

**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**

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT**  
*(Theo định hướng ứng dụng)*

**HÀ NỘI-2025**

**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**

**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**

## **LỜI CAM ĐOAN**

Tôi xin cam đoan mọi nội dung nghiên cứu, số liệu, kết quả trong đề án “Phân tích xu hướng giá cổ phiếu dựa trên các kỹ thuật học sâu” của tôi là công trình nghiên cứu của cá nhân tôi, mọi nội dung đều là trung thực và không sao chép từ bất kỳ báo cáo, công trình nào có trước.

Ký và ghi rõ họ tên

**NGUYỄN THẾ ANH**

## **LỜI CẢM ƠN**

Tôi xin gửi lời cảm ơn sâu sắc đến PGS.TS. Trần Đình Quế, Khoa Công Nghệ Thông Tin, vì sự hỗ trợ và hướng dẫn tâm huyết trong suốt quá trình thực hiện đề án của tôi. Sự đóng góp của thầy trong việc giảng dạy và hướng dẫn, từ việc góp ý tên đề tài đề án cho đến việc thực hiện và kiểm tra quá trình soạn thảo báo cáo này một cách tỉ mỉ, đều rất quý giá. Kiến thức chuyên môn và lời khuyên sâu sắc của thầy đã góp phần quan trọng giúp tôi hoàn thành thành công đề án của mình.

Ngoài ra, tôi cũng xin gửi lời cảm ơn chân thành tới toàn thể các thầy cô giáo trong Khoa Công Nghệ Thông Tin, và các thầy cô giáo khoa Sau Đại Học, những người đã tận tâm giảng dạy và hướng dẫn tôi trong suốt hai năm học tập tại trường. Trí tuệ và sự tâm huyết của mọi người là nền tảng cho sự phát triển của tôi.

**NGUYỄN THẾ ANH**

## MỤC LỤC

|                                                                    |           |
|--------------------------------------------------------------------|-----------|
| <b>LỜI CAM ĐOAN.....</b>                                           | <b>3</b>  |
| <b>LỜI CẢM ƠN.....</b>                                             | <b>4</b>  |
| <b>MỤC LỤC.....</b>                                                | <b>5</b>  |
| <b>DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT.....</b>                 | <b>7</b>  |
| <b>DANH MỤC CÁC BẢNG.....</b>                                      | <b>8</b>  |
| <b>DANH MỤC CÁC HÌNH.....</b>                                      | <b>10</b> |
| <b>MỞ ĐẦU.....</b>                                                 | <b>1</b>  |
| <b>CHƯƠNG 1: BÀI TOÁN PHÂN TÍCH GIÁ CỔ PHIẾU.....</b>              | <b>3</b>  |
| 1.1 Giới thiệu bài toán.....                                       | 3         |
| 1.1.1 Giới thiệu về các yếu tố kỹ thuật.....                       | 4         |
| 1.1.1.1 Nhóm chỉ báo về xu hướng.....                              | 5         |
| 1.1.1.2 Nhóm chỉ báo về biến động.....                             | 5         |
| 1.1.1.3 Nhóm chỉ báo về dao động.....                              | 6         |
| 1.1.1.4 Nhóm chỉ báo khối lượng.....                               | 6         |
| 1.2 Các nghiên cứu liên quan.....                                  | 7         |
| 1.2.1 Các nghiên cứu trong nước liên quan.....                     | 7         |
| 1.2.2 Các nghiên cứu Quốc tế liên quan.....                        | 8         |
| 1.3 Kết luận chương.....                                           | 9         |
| <b>CHƯƠNG 2: GIỚI THIỆU CÁC MÔ HÌNH HỌC SÂU CHO BÀI TOÁN.....</b>  | <b>10</b> |
| 2.1 Giới thiệu mô hình RNN.....                                    | 10        |
| 2.1.1 Kiến trúc của mô hình RNN.....                               | 10        |
| 2.1.2 Ứng dụng của mô hình RNN và các biến thể.....                | 11        |
| 2.1.3 Một số hạn chế của mô hình RNN.....                          | 13        |
| 2.2 Giới thiệu mô hình LSTM.....                                   | 13        |
| 2.2.1 Kiến trúc của LSTM.....                                      | 13        |
| 2.2.2 Một số hạn chế của LSTM.....                                 | 15        |
| 2.3 Kết luận chương.....                                           | 15        |
| <b>CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ.....</b>                      | <b>16</b> |
| 3.1 Thiết kế giải pháp.....                                        | 16        |
| 3.1.1 Tính toán các chỉ báo kỹ thuật.....                          | 18        |
| 3.1.2 Xây dựng mô hình đánh giá độ quan trọng của các chỉ báo..... | 22        |
| 3.1.2.1 Sử dụng thuật toán RF.....                                 | 22        |
| 3.1.2.2 Sử dụng điểm MI.....                                       | 23        |
| 3.2 Cài đặt thử nghiệm.....                                        | 24        |
| 3.2.1 Thử nghiệm với mô hình RNN.....                              | 27        |
| 3.2.1.1 Tinh chỉnh các siêu tham số của mô hình.....               | 28        |
| 3.2.1.2 Thay đổi tập dữ liệu đầu vào.....                          | 29        |

|                                                      |           |
|------------------------------------------------------|-----------|
| 3.2.2 Thử nghiệm với mô hình LSTM.....               | 30        |
| 3.2.2.1 Tinh chỉnh các siêu tham số của mô hình..... | 30        |
| 3.2.2.2 Thay đổi tập dữ liệu đầu vào.....            | 31        |
| 3.3 Đánh giá.....                                    | 32        |
| 3.4 Kết luận chương.....                             | 36        |
| <b>KẾT LUẬN.....</b>                                 | <b>37</b> |
| <b>TÀI LIỆU THAM KHẢO.....</b>                       | <b>39</b> |

## DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

| Viết tắt | Tiếng Anh                                | Tiếng Việt                                   |
|----------|------------------------------------------|----------------------------------------------|
| AI       | Artificial Intelligence                  | Trí tuệ nhân tạo                             |
| API      | Application Programming Interface        | Giao diện lập trình ứng dụng                 |
| ARIMA    | AutoRegressive Integrated Moving Average | Mô hình tự hồi quy tích hợp trung bình trượt |
| BRNN     | Bidirectional Recurrent Neural Network   | Mạng nơ-ron hồi quy hai chiều                |
| CNN      | Convolutional Neural Network             | Mạng nơ-ron tích chập                        |
| CSV      | Comma-Separated Values                   | Định dạng tệp chứa dữ liệu                   |
| EMA      | Exponential Moving Average               | Đường trung bình động hàm mũ                 |
| FFNN     | Feedforward Neural Network               | Mạng nơ-ron truyền thẳng                     |
| GRU      | Gated Recurrent Unit                     | Mạng nơ-ron hồi quy với nút có cổng          |
| JSON     | JavaScript Object Notation               | Kiểu dữ liệu mở trong JavaScript             |
| LSTM     | Long Short-Term Memory                   | Mạng nơ-ron bộ nhớ dài ngắn hạn              |
| MA       | Moving Average                           | Đường trung bình động                        |
| MACD     | Moving Average Convergence Divergence    | Đường trung bình hội tụ/phân kỳ              |
| MAE      | Mean Absolute Error                      | Sai số tuyệt đối trung bình.                 |
| MI       | Mutual Information                       | Thông tin tương hỗ                           |

|      |                           |                                               |
|------|---------------------------|-----------------------------------------------|
| MSE  | Mean Squared Error        | Sai số bình phương trung bình                 |
| OBV  | On-Balance Volume         | Chỉ báo khối lượng cân bằng                   |
| RF   | Random Forest             | Thuật toán gồm nhiều cây quyết định           |
| RMSE | Root Mean Squared Error   | Căn bậc hai của sai số bình phương trung bình |
| RNN  | Recurrent Neural Network  | Mạng nơ-ron hồi quy                           |
| RSI  | Relative Strength Index   | Chỉ báo sức mạnh tương đối                    |
| SMA  | Simple Moving Average     | Trung bình động đơn giản                      |
| SVM  | Support Vector Machine    | Thuật toán máy vector hỗ trợ                  |
| SVR  | Support Vector Regression | Thuật toán hồi quy vector hỗ trợ              |



## DANH MỤC CÁC BẢNG

|                                                                     |    |
|---------------------------------------------------------------------|----|
| Bảng 3.1 Thông tin chung của chỉ số VNIndex.....                    | 17 |
| Bảng 3.2 Bảng đánh giá độ tốt của mô hình trên hai tập dữ liệu..... | 26 |
| Bảng 3.3 Danh sách các đặc trưng tương ứng với tập dữ liệu.....     | 27 |
| Bảng 3.4 Bảng đo độ chính xác trên tập huấn luyện.....              | 32 |
| Bảng 3.5 Bảng đo độ chính xác trên tập kiểm tra.....                | 33 |
| Bảng 3.6 Bảng đo độ chính xác trên tập huấn luyện [5].....          | 34 |
| Bảng 3.7 Bảng đo độ chính xác trên tập kiểm tra [5].....            | 34 |

## DANH MỤC CÁC HÌNH

|                                                                                 |    |
|---------------------------------------------------------------------------------|----|
| Hình 2.1 Kiến trúc của mạng FFNN và RNN [20].....                               | 10 |
| Hình 2.2 Kiến trúc của LSTM [23].....                                           | 14 |
| Hình 3.1 Chỉ số VN Index.....                                                   | 17 |
| Hình 3.2 Lưu đồ quy trình xử lý của phương pháp đề xuất.....                    | 18 |
| Hình 3.3 Chỉ báo đường trung bình động.....                                     | 19 |
| Hình 3.4 Chỉ báo dải Bollinger.....                                             | 20 |
| Hình 3.5 Chỉ báo sức mạnh tương đối RSI.....                                    | 21 |
| Hình 3.6 Chỉ báo khối lượng cân bằng OBV.....                                   | 22 |
| Hình 3.7 Độ quan trọng của các chỉ báo theo RF.....                             | 23 |
| Hình 3.8 Độ quan trọng của các chỉ báo theo MI.....                             | 24 |
| Hình 3.9 Điểm MAE theo kích thước cửa sổ.....                                   | 25 |
| Hình 3.10 Điểm MSE theo kích thước cửa sổ.....                                  | 25 |
| Hình 3.11 Điểm RMSE theo kích thước cửa sổ.....                                 | 25 |
| Hình 3.12 Biểu đồ hàm mất mát của mô hình trên hai tập dữ liệu.....             | 27 |
| Hình 3.13 Biểu đồ hàm mất mát của mô hình RNN với 1 đặc trưng.....              | 28 |
| Hình 3.14 Biểu đồ hàm mất mát của mô hình RNN 1 đặc trưng với dropout=0.2.....  | 29 |
| Hình 3.15 Biểu đồ hàm mất mát của mô hình RNN với 4 và 10 đặc trưng.....        | 30 |
| Hình 3.16 Biểu đồ hàm mất mát của mô hình LSTM với 1 đặc trưng.....             | 31 |
| Hình 3.17 Biểu đồ hàm mất mát của mô hình RNN với 4 và 10 đặc trưng.....        | 32 |
| Hình 3.18 Biểu đồ so sánh của hàm loss [5].....                                 | 34 |
| Hình 3.19 Xu hướng giá của mô hình RNN 1 lớp trên toàn bộ tập dữ liệu.....      | 35 |
| Hình 3.20 Xu hướng giá của mô hình RNN 1 lớp trên tập dữ liệu kiểm tra.....     | 35 |
| Hình 3.21 Xu hướng giá của mô hình RNN 1 lớp trên một phần của tập dữ liệu..... | 35 |

## 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 nơ-ron nhân tạo chuyên xử lý cho dữ liệu dạng chuỗi thời gian. Sau đó, đề án sẽ tìm hiểu về kiến trúc của mạng nơ-ron LSTM, là sự cải tiến từ mạng nơ-ron 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 dự đoán giá cổ phiếu với dữ liệu dạng chuỗi thời gian được biết đến là một vấn đề khó khăn do thị trường chứng khoán là dạng thị trường hiệu quả trung bình với mức độ nhiễu cao [9]. Mặt khác, thị trường chứng khoán Việt Nam được coi là "phong vũ biểu" của nền kinh tế, những diễn biến trên thị trường chứng khoán có thể phản ánh những thay đổi và triển vọng của nền kinh tế trong tương lai. Theo giới phân tích thì thị trường chứng khoán sẽ đi trước sự thay đổi của nền kinh tế sáu tháng [30]. Thị trường chứng khoán là thị trường thể hiện niềm tin và kỳ vọng của các nhà đầu tư, là tổng hòa của rất nhiều yếu tố như nội tại của doanh nghiệp, những thông tin về kinh tế vĩ mô, tình hình chính trị - xã hội, tâm lý đám đông và thậm chí là những thông tin liên quan đến các cá nhân là lãnh đạo của doanh nghiệp.

