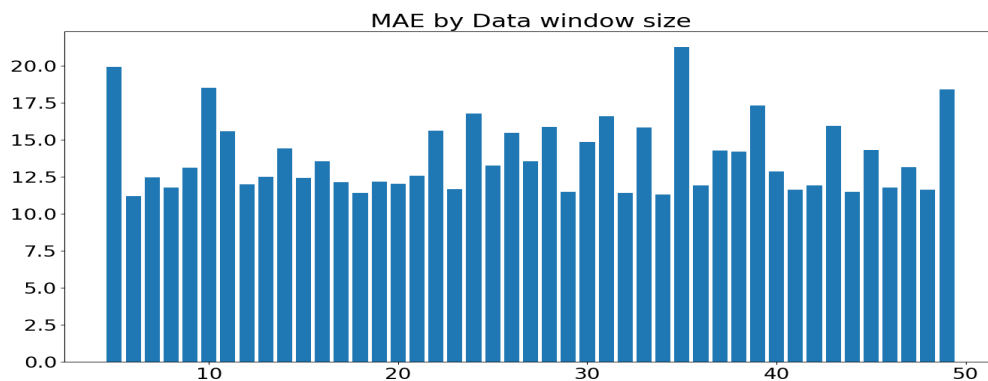
Dựa trên kết quả của việc đánh giá độ quan trọng của các chỉ báo, đề án sẽ chia dữ

liệu thành ba tập, tập có một đặc trưng, tập có bốn đặc trưng và tập sử dụng mười đặc trưng.

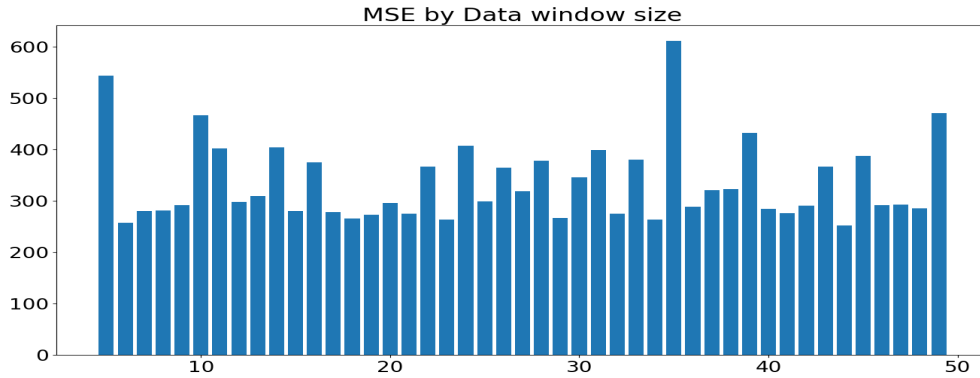
Vì tập có một chỉ báo đang có hai ứng cử nên đề án sẽ thực hiện loại trừ bằng việc chạy mô hình đơn giản trên cả hai tập, sau đó chọn ra chỉ báo cho kết quả tốt hơn. Hai tập đặc trưng quan trọng được xác định còn lại là bốn chỉ báo kỹ thuật cơ bản là giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa và tập thứ ba là mười chỉ báo quan trọng nhất theo MI sau khi loại bỏ yếu tố ngày.

3.2 Cài đặt thử nghiệm

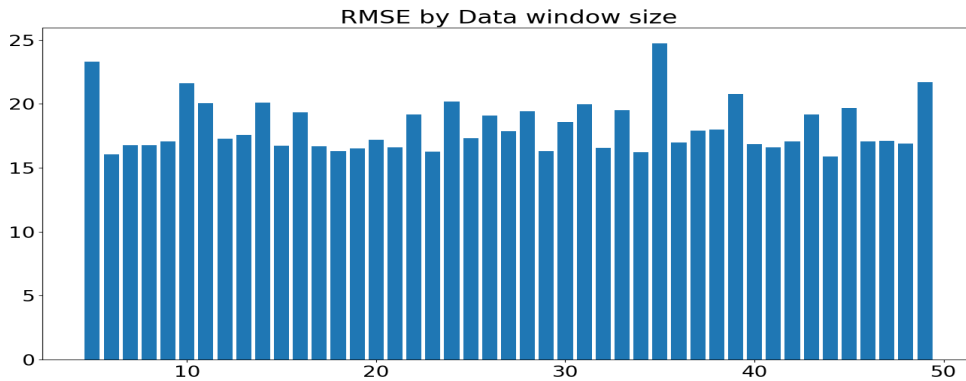
Đề án sử dụng phương pháp tìm kiếm lưới (GridsearchCV) để đi tìm mô hình tối ưu. Tuy nhiên, phương pháp này có đặc điểm là duyệt hết tất cả những tổ hợp của các siêu tham số dẫn đến số lượng lần chạy thử nghiệm sẽ tăng theo cấp số nhân tương ứng với số lượng giá trị thử nghiệm của siêu tham số. Do đó, để giảm chi phí cho quá trình huấn luyện, một số siêu tham số sẽ được xác định độc lập bằng một mô hình RNN đơn giản và tập dữ liệu một đặc trưng là giá đóng cửa, qua đó chọn giá trị tối ưu. Cụ thể dưới đây là khảo sát chọn giá trị tối ưu đối với siêu tham số là kích thước cửa sổ (window size) trên tập kiểm tra.