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 những hiểu biết nhất định về lĩnh vực tài chính và xã hội và đặc thù của từng doanh nghiệp. Trong khi đó, phân tích kỹ thuật rất trực quan dựa trên biểu đồ và các chỉ báo kỹ thuật, 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 cơ bản là việc nghiên cứu các chỉ số tài chính của doanh nghiệp kết hợp với tình hình kinh tế vĩ mô để khám phá tiềm năng, giá trị nội tại, nhằm mục đích xác định giá trị thực của doanh nghiệp. Qua đó so sánh với giá trị thị trường hiện tại của doanh nghiệp để đưa ra quyết định đầu tư.

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:

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

$X_t$  là ma trận kích thước  $M*N$  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:

$$\hat{p}_{t+1} = f(X_t)$$

Với hàm mục tiêu là:

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

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.

Dựa theo mục đích sử dụng, 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.

#### 1.1.1.1 Nhóm chỉ báo về xu hướng

Chỉ báo theo xu hướng giúp trả lời cho câu hỏi liệu giá của chứng khoán nhìn chung đang có xu hướng tăng, giảm hay đi ngang và độ mạnh của xu hướng đó.

Một số chỉ báo theo xu hướng phổ biến:

##### ***Đường trung bình động (MA)***

Đường trung bình động là mức trung bình của các giá trị (thường là giá) trong một khoảng thời gian cụ thể. Mỗi điểm dữ liệu được tính bằng dữ liệu từ  $n$  kỳ trước đó,  $n$  có thể là 10, 20, 50, 100, 200, ... Bởi vì tính giá trị trung bình dữ liệu trước đó nên các đường trung bình động sẽ làm mịn dữ liệu giá để tạo thành một chỉ báo theo xu hướng [24].

Đường trung bình động không dự đoán hướng giá. Thay vào đó, nó xác định hướng hiện tại. Tuy nhiên, đường trung bình động có xu hướng trễ vì nó dựa trên giá trong quá khứ. Mặc dù vậy, các nhà đầu tư sử dụng đường trung bình động để giúp hành động giá và lọc nhiễu.

Đường trung bình động có thể được sử dụng để xác định xu hướng hoặc xác định các mức hỗ trợ và kháng cự tiềm năng. Chúng cũng tạo thành các khối xây dựng cho nhiều chỉ báo kỹ thuật chẳng hạn như dải Bollinger, MACD.

Hai đường trung bình động phổ biến nhất là đường trung bình động đơn giản (SMA) và đường trung bình động hàm mũ (EMA). Đường trung bình động đơn giản (SMA) có giá trung bình trong khung thời gian được chỉ định, trong khi đường trung bình động hàm mũ (EMA) sử dụng trọng số tỷ lệ với thời gian.

#### 1.1.1.2 Nhóm chỉ báo về biến động

Chỉ báo biến động là các chỉ báo kỹ thuật dùng để đo lường mức độ dao động của giá trong một khoảng thời gian. Thay vì cho biết xu hướng tăng hay giảm, các chỉ báo này tập trung vào việc đánh giá biên độ biến động – tức mức độ giá thay đổi mạnh hay yếu của xu hướng.

##### ***Dải Bollinger***

Được phát triển bởi John Bollinger, dải Bollinger là các dải biến động được đặt trên và dưới đường trung bình động. Độ biến động dựa trên độ lệch chuẩn, độ lệch này thay đổi khi độ biến động tăng và giảm. Các dải tự động mở rộng khi độ biến động tăng và co lại khi độ biến động giảm. Dải Bollinger bao gồm một dải giữa với hai dải biên trên và dưới. Dải giữa là một đường trung bình động đơn giản (SMA) thường được đặt ở mức 20. Đường trung bình động đơn giản được sử dụng vì công thức độ lệch chuẩn cũng sử dụng đường trung bình động đơn giản. Hai dải trên và dưới được đặt cách giải giữa một khoảng bằng với độ lệch chuẩn [26].

Dải Bollinger phản ánh hướng với đường trung bình động đơn giản và sự biến động với biên trên và biên dưới. Dùng để xác định giá tương đối cao hay thấp. Giá cổ phiếu trong phần lớn thời điểm sẽ nằm trong dải Bollinger, điều này khiến cho việc di chuyển ra ngoài dải là đáng để chú ý. Về mặt kỹ thuật, giá tương đối cao khi ở trên dải Bollinger trên và tương đối thấp khi ở dưới dải Bollinger dưới.

#### 1.1.1.3 Nhóm chỉ báo về dao động

Chỉ báo dao động tập hợp những chỉ báo có giá trị dao động trong một khoảng nhất định, thường là từ 0 đến 100. Khi chỉ báo chạm đến các ngưỡng cực cao hoặc cực thấp, nó phát tín hiệu rằng giá có thể sắp đảo chiều.

Khi giá trị của chỉ báo dao động tiếp cận giá trị tối đa, các nhà phân tích kỹ thuật giải thích thông tin đó có nghĩa là tài sản bị mua quá mức và khi chỉ báo tiếp cận mức giá trị tối thiểu, điều đó có nghĩa là tài sản bị bán quá mức.

Một trong những chỉ báo dao động được sử dụng rộng rãi là chỉ số sức mạnh tương đối RSI, một chỉ báo dao động động lượng do J. Welles Wilder phát triển, đo tốc độ và sự thay đổi của biến động giá. Chỉ số chỉ số sức mạnh tương đối RSI di chuyển lên và xuống (dao động) trong khoảng từ 0 đến 100. Khi chỉ số RSI trên 70, nó thường biểu thị tình trạng cổ phiếu đi vào vùng quá mua. Khi chỉ số RSI dưới 30, thể hiện tình trạng cổ phiếu đi vào vùng quá bán [27].

#### 1.1.1.4 Nhóm chỉ báo khối lượng

Chỉ báo khối lượng là các chỉ báo kỹ thuật sử dụng dữ liệu khối lượng giao dịch (lượng cổ phiếu được mua bán trong một khoảng thời gian nhất định) để đánh giá sức mạnh hoặc độ tin cậy của một xu hướng giá. Bởi vì xu hướng giá phải đi kèm với biến động về khối lượng tương ứng mới là xu hướng đáng tin cậy, nếu giá tăng hoặc giảm mà



không có sự đồng thuận từ khối lượng, thì xu hướng đó có thể yếu hoặc không bền vững.

Chỉ báo khối lượng cân bằng (OBV) là một chỉ báo đánh giá áp lực mua và bán chứng khoán bằng cách phân tích khối lượng tích lũy. Nó tăng thêm khối lượng vào những ngày giá tăng và trừ đi vào những ngày giá giảm [28].

## 1.2 Các nghiên cứu liên quan

Dữ liệu dạng chuỗi thời gian nói chung hay dữ liệu giá cổ phiếu trên thị trường chứng khoán nói riêng đã được quan tâm và nghiên cứu từ lâu. Những mô hình thống kê như trung bình động tích hợp hồi quy tự động (ARIMA) đã được sử dụng. Tuy nhiên, một hạn chế của ARIMA là xu hướng tập trung vào các giá trị trung bình của dữ liệu chuỗi quá khứ. Do đó, vẫn còn nhiều thách thức để nắm bắt một quá trình thay đổi nhanh chóng.

Hồi quy vector hỗ trợ (SVR), một công cụ mạnh mẽ cho hồi quy phi tuyến đã được áp dụng cho bài toán dự báo giá cổ phiếu, tuy nhiên, SVR đòi hỏi khá nhiều ở phần tiền xử lý dữ liệu vì không tự xử lý được thông tin dạng chuỗi thời gian và mô hình có nhiều siêu tham số.

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

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.

#### **Phương pháp Voting hồi quy**

Trong nghiên cứu của được công bố vào tháng 6 năm 2023, TS. Nguyễn Hữu Xuân Trường, Trương Hải Nam, Nguyễn Ngọc Ánh (Khoa Kinh tế số, Học viện Chính sách và Phát triển) và ThS. Nguyễn Sao Mai (Bộ môn Công nghệ thông tin, Đại học Lao động Xã hội) đã thực hiện nghiên cứu dự báo giá cổ phiếu sử dụng phương pháp Voting hồi quy, một phương pháp kết hợp 3 mô hình học máy là hồi quy Ridge, hồi quy BayesianRidge và hồi quy ElasticNet, kết quả cuối cùng là trung bình kết quả của cả 3

mô hình. Sử dụng cả những chỉ báo kỹ thuật làm đặc trưng, mô hình cho độ dự báo chính xác tương đối cao [4].

### **Mô hình XGBoost**

Trong nghiên cứu của mình năm 2023, Trương Thị Thùy Dương đã thực hiện sử dụng thuật toán tăng cường XGBoost, kết hợp với việc sử dụng các chỉ báo kỹ thuật để tăng độ tốt của mô hình dự đoán xu hướng của chỉ số VNIndex. Điều chú ý của nghiên cứu là việc nghiên cứu đã đơn giản hóa bài toán bằng việc chỉ xác định xu hướng tăng hoặc giảm, chứ không xác định cụ thể số điểm tăng hoặc giảm. Nghiên cứu đã chỉ ra bằng chứng thực nghiệm về tầm quan trọng của các chỉ báo kỹ thuật trong dự báo chiều biến động của chỉ số VNIndex [1].

### **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. Các mô hình được đánh giá kết quả trên tập dữ liệu của chỉ số VNIndex, kéo dài từ ngày 14 tháng 7 năm 2008 đến 14 tháng 7 năm 2023, với 3747 dữ liệu, 80% cho việc huấn luyện và 20% còn lại để đánh giá. Nghiên cứu cho thấy việc sử dụng các biến thể mô hình mạng nơ-ron 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 nơ-ron 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 2021, A. Chatterjee, H. Bhowmick và J. Sen đã thử nghiệm trên dữ liệu của 3 cổ phiếu đại diện cho 3 ngành công nghiệp khác nhau gồm ICICI Bank (ngân hàng), Sun Pharma (y tế), Infosys (công nghệ) được niêm yết trên sàn chứng khoán Ấn Độ. Có 4 loại mô hình khác nhau được thử nghiệm, cho thấy các mô hình học sâu là RNN và LSTM cho hiệu quả ổn định với các tập dữ liệu khác nhau và LSTM có độ chính xác cao hơn RNN nhờ khả năng xử lý dữ liệu tuần tự phức tạp do không gặp phải vấn đề về độ dốc biến mất và độ dốc bùng nổ [6].

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**

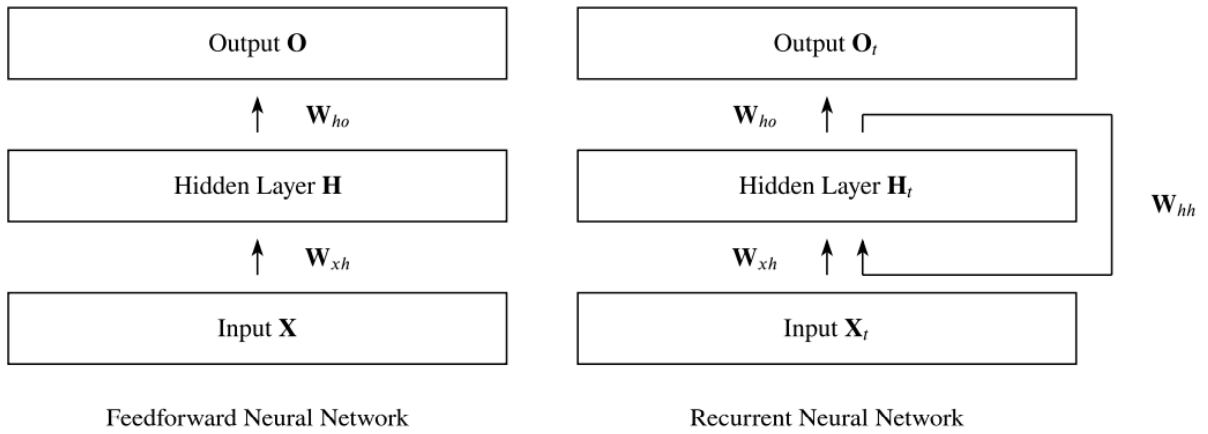
Chương này tạo tiền đề cho đề án bằng cách giới thiệu bài toán phân tích giá cổ phiếu, trong phần 1.2, đề án đi sâu vào bối cảnh nghiên cứu của bài toán ở cả trong và ngoài nước. Chương tiếp theo đề án sẽ trình bày khung cơ sở lý thuyết của các thành phần trong giải pháp được đề xuất.

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

### 2.1 Giới thiệu mô hình RNN

Những năm gần đây, lĩnh vực trí tuệ nhân tạo đã có được rất nhiều bước tiến đột phá với đóng góp không nhỏ của mạng nơ-ron hồi quy RNN. RNN là một dạng mạng nơ-ron nhân tạo chuyên dùng để xử lý dữ liệu tuần tự như văn bản, chuỗi thời gian, âm thanh, hoặc video. Khác với mạng nơ-ron truyền thống (Feedforward Neural Network), RNN có cơ chế ghi nhớ thông tin từ các bước trước đó thông qua trạng thái ẩn, giúp mô hình hiểu được mối liên hệ theo thời gian giữa các phần tử trong chuỗi.

#### 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:

$$H_t = \phi_h(X_t W_{xh} + H_{t-1} W_{hh} + b_h) \quad (1) \quad [20]$$

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

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

$$H = \phi_h(X W_{xh} + b_h) \quad (3) \quad [20]$$

$$O = \phi_o(H W_{ho} + b_o) \quad (4)$$

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

$\phi_h$  và  $\phi_o$ : 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 nơ-ron hồi quy hai chiều (BRNN), mạng nơ-ron 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**

Mạng nơ-ron hồi quy (RNN) là mô hình học máy mạnh mẽ và đã được sử dụng trong nhiều lĩnh vực. Nó khác biệt rõ rệt với các mô hình mạng tích chập (CNN) như GoogleNet được ứng dụng trong lĩnh vực thị giác máy tính. Sau đây là các ứng dụng khác nhau của RNN và các biến thể, với trọng tâm chính là khả năng sử dụng dữ liệu tuần tự.

Sinh văn bản: với một chuỗi các từ làm đầu vào cho trước, mạng sẽ dự đoán khả

năng xuất hiện của từ tiếp theo. Mô hình hoàn thiện những văn bản chưa hoàn chỉnh như là viết tiếp một câu chuyện hoặc bài thơ,... Ví dụ cụ thể có thể kể đến như nền tảng MosaicML, một nền tảng phát triển AI được thiết kế để giải quyết những thách thức trong việc huấn luyện và cung cấp các mô hình AI.

Tóm tắt văn bản: mô hình RNN và các biến thể được ứng dụng rộng rãi vào bài toán tóm tắt văn bản, ví dụ như việc tóm tắt bài báo, mẫu chuyện hoặc phức tạp hơn như là đoạn hội thoại.

Dịch máy: các biến thể của RNN có thể được sử dụng để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác. Phần lớn các hệ thống dịch thuật đang được sử dụng ngày nay đều sử dụng một số phiên bản nâng cao của RNN. Đầu vào có thể là ngôn ngữ nguồn và đầu ra sẽ là ngôn ngữ đích mà người dùng muốn. Hiện tại, một trong những ứng dụng dịch máy phổ biến và nổi bật nhất là Google Translate. Thậm chí còn có nhiều ứng dụng RNN được sử dụng để tinh chỉnh và giới hạn nội dung trên nhiều nền tảng khác nhau. Các nền tảng thương mại điện tử như Flipkart, Amazon và eBay sử dụng dịch máy trong nhiều lĩnh vực và nó cũng giúp nâng cao hiệu quả của kết quả tìm kiếm.

Nhận dạng giọng nói: RNN có thể được sử dụng để dự đoán các phân đoạn ngữ âm coi sóng âm thanh từ môi trường làm nguồn đầu vào. Bộ đầu vào là tín hiệu âm thanh được xử lý theo cách thích hợp. Tín hiệu âm thanh đầu vào được phân loại thành các phân đoạn ngữ âm và được ghép lại với nhau thành các từ gắn kết bằng ứng dụng RNN. Đầu ra bao gồm một mẫu các phân đoạn ngữ âm được ghép lại với nhau thành một tổng thể duy nhất một cách hợp lý.

Tạo mô tả hình ảnh: với việc kết hợp giữa CNN và RNN được sử dụng để cung cấp mô tả chính xác những gì đang xảy ra bên trong hình ảnh. CNN thực hiện phân đoạn và RNN sau đó sử dụng dữ liệu được phân đoạn để tạo mô tả cho ảnh.

Gắn thẻ video: RNN có thể được sử dụng để tìm kiếm video. Bằng việc thêm mô tả cho từng đoạn các ảnh.

Ngoài ra, các biến thể của RNN còn được sử dụng trong việc dự đoán các dữ liệu dạng chuỗi thời gian khác như là nhiệt độ thời tiết và đặc biệt là việc 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. Trong quá trình huấn luyện, mô hình RNN thực hiện phép nhân ma trận trên chuỗi dữ liệu (có thể rất dài), nếu có các giá trị nhỏ ( $< 1$ ) trong phép nhân ma trận, điều này khiến độ dốc giảm theo từng lớp (hoặc bước thời gian) và cuối cùng biến mất. Điều này sẽ làm giảm dần sự đóng góp của các thông tin ở xa so với bước thời gian hiện tại. Ngược lại, nếu các giá trị lớn ( $> 1$ ) trong phép nhân ma trận có thể gây ra hiện tượng độ dốc bùng nổ [20]. Việc độ dốc bùng nổ hoặc biến mất dẫn đến việc các mô hình RNN khó hoạt động tốt với các chuỗi dữ liệu dài và xử lý các phụ thuộc xa.

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.

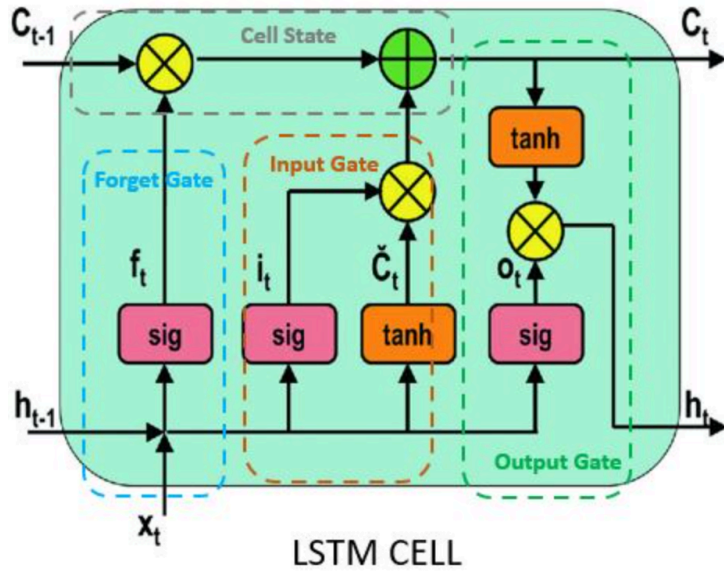
Ngoài ra, cũng giống như những mô hình học sâu khác, với nhiều cổng, ô trạng thái và trạng thái ẩn khiến việc nắm bắt trực quan cách thông tin được xử lý qua mạng là một vấn đề lớn, dẫn đến sự khó khăn trong việc lý giải mô hình.

## **2.2 Giới thiệu mô hình LSTM**

Mạng bộ nhớ dài ngắn hạn (LSTM) là một biến thể mạng nơ-ron hồi quy (RNN), có khả năng xử lý tình trạng mất mát thông tin trong quá trình học bằng việc xử lý vấn đề biến mất và bùng nổ độ dốc mà RNN gặp phải. LSTM được thiết kế bởi Hochreiter và Schmidhuber [13] nhằm giải quyết vấn đề do thuật toán học máy và RNN truyền thống gây ra. Sự ra đời của mạng LSTM đánh dấu một tiến bộ đáng kể trong lĩnh vực trí tuệ nhân tạo nói chung và học sâu nói riêng.

### **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.

**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.

Điểm đột phá của LSTM ở chỗ đã tạo thêm một nơi để lưu trữ thông tin là ô



trạng thái, thay vì chỉ trạng thái ẩn như RNN. Kiến trúc của mạng LSTM là một tiến bộ đáng kể trong thiết kế mạng nơ ron nhân tạo, đặc biệt đối với các bài toán liên quan đến dữ liệu tuần tự. Bằng cách tạo thêm ô trạng thái để lưu trữ thông tin có chọn lọc, LSTM có thể duy trì thông tin liên quan của các chuỗi dài, khiến chúng trở nên lý tưởng cho nhiều ứng dụng trong xử lý ngôn ngữ tự nhiên, phân tích dữ liệu dạng chuỗi thời gian. Khả năng giải quyết những thách thức về sự phụ thuộc dài của các dữ liệu dạng chuỗi khiến LSTM được sử dụng rộng rãi.

### **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. Khả năng khái quát hóa của mô hình đối với dữ liệu mới, chưa biết bị giảm đi khi nó khớp quá mức với huấn luyện, bao gồm nhiễu và giá trị ngoại lai [16].