Hình 3.9 Điểm MAE theo kích thước cửa sổ



Hình 3.10 Điểm MSE theo kích thước cửa sổ



Hình 3.11 Điểm RMSE theo kích thước cửa sổ

Như khảo sát ở trên, kích thước cửa sổ có những ảnh hưởng đến độ tốt của mô hình, do đó đề án chọn giá trị kích thước cửa sổ là 20, bởi con số 20 cũng tương đương với số ngày giao dịch của bốn tuần và xấp xỉ số ngày giao dịch trong một tháng.

Ngoài ra, đề án sử dụng mô hình mạng nơ-ron với kích thước ô nhớ (batch size) là 32. Hàm relu được sử dụng làm hàm kích hoạt trong quá trình đào tạo mô hình. Sử dụng MSE cho hàm mất mát, và thuật toán Adam làm phương pháp tối ưu các tham số của mô hình.

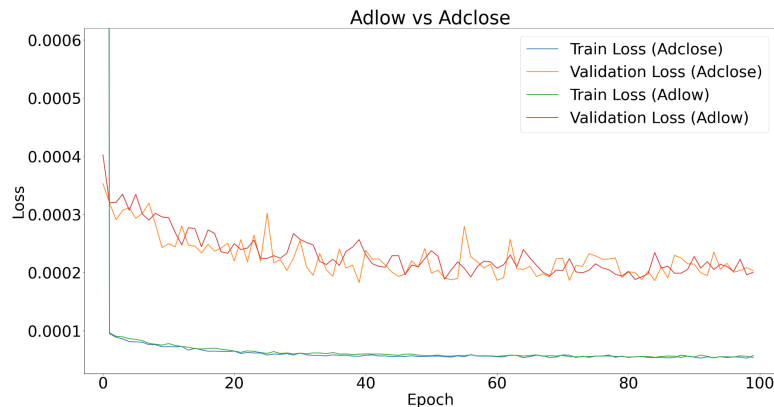
Tập có duy nhất một đặc trưng đang có hai đề cử là adlow (giá thấp nhất phiên giao dịch) dựa theo RF và adclose (giá đóng cửa phiên giao dịch) theo MI. Do đó, đề án thực hiện chạy tìm kiếm tối ưu bằng phương pháp tìm kiếm lưới (GridsearchCV) với mô hình RNN một lớp trên cả hai tập dữ liệu để tìm số nơ-ron tối ưu cho mỗi tập dữ liệu, sau đó đánh giá kết quả của hai mô hình.

Sau khi chạy tìm kiếm tối ưu, với tập dữ liệu sử dụng chỉ báo adclose, số lượng

neuron tối ưu là 128 và 64 đối với tập adlow. Dưới đây là bảng kết quả đánh giá hai mô hình với epoch là 100. Và để giảm tính ngẫu nhiên của mô hình, đề án sẽ đánh giá các mô hình bằng điểm trung bình của mười mô hình có cùng bộ siêu tham số.

Bảng 3.2 Bảng đánh giá độ tốt của mô hình trên hai tập dữ liệu

Chỉ báo kỹ thuật	Điểm đánh giá (Tập huấn luyện)			Điểm đánh giá (Tập kiểm tra)		
	MSE	RMSE	MAE	MSE	RMSE	MAE
adclose	66.17	8.13	5.42	249.20	15.79	11.15
adlow	67.74	8.23	5.49	250.14	15.82	11.44



Hình 3.12 Biểu đồ hàm mất mát của mô hình

Có thể thấy hai mô hình có điểm đánh giá khá tương đồng nhau, tuy nhiên, mô hình sử dụng tập dữ liệu với giá đóng cửa làm đặc trưng vẫn tốt hơn dù không nhiều. Do đó, tập sử dụng một đặc trưng sẽ là giá đóng cửa.

Đối với việc chọn tập có bốn đặc trưng tốt nhất và mười đặc trưng tốt nhất thì hai mô hình đánh giá có sự tương đồng.

Bảng 3.3 Danh sách các đặc trưng ứng với các tập dữ liệu

Số lượng đặc trưng	Danh sách đặc trưng
1	adclose
4	adclose,others_cr,adlow,adhigh
10	Adclose,others_cr,adlow,adhigh,adopen,trend_ichi

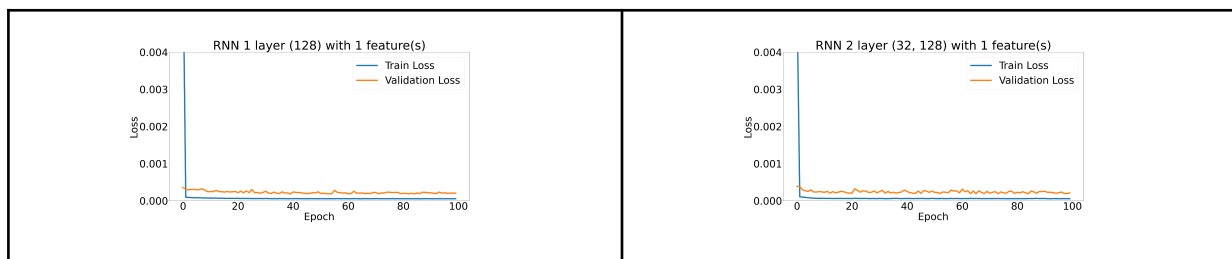
	moku_conv, momentum_kama,trend_ema_fast,volatility_kcl,volatility_kcc
--	--

3.2.1 Thử nghiệm với mô hình RNN

Thử nghiệm với mô hình RNN từ đơn giản có 1 lớp RNN sau đó đến những mô hình phức tạp hơn với 2 lớp RNN với số nút khác nhau. Việc xác định số nút cho mỗi mô hình sẽ được tối ưu bằng phương pháp tìm kiếm lưới (GridsearchCV).

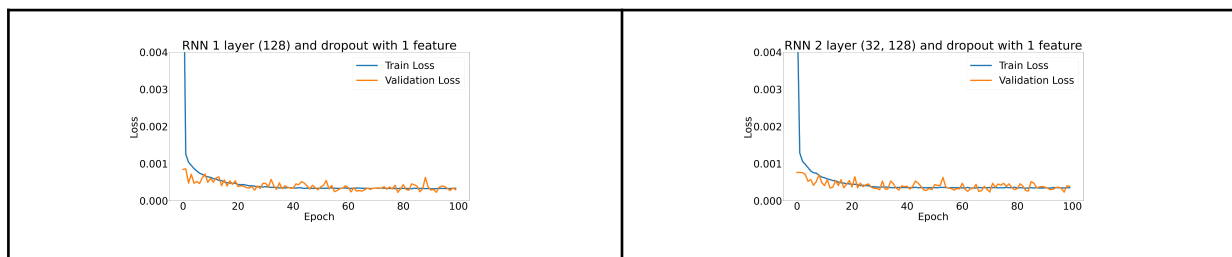
Đề án sử dụng phương pháp tìm kiếm lưới (GridsearchCV) sẽ sử dụng số lần lặp $n=5$. Với mô hình một lớp, số nơ ron trong lưới tìm kiếm là 32, 64, 128, 256 và 512 còn với mô hình hai lớp, số nơ ron trong lưới tìm kiếm là 32, 128 và 256. Trong khi đó, tham số epoch trong lưới tìm kiếm là 50, 80 và 100.

3.2.1.1 Tinh chỉnh các siêu tham số của mô hình



Hình 3.13 Biểu đồ hàm mất mát của mô hình RNN với 1 đặc trưng

Dựa vào hình 3.10, ta có thể thấy hai mô hình có giá trị hàm mất mát hội tụ khá nhanh cả trên tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Sau khoảng 35-40 epoch thì đi ngang với biên độ hẹp.