Sự ổn định của các mô hình LSTM cũng là một vấn đề, đặc biệt là trong các kiến trúc nhiều lớp. Các khó khăn về độ dốc biến mất và bùng nổ là những vấn đề thường gặp có thể dẫn đến việc cập nhật không ổn định cho các tham số mô hình hoặc mô hình không hội tụ [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 được lấy về dưới dạng JSON theo từng tệp riêng bằng việc gọi API theo kỹ thuật phân trang dữ liệu. 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. Các tệp dữ liệu sau khi được tải về, sẽ được kiểm tra và đổ vào trong cơ sở dữ liệu dạng bảng, với 7 thông tin quan trọng gồm có mã cổ phiếu, ngày giao dịch, giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa và khối lượng giao dịch. Để tổng quát nhất, bài toán sẽ tập trung vào dữ liệu của chỉ số VNIndex, chỉ số chung đại diện cho thị trường chứng khoán Việt Nam nhằm tránh những cổ phiếu cụ thể dễ bị ảnh hưởng bởi một nhóm nhỏ các nhà đầu tư.

**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. Kết quả của bước này, là các tệp chứa tên những đặc trưng quan trọng nhất theo số lượng nhất định, cụ thể là tệp chứa tên một chỉ báo quan trọng nhất, bốn chỉ báo quan trọng nhất và mười chỉ báo quan trọng nhất sử dụng cho việc xây dựng mô hình dự báo.

**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.

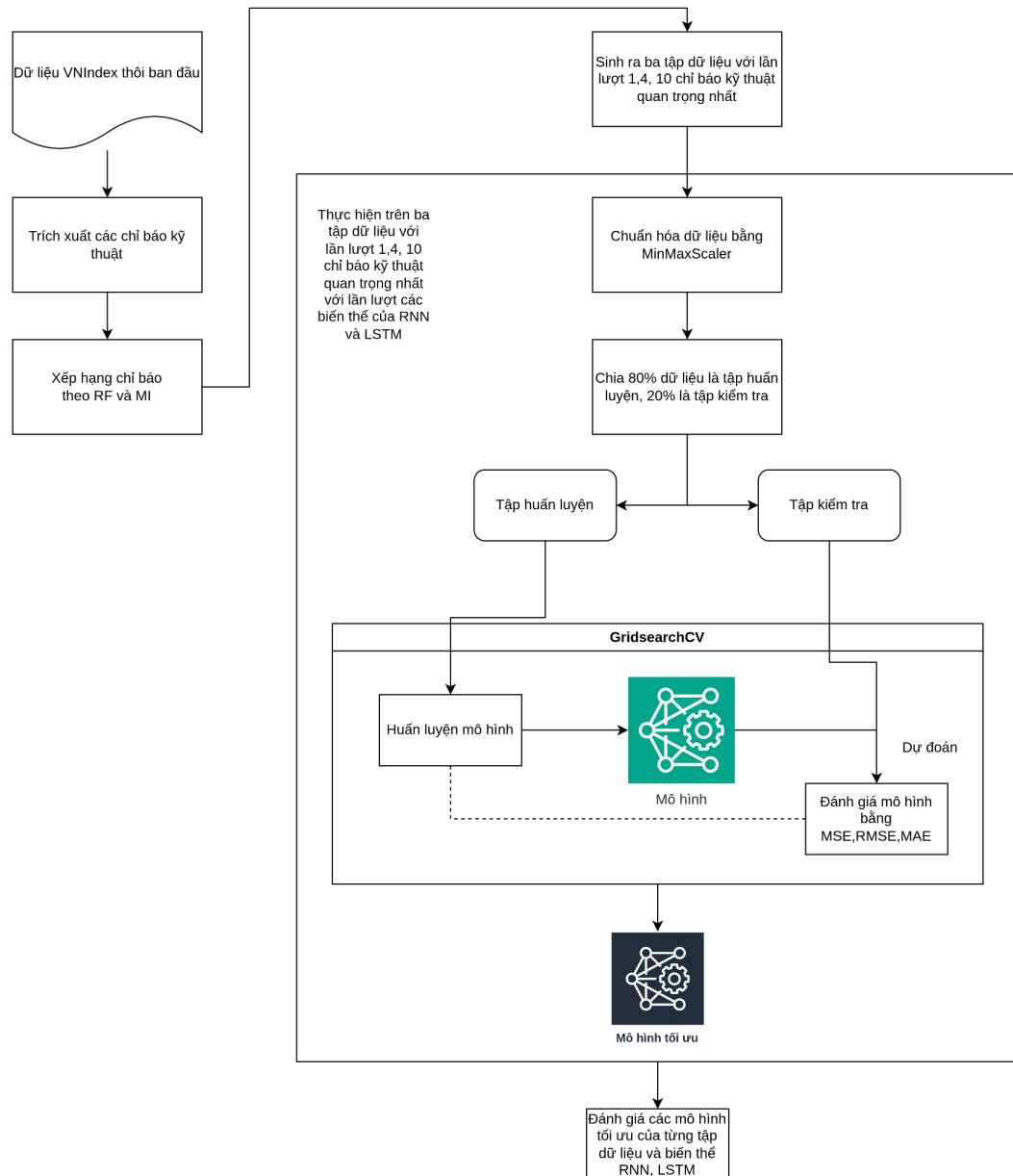
Đề án thực hiện trích xuất các chỉ báo kỹ thuật và sử dụng kỹ thuật học máy RF và MI để đánh giá độ liên quan của các chỉ báo kỹ thuật đối với giá trị đóng cửa của cổ phiếu để lọc bỏ những chỉ báo kỹ thuật gây nhiễu. Sau đó dữ liệu sẽ được chuẩn hóa để các giá trị về trong khoảng  $[0,1]$  sau đó được đưa vào để phân tách tập huấn luyện với 80% dữ liệu và tập kiểm tra sẽ là 20%



**Hình 3.1 Chỉ số VN Index**

**Bảng 3.1 Thông tin chung của chỉ số VNIndex**

| count | mean   | 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.

### **Đường trung bình động (MA)**

Đường trung bình động là mức trung bình của các điểm dữ liệu (thường là giá) trong một khoảng thời gian cụ thể. Mỗi điểm dữ liệu được tính bằng dữ liệu từ  $n$  kỳ trước đó,  $n$  có thể là 10, 20, 50, 100, 200, ... Vì sử dụng phép tính trung bình dữ liệu trước đó nên các đường trung bình động sẽ làm mịn dữ liệu giá để tạo thành một chỉ báo theo xu hướng.

#### **Đường trung bình động đơn giản (SMA)**

Tính bằng cách lấy trung bình cộng của giá trong khoảng thời gian  $n$  khi biết giá trị cổ phiếu tại thời điểm  $n$  là  $A_n$ :

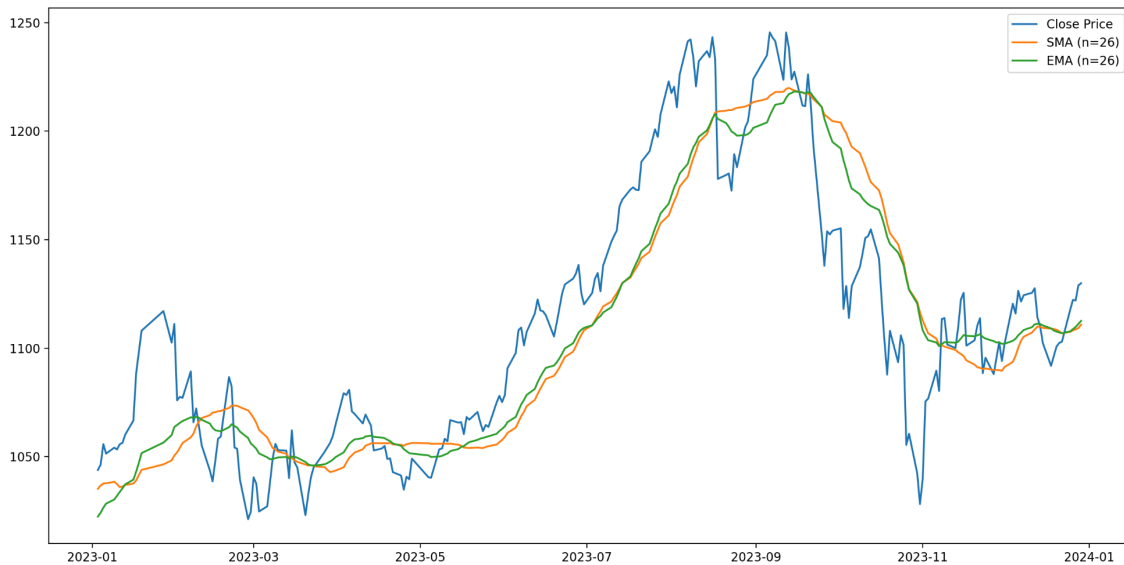
$$SMA = \frac{A_1 + A_2 + \dots + A_n}{n} \quad [24]$$

#### **Đường trung bình động lũy thừa (EMA)**

Gán trọng số để làm tăng mức độ ảnh hưởng cho các giá trị gần, EMA tại thời điểm  $t$  được xác định khi biết giá trị  $P_t$  được tính bởi:

$$EMA_t = \alpha * P_t + (1 - \alpha)EMA_{t-1} \quad [25]$$

$$\text{với } \alpha = \frac{2}{n+1}$$



**Hình 3.3 Chỉ báo đường trung bình động**

### **Dải Bollinger**

Dải Bollinger là các dải biến động được đặt trên và dưới đường trung bình động. Độ biến động dựa trên độ lệch chuẩn, độ lệch này thay đổi khi độ biến động tăng và giảm. Dải Bollinger sẽ mở rộng khi độ biến động tăng và co lại khi độ biến động giảm.

Với:

- $P_t$  là giá của cổ phiếu ngày  $t$
- $\sigma$  là độ lệch chuẩn của giá trong  $n$  phiên
- $k$  là hệ số độ lệch (thường là 2)

Ta có công thức tính các chỉ báo của dải Bollinger:

Đường trung tâm chính là SMA (thường với  $n=20$ ):

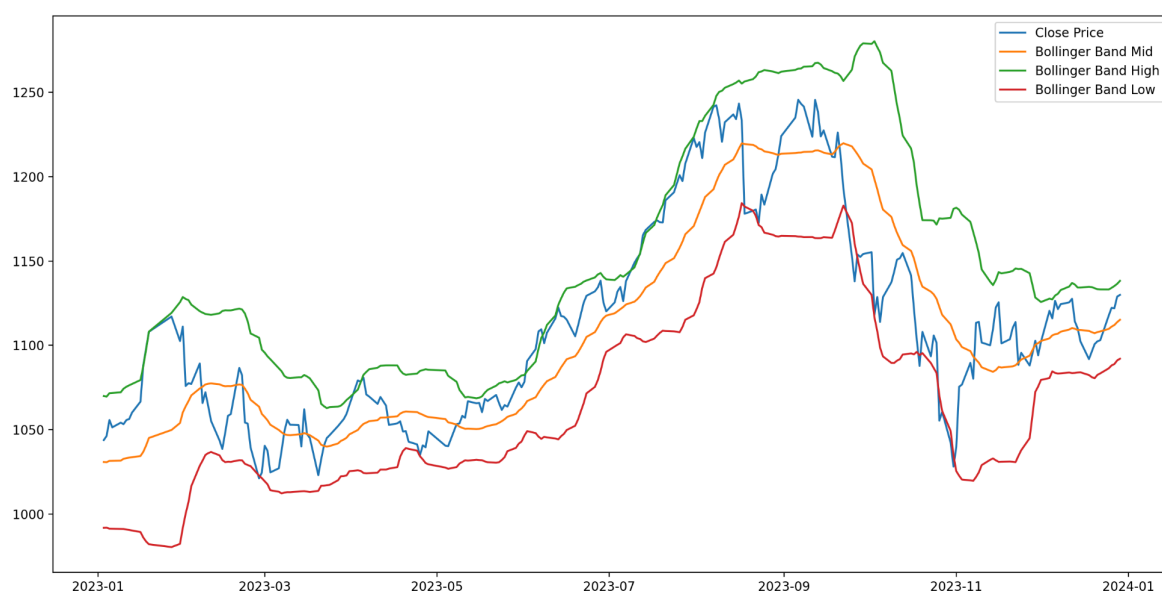
$$SMA_t = \frac{P_t + P_{t-1} + \dots + P_{t-n+1}}{n} \quad [26]$$