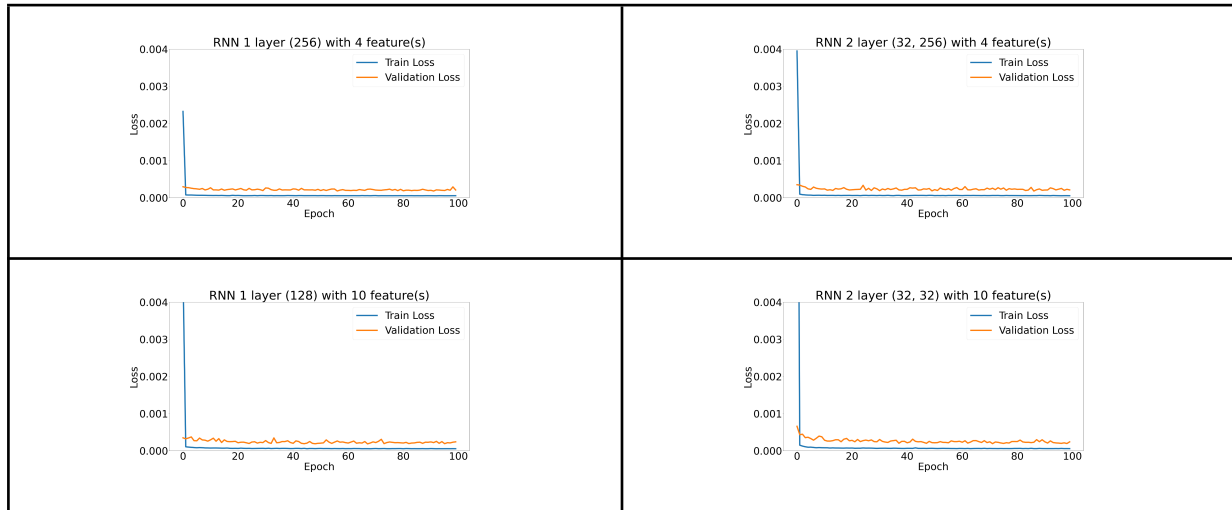
Hình 3.14 Biểu đồ hàm mất mát của mô hình RNN 1 đặc trưng với dropout=0.2

Dựa vào hình 3.11, việc thêm một lớp dropout vào sau lớp RNN với tỷ lệ là 20% đã khiến cho việc hội tụ của hàm mất mát trên tập dữ liệu huấn luyện bị chậm đi. Có thể thấy là việc sử dụng dropout ở đây tuy làm giảm sự quá khớp với dữ liệu huấn luyện nhưng cũng giảm cả sự ổn định của mô hình với tập dữ liệu kiểm tra.

3.2.1.2 Thay đổi tập dữ liệu đầu vào

Bước này ta sẽ thực hiện việc thay đổi tập dữ liệu đầu vào, cụ thể là tìm kiếm mô

hình tối ưu trên hai tập dữ liệu là tập với bốn đặc trưng và tập với mười đặc trưng đã được xây dựng. Dưới đây là kết quả các mô hình tối ưu thu được tương ứng với các tập dữ liệu.



Hình 3.15 Biểu đồ hàm mất mát của mô hình RNN tối ưu trên tập với 4 và 10 đặc trưng

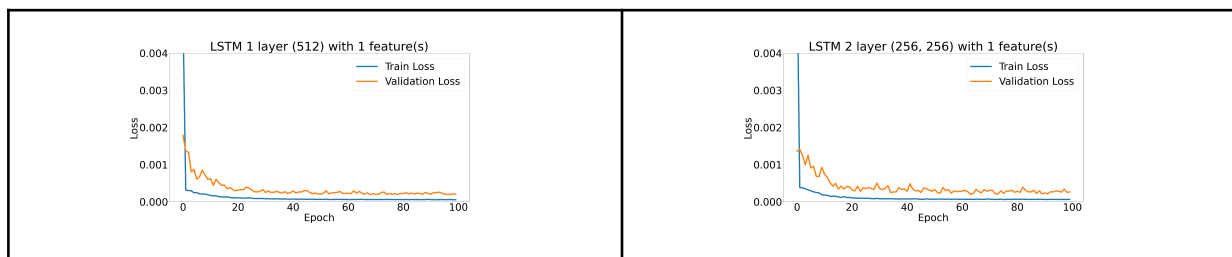
Dựa vào hình 3.12, ta có thể thấy cả bốn biểu đồ khá tương đồng. Cho thấy tuy là khác nhau về bộ dữ liệu hay là số lượng nơron, thì sự hội tụ trên tập dữ liệu huấn luyện và tập dữ liệu kiểm tra không bị ảnh hưởng nhiều.

3.2.2 Thử nghiệm với mô hình LSTM

Phần này sẽ đi vào việc thử nghiệm với mô hình LSTM từ đơn giản có 1 lớp LSTM sau đó đến những mô hình phức tạp hơn với 2 lớp LSTM với số nút khác nhau. Việc xác định số nút cho mỗi mô hình sẽ được tối ưu bằng phương pháp tìm kiếm lưới (GridsearchCV) sử dụng số lần lặp $n=5$. Với mô hình một lớp, số nơ ron trong lưới tìm kiếm là 32, 64, 128, 256 và 512 còn với mô hình hai lớp, số nơ ron trong lưới tìm kiếm là 32, 128 và 256. Trong khi đó, tham số epoch trong lưới tìm kiếm là 50, 80 và 100.

3.2.1.1 Tinh chỉnh các siêu tham số của mô hình

Sau quá trình tìm kiếm tham số tối ưu cho tập dữ liệu có một đặc trưng, đối với mô hình có số lớp LSTM là một, số nơron tối ưu là 512. Đối với mô hình có hai lớp LSTM, số nơron tối ưu là của mỗi lớp là 256.

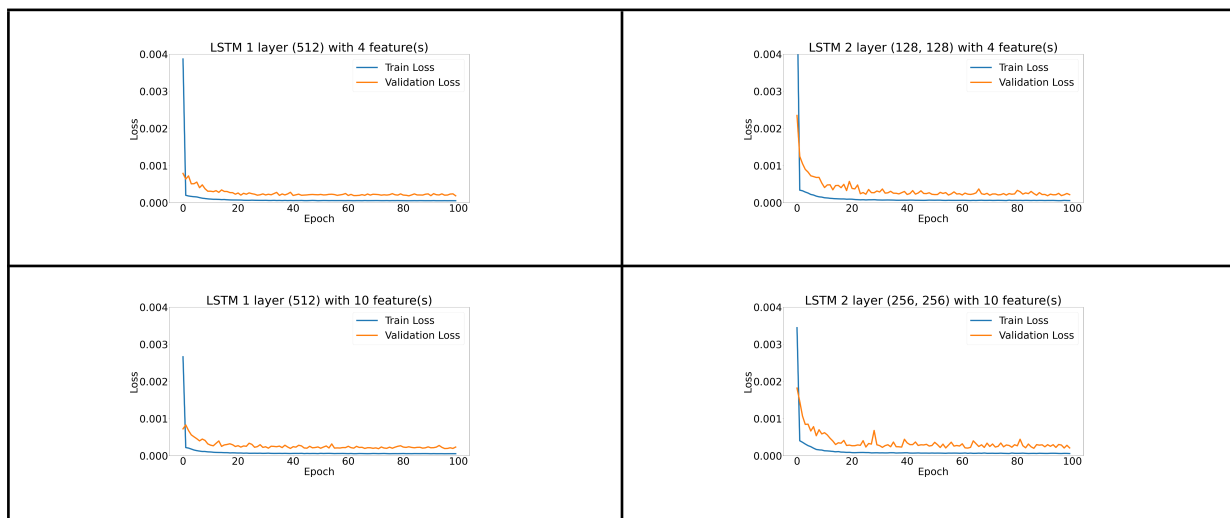


Hình 3.16 Biểu đồ hàm mất mát của mô hình LSTM với 1 đặc trưng

Dựa vào hình 3.13, ta có thể thấy hai mô hình có giá trị hàm mất mát hội tụ khá nhanh trên tập dữ liệu huấn luyện. Sau khoảng 35-40 epoch, đường màu xanh đi ngang với biên độ hẹp khá giống với mô hình RNN. Đường màu cam thể hiện giá trị hàm mất mát trên tập dữ liệu kiểm tra hội tụ với biên độ khá lớn, thể hiện sự kém ổn định so với mô hình RNN.

3.2.1.2 Thay đổi tập dữ liệu đầu vào

Giống như với RNN, bước này ta sẽ thực hiện việc thay đổi tập dữ liệu đầu vào, cụ thể là tìm kiếm mô hình tối ưu trên hai tập dữ liệu là tập với bốn đặc trưng và tập với mười đặc trưng đã được xây dựng. Dưới đây là kết quả các mô hình tối ưu thu được tương ứng với các tập dữ liệu.



Hình 3.17 Biểu đồ hàm mất mát của mô hình RNN với 4 và 10 đặc trưng

Dựa vào hình 3.14, ta có thể thấy hai mô hình có giá trị hàm mất mát hội tụ khá nhanh trên tập dữ liệu huấn luyện và kiểm tra. Đối với mô hình có hai lớp LSTM, biên độ dao động của giá trị hàm mất mát trên tập dữ liệu kiểm tra lớn hơn rõ rệt so với mô hình có một lớp LSTM, và sự hội tụ trên tập huấn luyện cũng chậm hơn.

3.3 Đánh giá

Phần này đi vào đánh giá độ tốt của từng loại mô hình trên hai tập dữ liệu là tập huấn luyện và tập kiểm tra. Kết quả thể hiện ở hình bên dưới là khi các mô hình được huấn luyện với siêu tham số epoch=100.