Dải trên:

$$BBH_t = SMA_t + k \cdot \sigma \quad [26]$$

Dải dưới:

$$BBL_t = SMA_t - k \cdot \sigma \quad [26]$$



**Hình 3.4 Chỉ báo dải Bollinger**

### **Chỉ số sức mạnh tương đối (RSI)**

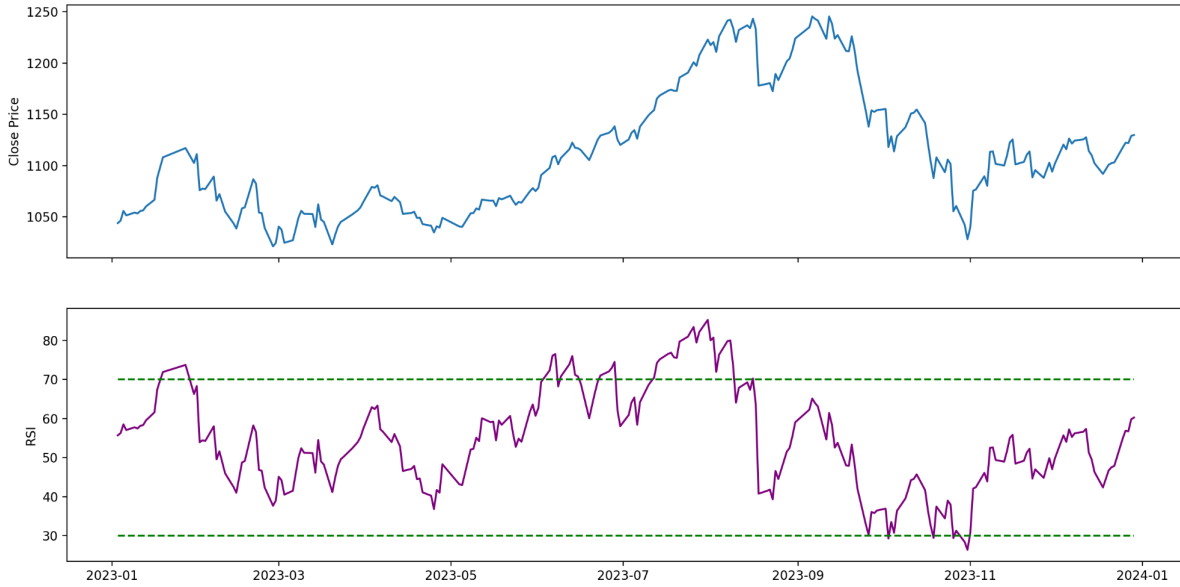
Chỉ số chỉ số sức mạnh tương đối RSI di chuyển lên và xuống (dao động) trong khoảng từ 0 đến 100. Khi chỉ số RSI trên 70, nó thường biểu thị tình trạng cổ phiếu đi

vào vùng quá mua. Khi chỉ số RSI dưới 30, thể hiện tình trạng cổ phiếu đi vào vùng quá bán. Với

- $P_t$  là giá của cổ phiếu ngày  $t$
- $\Delta P_t = P_t - P_{t-1}$  [27]
- $AVG\_GAIN_n = \frac{\sum_{i=1}^n \max(0, \Delta P_i)}{n}$  [27]
- $AVG\_LOSS_n = \frac{\sum_{i=1}^n \text{abs}(\min(0, \Delta P_i))}{n}$  [27]
- $RS_n = \frac{AVG\_GAIN_n}{AVG\_LOSS_n}$  [27]

Công thức tính chỉ số sức mạnh tương đối được tính bởi công thức ( $n$  là số ngày quan sát, thường là 14):

$$RSI = 100 - \frac{100}{1 + RS_n} \quad [27]$$



**Hình 3.5 Chỉ báo sức mạnh tương đối RSI**

### ***Chỉ báo khối lượng cân bằng (OBV)***

Chỉ báo khối lượng cân bằng (OBV) là một chỉ báo đánh giá áp lực mua và bán chứng khoán bằng cách phân tích khối lượng tích lũy. Nó tăng thêm khối lượng vào những ngày giá tăng và trừ đi vào những ngày giá giảm. Với

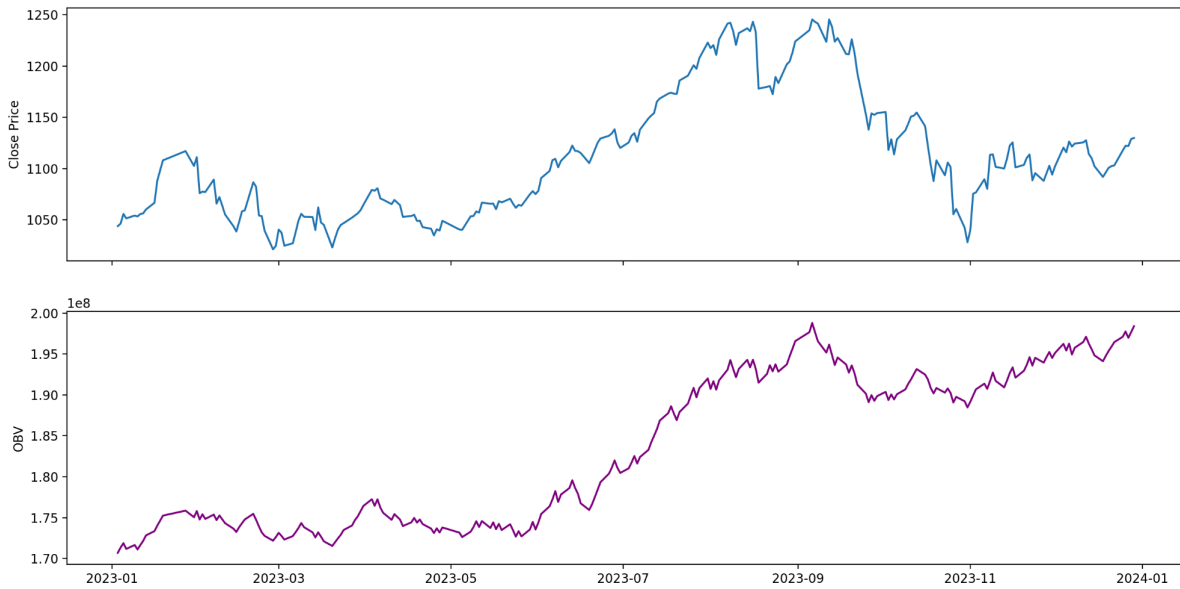
- $P_t$  là giá của cổ phiếu ngày  $t$

- $V_t$  là khối lượng giao dịch của cổ phiếu ngày  $t$
- Hàm  $sign(x)$ :

$$sign(x) = \begin{cases} -1 & \text{nếu } x < 0 \\ 1 & \text{nếu } x > 0 \\ 0 & \text{nếu } x = 0 \end{cases}$$

Chỉ báo khối lượng cân bằng (OBV) được tính bởi:

$$OBV_t = OBV_{t-1} + V_t \cdot sign(P_t - P_{t-1}) \quad [28]$$



**Hình 3.6 Chỉ báo khối lượng cân bằng OBV**

### 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

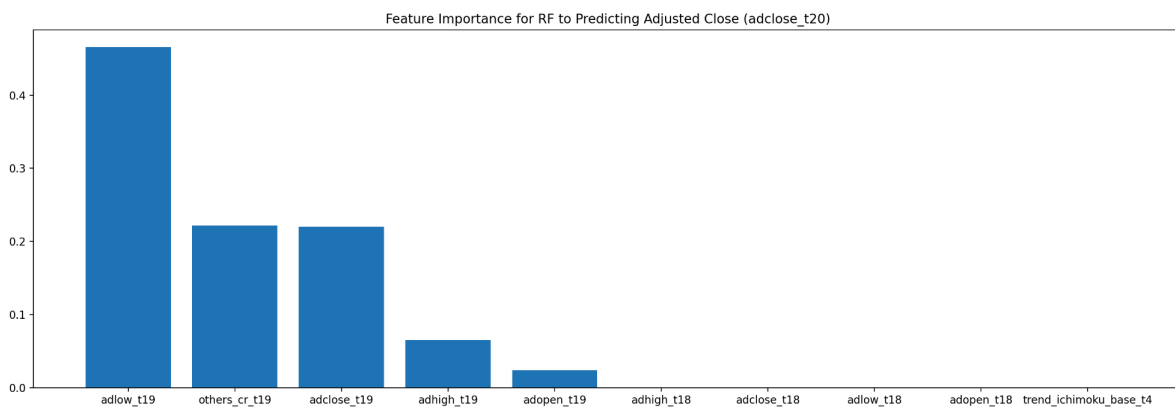
Trong nghiên cứu của mình năm 2023 về những kỹ thuật trích chọn và tổng hợp đặc trưng cho bài toán giá cổ phiếu của Htun và cộng sự, đã nghiên cứu tổng cộng 208



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.

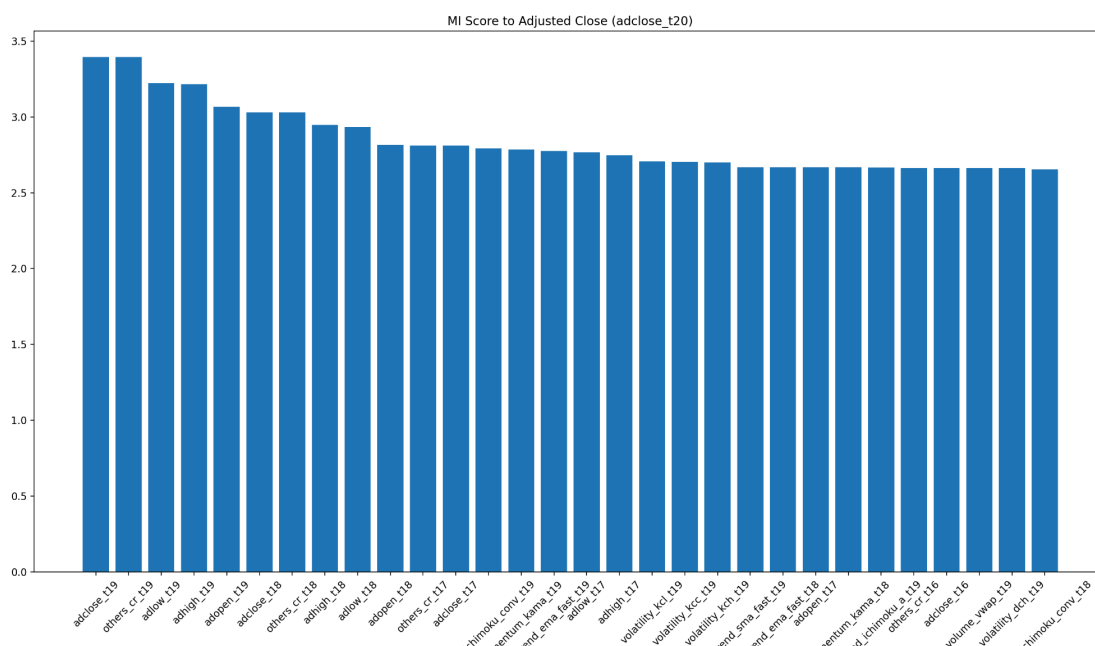
Dữ liệu ban đầu được chuẩn hóa bằng bộ chuẩn hóa dữ liệu theo phân phối chuẩn, sau đó sẽ được đưa vào mô hình RF hồi quy với số lượng cây quyết định là 100. Dưới đây là độ quan trọng của các chỉ báo đối với mô hình RF hồi quy với 4 chỉ báo kỹ thuật cơ bản là giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa.