Bảng 3.4 Bảng đo độ chính xác trên tập huấn luyện

Mô hình	Số lớp ẩn	Số lượng nơ ron	Số lượng tham số	Số lượng đặc trưng	Điểm đánh giá (Tập huấn luyện)		
					MSE	RMS E	MAE
RNN	1	128	16,76	1	66.17	8.13	5.42

			9				
	1	256	67,07 3	4	65.81	8.11	5.41
	1	128	17,92 1	10	66.22	8.14	5.47
	2	(32, 128)	21,82 5	1	69.69	8.35	5.60
	2	(32, 256)	75,42 5	4	65.77	8.11	5.47
	2	(32, 32)	3,489	10	77.01	8.78	5.93
LST M	1	512	1,053, 185	1	66.89	8.18	5.42
	1	512	1,059, 329	4	68.76	8.29	5.51
	1	512	1,071, 617	10	73.05	8.55	5.88
	2	(256, 256)	789,7 61	1	66.87	8.18	5.46
	2	(128, 128)	199,8 09	4	67.58	8.22	5.50
	2	(256, 256)	798,9 77	10	70.94	8.42	5.63

Bảng 3.5 Bảng đo độ chính xác trên tập kiểm tra

Model	Số lớp ẩn	Số lượng nơ ron	Số lượng tham số	Số lượng đặc trung	Điểm đánh giá (Tập kiểm tra)		
					MSE	RMS E	MAE
RNN	1	128	16,76 9	1	249.2 0	15.79	11.15
	1	256	67,07 3	4	241.6 5	15.55	11.15
	1	128	17,92 1	10	243.0 7	15.59	11.23

	2	(32, 128)	21,825	1	256.65	16.02	11.26
	2	(32, 256)	75,425	4	243.33	15.60	11.39
	2	(32, 32)	3,489	10	294.01	17.15	12.13
LSTM	1	512	1,053,185	1	248.61	15.77	11.31
	1	512	1,059,329	4	247.61	15.74	11.43
	1	512	1,071,617	10	292.90	17.11	13.36
	2	(256, 256)	789,761	1	288.47	16.98	12.78
	2	(128, 128)	199,809	4	259.76	16.12	11.91
	2	(256, 256)	798,977	10	254.25	15.95	11.57

Dựa vào bảng đánh giá trên, có thể thấy các mô hình có độ tốt khá tương đồng, dù mô hình đó có số lượng tham số từ vài nghìn đến hơn một triệu tham số hay tập dữ liệu có một hay mười đặc trưng. Đặc biệt là trên tập huấn luyện, điểm MSE, RMSE và MAE của các mô hình khá sát nhau. Điểm đánh giá trên tập dữ liệu huấn luyện và tập dữ liệu kiểm tra có khoảng cách nhất định.

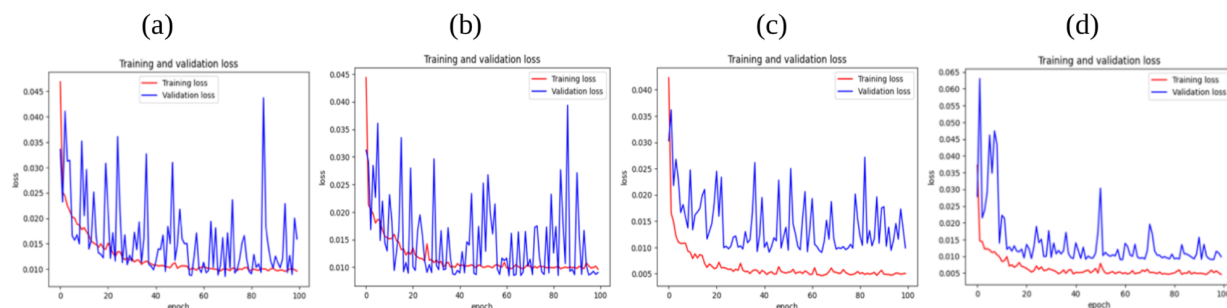
Năm 2024, Trần Đăng Tuyên công bố nghiên cứu của mình về việc sử dụng mô hình phức hợp LSTM-GRU để dự báo chỉ số VNIndex, tuy nhiên có một số điểm khác như tập dữ liệu là từ ngày 14/7/2008 đến 14/7/2023 với số quan sát là 3747 và chọn kích thước bước thời gian là 50 [4]. Dưới đây là kết quả được Trần Đăng Tuyên công bố:

Bảng 3.6 Bảng đo độ chính xác trên tập huấn luyện [4]

Thông số	LSTM(0:18:50)	GRU(0:13:06)	GRU-LSTM(0:14:21)	LSTM-GRU(0:12:28)
r2_score	0,99794	0,99866	0,99833	0,99868
ev_score	0,99801	0,99867	0,99868	0,99869
Gamma Deviance	0,00024	0,00017	0,00021	0,00017
Poisson Deviance	0,13842	0,09588	0,11958	0,09361
MSE	93,70231	60,83519	76,14630	59,88852
RMSE	9,68000	7,79969	8,72618	7,73877
MAE	7,16149	5,47041	6,34235	5,37542
MAPE	0,01681	0,00946	0,01093	0,00926

Bảng 3.7 Bảng đo độ chính xác trên tập kiểm tra [4]