**Hình 3.7 Độ quan trọng của các chỉ báo theo RF**

### 3.1.2.2 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.

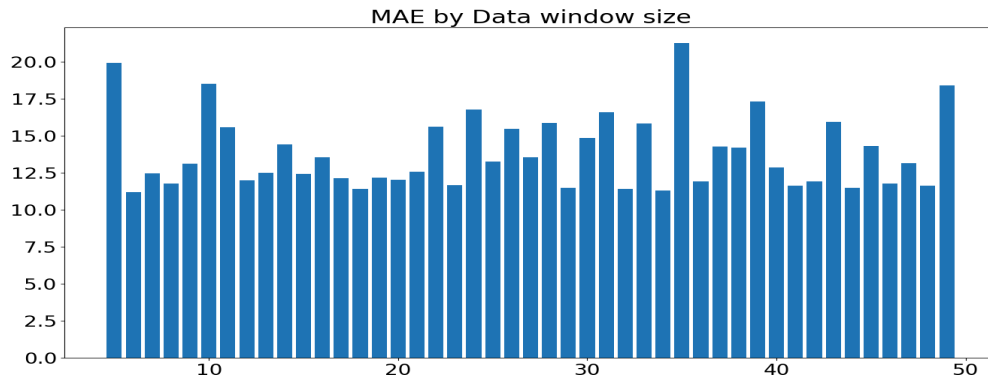
Dựa trên kết quả của việc đánh giá độ quan trọng của các chỉ báo, đề án sẽ chia dữ liệu thành ba tập, tập có một đặc trưng, tập có bốn đặc trưng và tập sử dụng mười đặc trưng.

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.

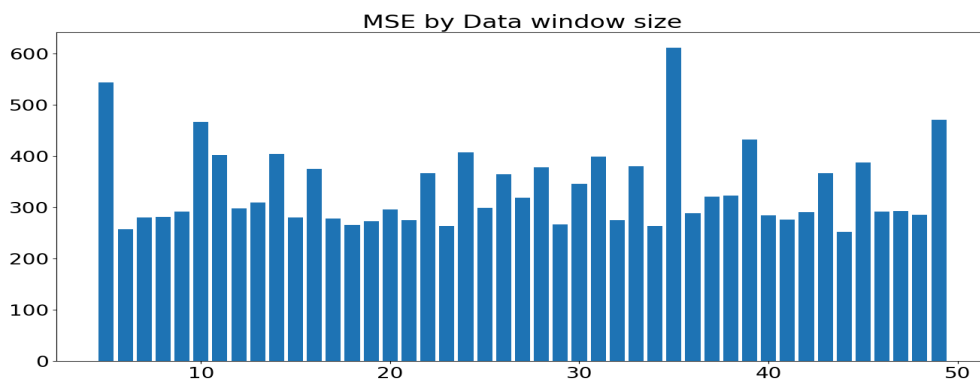
### 3.2 Cài đặt thử nghiệm

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

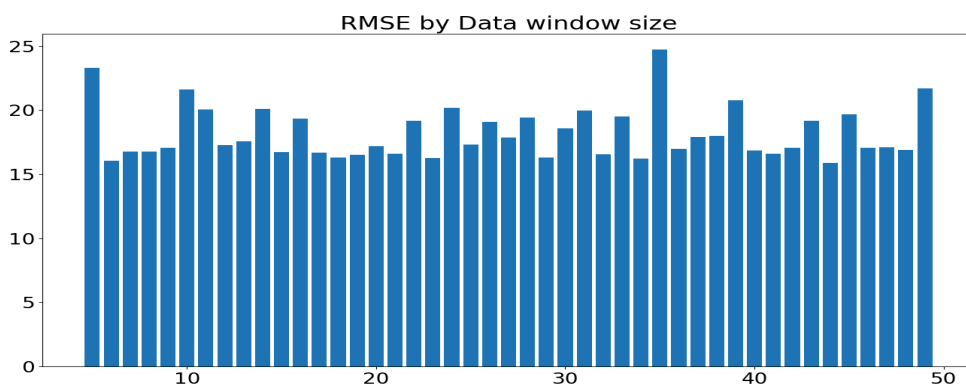
đây là khảo sát chọn giá trị tối ưu đối với siêu tham số là kích thước cửa sổ (window size) trên tập kiểm tra.



**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.

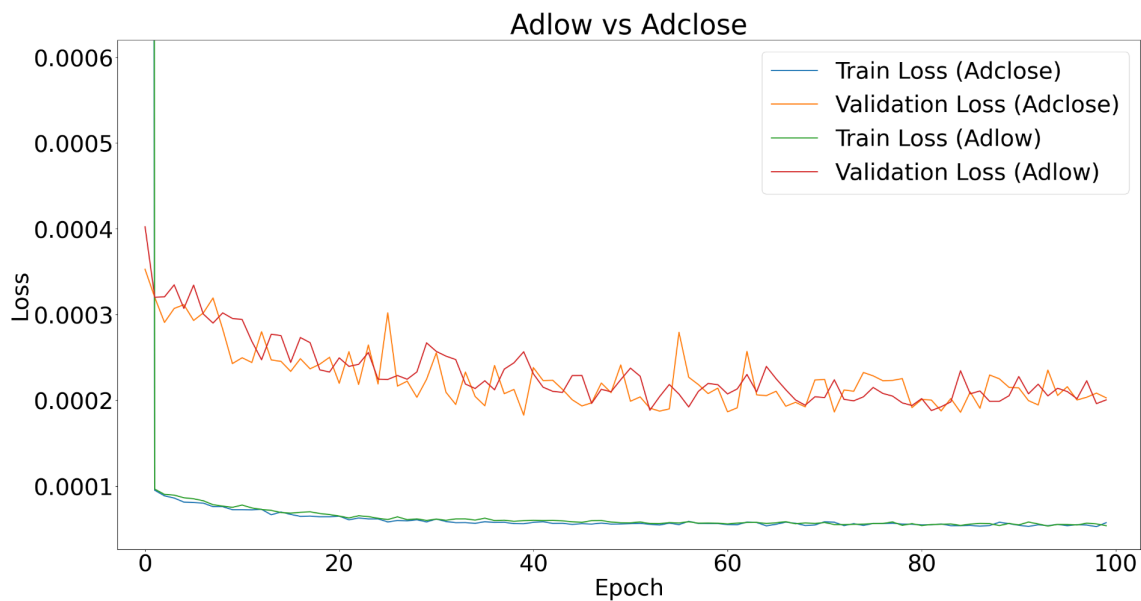
Ngoài ra, đề án sử dụng mô hình mạng nơ-ron với kích thước ô nhớ (batch size) là 32. Hàm relu được sử dụng làm hàm kích hoạt trong quá trình đào tạo mô hình. Sử dụng MSE cho hàm mất mát, và thuật toán Adam làm phương pháp tối ưu các tham số của mô hình. Với  $y_i$  là giá trị chính xác của dữ liệu thứ  $i$ , và  $\bar{y}_i$  là giá trị mô hình dự đoán đối với dữ liệu thứ  $i$ , công thức của hàm mất mát như sau:

$$MSE = \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{n} \quad [29]$$

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 nơron 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 nơron 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         |
| adclos           | 66.17155475                    | 8.134590017 | 5.415159344 | 249.2014303                  | 15.78611511 | 11.14971804 |
| adlow            | 67.73525169                    | 8.230142872 | 5.492457143 | 250.1422082                  | 15.81588468 | 11.43820282 |



**Hình 3.12 Biểu đồ hàm mất mát của mô hình trên hai tập dữ liệu**

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 tương ứng với 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_ichimoku_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).

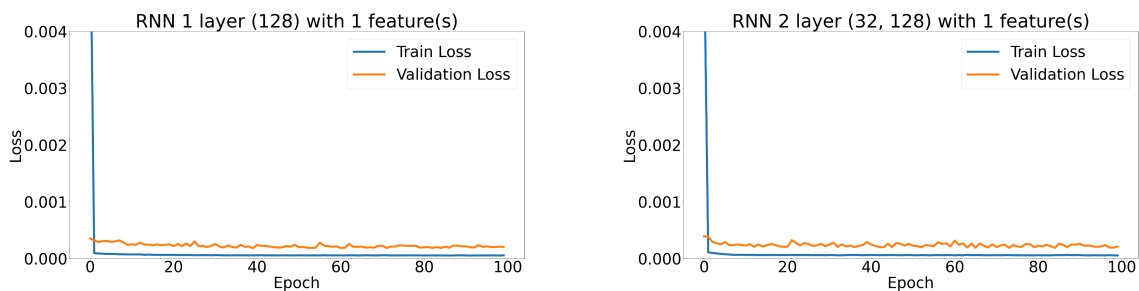
Đề án sử dụng phương pháp tìm kiếm lưới (GridsearchCV) sẽ sử dụng số lần lặp  $n=5$ . Với mô hình một lớp, số nơ ron trong lưới tìm kiếm là 32, 64, 128, 256 và 512 còn với mô hình hai lớp, số nơ ron trong lưới tìm kiếm là 32, 128 và 256. Trong khi đó, tham số epoch trong lưới tìm kiếm là 50, 80 và 100.

### 3.2.1.1 Tinh chỉnh các siêu tham số của mô hình

Sau khi đã triển khai cài đặt ban đầu, bước tiếp theo là tinh chỉnh các siêu tham số của từng phiên bản và theo dõi sự thay đổi độ tốt của các mô hình. Đặc biệt là sự ảnh hưởng của số nút của từng lớp RNN.

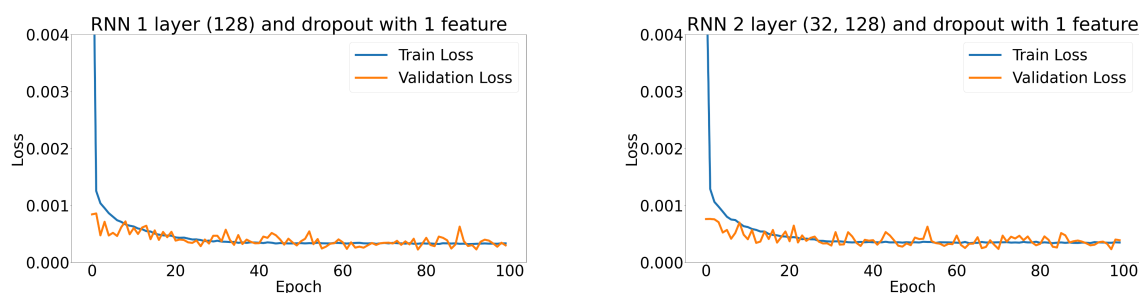
Vì mô hình RNN có một lớp RNN với tập dữ liệu có một đặc trưng đã được tối ưu ở phần trước, phần này sẽ tập trung đi tìm bộ siêu tham số tối ưu cho mô hình RNN có hai lớp RNN. Tập dữ liệu sử dụng ở phần này là tập dữ liệu có một đặc trưng.

Sau khi chạy tìm kiếm tối ưu với mô hình RNN hai lớp ẩn, ta tìm được số nơ ron tối ưu cho lớp RNN thứ nhất là 32 và lớp RNN thứ hai là 128. Bước tiếp theo là chạy lại mô hình và tiến hành khảo sát.



**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. Đường màu cam đi xuống chậm, sau khoảng 35-40 epoch thì đi ngang với biên độ hẹp. Đường màu xanh thể hiện giá trị hàm mất mát của mô hình trên tập dữ liệu huấn luyện vẽ góc vuông thể hiện việc hội tụ rất nhanh ngay từ những epochs đầu tiên.



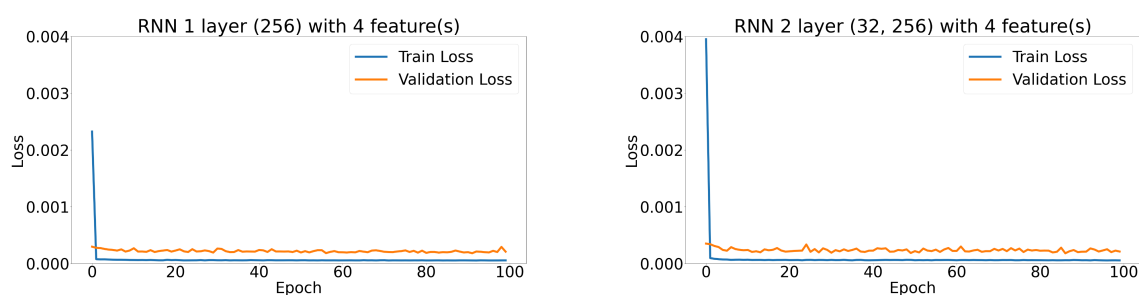
**Hình 3.14 Biểu đồ hàm mất mát của mô hình RNN 1 đặc trưng với dropout=0.2**

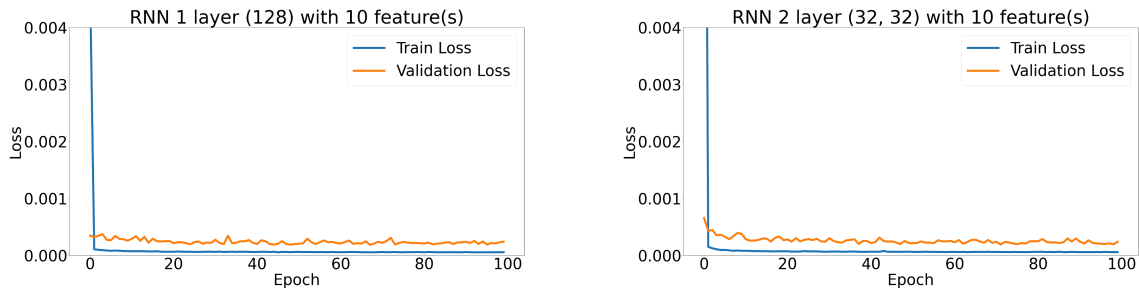
Dựa vào hình 3.11, việc thêm một lớp dropout vào sau lớp RNN với tỷ lệ là 20% đã khiến cho việc hội tụ của hàm mất mát trên tập dữ liệu huấn luyện bị chậm đi, thay vì góc vuông thì giờ đã thành một đường cong và sau khoảng 30 epoch thì đi ngang. Trong khi đường màu cam thể hiện giá trị hàm mất mát của mô hình trên tập dữ liệu kiểm tra hội tụ và gần như đi ngang sau khoảng 20 epoch, tuy nhiên với biên độ khá lớn. Như vậy, 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 trên 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.

Sau quá trình tìm kiếm tham số tối ưu, đối với mô hình có một lớp RNN, tập dữ liệu gồm bốn đặc trưng, số nơron tối ưu là 256. Đối với mô hình có hai lớp RNN, tập dữ liệu gồm bốn đặc trưng, số nơron tối ưu của lớp thứ nhất là 32 và của lớp thứ hai là 256. Đối với mô hình có một lớp RNN, tập dữ liệu gồm mười đặc trưng, số nơron tối ưu là 128. Đối với mô hình có hai lớp RNN, tập dữ liệu gồm mười đặc trưng, số nơron tối ưu của lớp thứ nhất là 32 và của lớp thứ hai là 32.





**Hình 3.15 Biểu đồ hàm mất mát của mô hình RNN 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 nơ-ron, thì sự hội tụ trên tập dữ liệu huấn luyện và tập dữ liệu kiểm tra không bị ảnh hưởng nhiều. Sau khoảng 35-40 epoch thì cả hai đường màu cam và xanh đều đi ngang với biên độ hẹp. Tuy nhiên, đối với mô hình RNN có hai lớp RNN, ta có thể thấy sự hội tụ trên tập huấn luyện chậm hơn dù không nhiều.

### 3.2.2 Thử nghiệm với mô hình LSTM

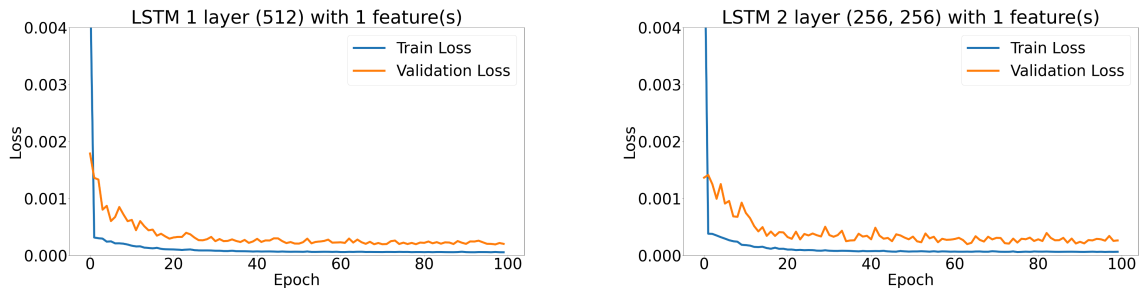
Phần này sẽ đi vào việc thử nghiệm với mô hình LSTM từ đơn giản có 1 lớp LSTM sau đó đến những mô hình phức tạp hơn với 2 lớp LSTM với số nút khác nhau. Việc xác định số nút cho mỗi mô hình sẽ được tối ưu bằng phương pháp tìm kiếm lưới (GridsearchCV) sử dụng số lần lặp  $n=5$ . Với mô hình một lớp, số 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.

#### 3.2.2.1 Tinh chỉnh các siêu tham số của mô hình

Cũng giống như với mô hình RNN, sau khi đã triển khai cài đặt ban đầu, bước tiếp theo là tinh chỉnh các siêu tham số của từng phiên bản và theo dõi sự thay đổi độ tốt của các mô hình. Đặc biệt là sự ảnh hưởng của số nút của từng lớp LSTM.

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 nơ ron tối ưu là 512. Đối với mô hình có hai lớp LSTM, số nơ ron tối ưu là của mỗi lớp là 256.





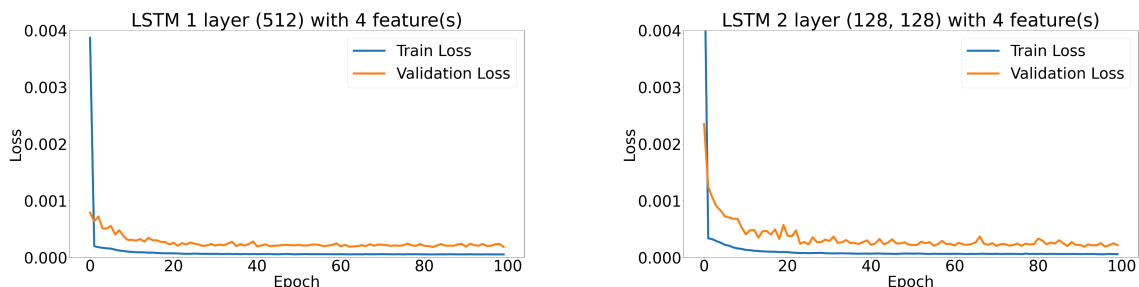
**Hình 3.16 Biểu đồ hàm mất mát của mô hình LSTM với 1 đặc trưng**

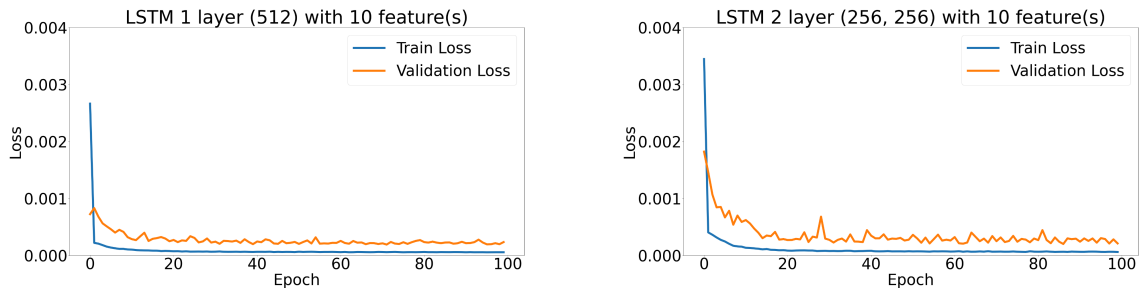
Dựa vào hình 3.13, ta có thể thấy hai mô hình có giá trị hàm mất mát hội tụ khá nhanh trên tập dữ liệu huấn luyện. Sau khoảng 35-40 epoch, đường màu xanh đi ngang với biên độ hẹp khá giống với mô hình RNN. Đường màu cam thể hiện giá trị hàm mất mát trên tập dữ liệu kiểm tra đi ngang sau khoảng 35-40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

### 3.2.2.2 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.

Sau quá trình tìm kiếm tham số tối ưu, đối với mô hình có một lớp LSTM, tập dữ liệu gồm bốn đặc trưng, số nơron tối ưu là 512. Đối với mô hình có hai lớp LSTM, tập dữ liệu gồm bốn đặc trưng, số nơron tối ưu của lớp thứ nhất là 128 và của lớp thứ hai là 128. Đối với mô hình có một lớp LSTM, tập dữ liệu gồm mười đặc trưng, số nơron tối ưu là 512. Đối với mô hình có hai lớp LSTM, tập dữ liệu gồm mười đặc trưng, số nơron tối ưu của lớp thứ nhất và của lớp thứ hai là 256.