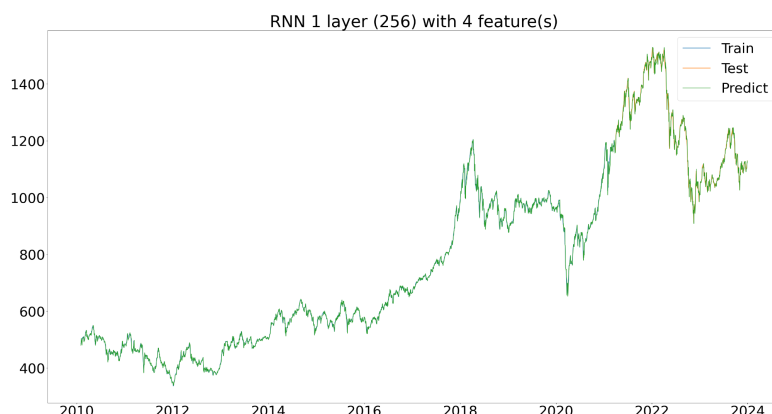
Thông số	LSTM	GRU	GRU-LSTM	LSTM-GRU
r2_score	0,98028	0,99052	0,98949	0,99108
ev_score	0,99064	0,99061	0,98960	0,99136
Gamma Deviance	0,00039	0,00019	0,00021	0,00018
Poisson Deviance	0,47032	0,23077	0,25072	0,21672
MSE	578,51989	278,18633	308,34858	261,68583
RMSE	24,05244	17,37117	17,55986	16,17671
MAE	20,57606	11,78603	12,81945	11,69156
MAPE	0,01681	0,00979	0,01050	0,00970

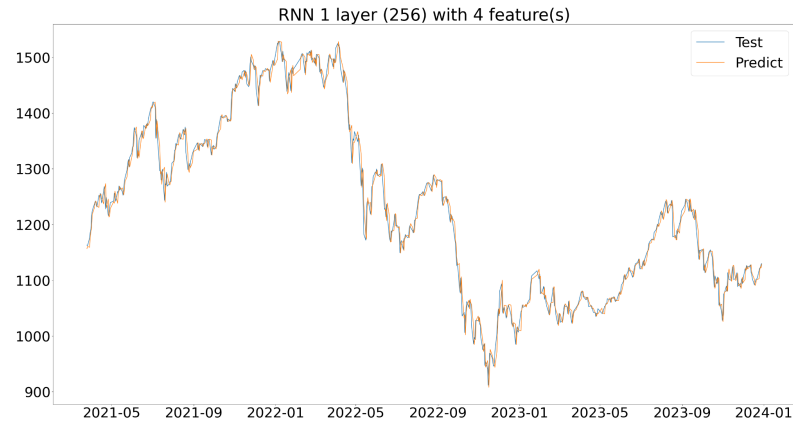
**Hình 3.18 Biểu đồ so sánh của hàm loss [4]**

Ghi chú: (a) LSTM, (b) GRU, (c) GRU-LSTM, (d) LSTM-GRU

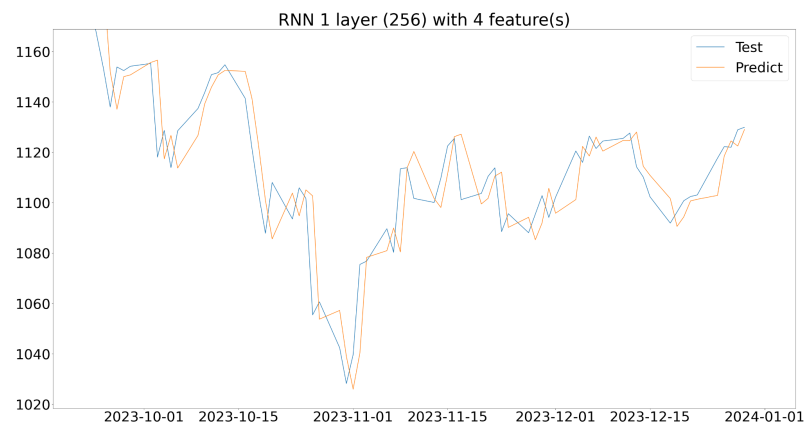
Khi xét trên thang điểm là MSE, RMSE và MAE, có thể thấy mô hình phức hợp LSTM-GRU của Trần Đăng Tuyên cho điểm số trên tập dữ liệu kiểm tra khá tương đồng với các mô hình được thực nghiệm trong đề án, trong khi điểm số trên tập dữ liệu huấn luyện lại tốt hơn đôi chút, điều này thể hiện sự quá khớp của mô hình.

Dưới đây là một số hình ảnh ứng dụng mô hình đề án nghiên cứu vào dự báo chỉ số VNIndex.

**Hình 3.19 Xu hướng giá của mô hình RNN 1 lớp trên toàn bộ tập dữ liệu**



Hình 3.20 Xu hướng giá của mô hình RNN 1 lớp trên tập dữ liệu kiểm tra



Hình 3.21 Xu hướng giá của mô hình RNN 1 lớp trên một phần của tập dữ liệu

3.4 Kết luận chương

Chương này đề án trình bày giải pháp và các bước thực nghiệm, từ thu thập dữ liệu đến đo lường đánh giá các phiên bản khác nhau của mô hình. Kết quả cho thấy tính hiệu quả của các mô hình học sâu RNN và LSTM, trong việc dự báo giá cổ phiếu. Những thử nghiệm này không chỉ khẳng định phương pháp đề xuất là đúng hướng, mà còn mở ra nhiều hướng nghiên cứu trong tương lai về dự báo giá các cổ phiếu trên sàn chứng khoán Việt Nam.

KẾT LUẬN

Bài toán dự báo giá cổ phiếu nói chung và dự báo giá cổ phiếu trên thị trường chứng khoán Việt Nam nói riêng, là một bài toán được nghiên cứu từ lâu bằng nhiều kỹ thuật học máy khác nhau, và có tính thực tiễn cao. Đề án này nhằm nghiên cứu sử dụng những khả năng của các kỹ thuật học sâu tiên tiến trong bài toán dự báo giá cổ phiếu. Với việc ứng dụng các chỉ báo kỹ thuật và tối ưu mô hình bằng tìm kiếm lưới, đề án đã đạt được một số kết quả trong bài toán dự báo giá cổ phiếu, ngoài ra cũng đối mặt với những thách thức vốn có và mở ra con đường nghiên cứu trong tương lai.

Một trong những đạt được của đề án là triển khai thành công các kỹ thuật học sâu RNN và LSTM với tập dữ liệu sử dụng các chỉ báo kỹ thuật của cổ phiếu. Mô hình đã được thiết kế và tinh chỉnh để có được những dự báo giá cổ phiếu cho kết quả tốt, gần với thực tế.

Một khía cạnh quan trọng của đề án này là việc tạo ra các mô hình học sâu đơn giản chỉ sử dụng từ một đến hai lớp ẩn, với số lượng tham số không lớn, nhưng đã cho kết quả ấn tượng. Đây là đóng góp cho cộng đồng nghiên cứu, cung cấp nền tảng cho các nghiên cứu và phát triển mô hình trong tương lai.

Đề án cũng gặp phải những thách thức, đặc biệt là điểm yếu của các mô hình học sâu nói chung, đó là vấn đề về dữ liệu. Việc số lượng dữ liệu về giá cổ phiếu tại thị trường Việt Nam, chỉ số VNIndex trong mười ba năm, giai đoạn từ 01/01/2010 đến 31/12/2023, là hơn 3400 dữ liệu, tương ứng với hơn 3400 ngày giao dịch, lượng dữ liệu khá hạn chế khi so sánh với số lượng các tham số của các mô hình học sâu. Việc các mô hình thử nghiệm cho kết quả khá tương đồng dù với những tập đặc trưng khác nhau cũng đặt ra bài toán về đánh giá mức độ quan trọng của các đặc trưng trong việc xây dựng mô hình hiệu quả.

Chi phí tính toán liên quan đến các mô hình học sâu như RNN và LSTM cũng là một thách thức lớn khác, đặt ra những giới hạn đối với số lượng các thử nghiệm của đề án, trong khi các mô hình học sâu này lại có thể tinh chỉnh, kết hợp rất nhiều biến thể.

Trong tương lai đề án có thể tập trung vào việc phát triển thêm các biến thể của mô hình RNN và LSTM thông qua việc thêm số lớp ẩn, và sử dụng các mô hình phức hợp khác nhau và tìm kiếm những phương pháp đánh giá đặc trưng mới để tìm ra tập đặc trưng tốt hơn. Những điều này có thể nâng cao hơn nữa chất lượng dự báo của mô hình.

Ngoài việc tập trung vào một chỉ số là VNIndex, mô hình có thể thử nghiệm trên các mã cổ phiếu khác trên thị trường.

Tóm lại, đề án này đã đóng góp một phần nhỏ trong bài toán dự báo giá cổ phiếu trên thị trường chứng khoán Việt Nam, kiểm chứng được khả năng ứng dụng các mô hình học sâu cho bài toán dự báo giá cổ phiếu. Những khó khăn và vấn đề gặp phải

trong quá trình nghiên cứu đã giúp mang lại những hiểu biết quý giá trong bài toán dự báo giá cổ phiếu.