**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. 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 đi ngang sau khoảng 35-40 epoch. 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                            | RMSE        | MAE         |
| RNN     | 1         | 128             | 16,769           | 1                  | 66.17155475                    | 8.134590017 | 5.415159344 |
|         | 1         | 256             | 67,073           | 4                  | 65.80623792                    | 8.112104407 | 5.408831615 |
|         | 1         | 128             | 17,921           | 10                 | 66.21609244                    | 8.137327106 | 5.466874152 |
|         | 2         | (32, 128)       | 21,825           | 1                  | 69.68641272                    | 8.347838806 | 5.597459111 |
|         | 2         | (32, 256)       | 75,425           | 4                  | 65.76546545                    | 8.109590955 | 5.470772448 |
|         | 2         | (32, 32)        | 3,489            | 10                 | 77.01293652                    | 8.775701483 | 5.925599524 |
| LSTM    | 1         | 512             | 1,053,185        | 1                  | 66.89107872                    | 8.178696639 | 5.418752841 |
|         | 1         | 512             | 1,059,329        | 4                  | 68.75721691                    | 8.291997161 | 5.513469583 |
|         | 1         | 512             | 1,071,617        | 10                 | 73.04949617                    | 8.546899799 | 5.876194065 |

|  |   |            |         |    |             |             |             |
|--|---|------------|---------|----|-------------|-------------|-------------|
|  | 2 | (256, 256) | 789,761 | 1  | 66.871328   | 8.177489101 | 5.455156859 |
|  | 2 | (128, 128) | 199,809 | 4  | 67.58342843 | 8.220914087 | 5.495880467 |
|  | 2 | (256, 256) | 798,977 | 10 | 70.93502739 | 8.422293476 | 5.630779517 |

**Bảng 3.5 Bảng đo độ chính xác trên tập kiểm tra**

| Mô hình | Số lớp ẩn | Số lượng<br>nơ ron | Số lượng<br>tham số | Số lượng<br>đặc trưng | Điểm đánh giá (Tập kiểm tra) |             |             |
|---------|-----------|--------------------|---------------------|-----------------------|------------------------------|-------------|-------------|
|         |           |                    |                     |                       | MSE                          | RMSE        | MAE         |
| RNN     | 1         | 128                | 16,769              | 1                     | 249.2014303                  | 15.78611511 | 11.14971804 |
|         | 1         | 256                | 67,073              | 4                     | 241.651444                   | 15.54514214 | 11.1533188  |
|         | 1         | 128                | 17,921              | 10                    | 243.0733194                  | 15.59080881 | 11.23058325 |
|         | 2         | (32, 128)          | 21,825              | 1                     | 256.6463031                  | 16.02018424 | 11.25858777 |
|         | 2         | (32, 256)          | 75,425              | 4                     | 243.3260323                  | 15.59891126 | 11.38724381 |
|         | 2         | (32, 32)           | 3,489               | 10                    | 294.0147843                  | 17.14685931 | 12.13193968 |
| LSTM    | 1         | 512                | 1,053,185           | 1                     | 248.6087317                  | 15.76733115 | 11.30907738 |
|         | 1         | 512                | 1,059,329           | 4                     | 247.6136624                  | 15.73574473 | 11.42911978 |
|         | 1         | 512                | 1,071,617           | 10                    | 292.9039599                  | 17.11443718 | 13.36275713 |
|         | 2         | (256, 256)         | 789,761             | 1                     | 288.4670523                  | 16.98431784 | 12.77683638 |
|         | 2         | (128, 128)         | 199,809             | 4                     | 259.7607215                  | 16.11709408 | 11.90671968 |
|         | 2         | (256, 256)         | 798,977             | 10                    | 254.2454237                  | 15.94507522 | 11.56734382 |

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.

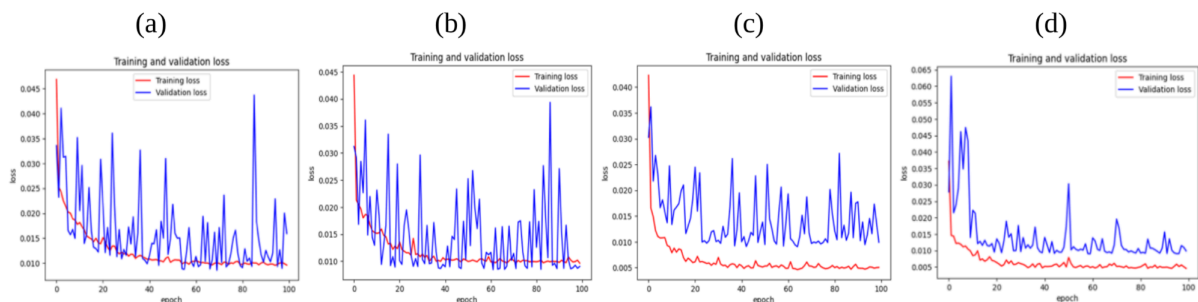
Năm 2024, Trần Đăng Tuyên công bố nghiên cứu của mình về việc sử dụng mô hình phức hợp LSTM-GRU để dự báo chỉ số VNIndex, tuy nhiên có một số điểm khác như tập dữ liệu là từ ngày 14/7/2008 đến 14/7/2023 với số quan sát là 3747 và chọn kích thước bước thời gian là 50 [5]. Dưới đây là kết quả được Trần Đăng Tuyên công bố:

**Bảng 3.6** Bảng đo độ chính xác trên tập huấn luyện [5]

| <i>Thông số</i>  | <i>LSTM(0:18:50)</i> | <i>GRU(0:13:06)</i> | <i>GRU-LSTM(0:14:21)</i> | <i>LSTM-GRU(0:12:28)</i> |
|------------------|----------------------|---------------------|--------------------------|--------------------------|
| 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 [5]

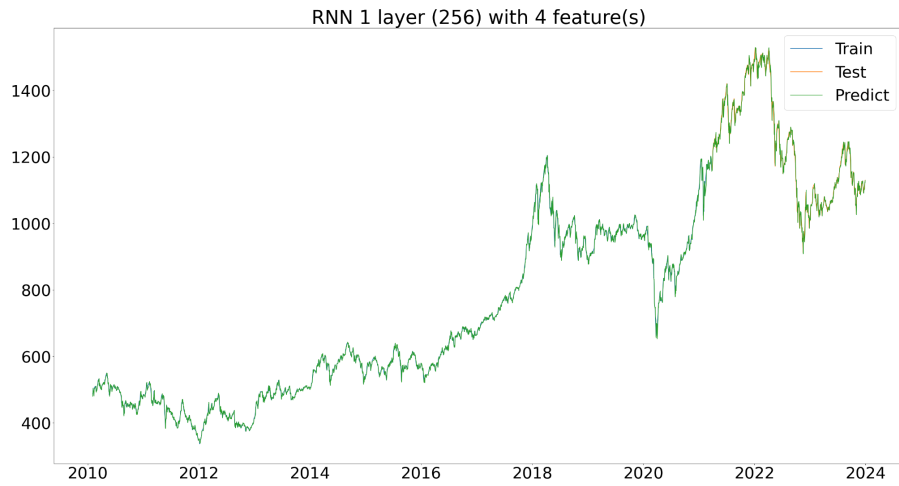
| <i>Thông số</i>  | <i>LSTM</i> | <i>GRU</i> | <i>GRU-LSTM</i> | <i>LSTM-GRU</i> |
|------------------|-------------|------------|-----------------|-----------------|
| r2_score         | 0,98028     | 0,99052    | 0,98949         | 0,99108         |
| ev_score         | 0,99064     | 0,99061    | 0,98960         | 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 [5]

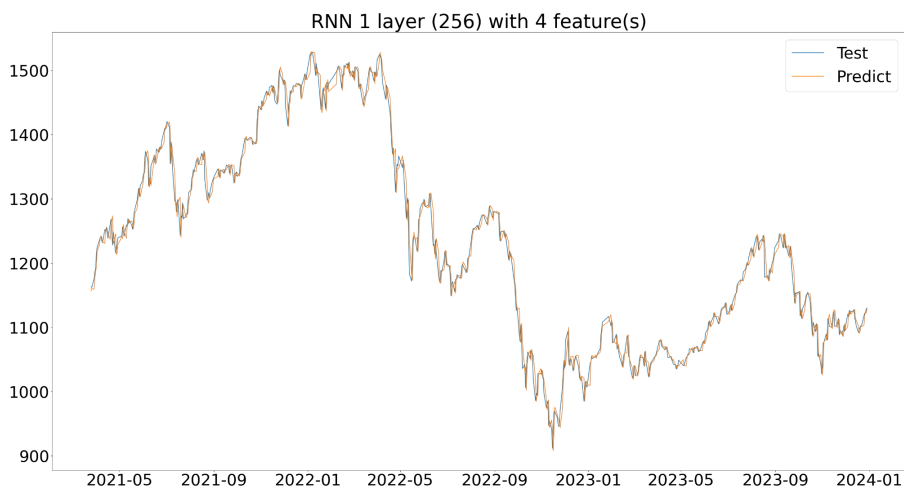
*Ghi chú: (a) LSTM, (b) GRU, (c) GRU-LSTM, (d) LSTM-GRU*

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.

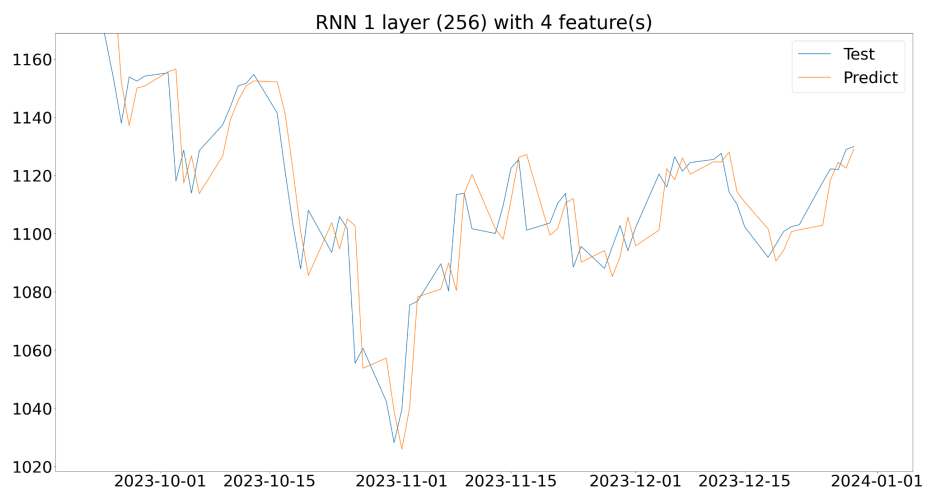
Dưới đây là một số hình ảnh ứng dụng mô hình đề án nghiên cứu vào dự báo chỉ số VNIndex.



**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.



## TÀI LIỆU THAM KHẢO

### Tiếng Việt

- [1] Trương Thị Thùy Dương (2023), “Dự báo chiều biến động của chỉ số chứng khoán bằng thuật toán tăng cường”, *Tạp chí Khoa học và Đào tạo Ngân hàng*, số 252, 40-46.
- [2] Trịnh Viết Giang (2024), “Dự báo giá cổ phiếu trên thị trường chứng khoán thông qua mô hình phức hợp LSTM-GRU hybrid”, *Tạp chí Tài chính*, 834, 59-62.
- [3] Đào Lê Kiều Oanh, Nguyễn Thị Minh Châu (2024), “Dự báo chỉ số chứng khoán bằng học máy: Bằng chứng thực nghiệm từ thị trường chứng khoán Việt Nam”, *Tạp chí Kinh tế và Dự báo*
- [4] Nguyễn Hữu Xuân Trường, Trương Hải Nam, Nguyễn Ngọc Ánh, Nguyễn Sao Mai (2023), “Ứng dụng kỹ thuật học kết hợp trong học máy dự báo giá cổ phiếu”, *Tạp chí ngân hàng*
- [5] Trần Đăng Tuyên (2024), “Đánh giá hiệu suất mô hình phức hợp LSTM-GRU: nghiên cứu điển hình về dự báo chỉ số đo lường xu hướng biến động giá cổ phiếu trên sàn giao dịch chứng khoán Hồ Chí Minh”, *Tạp chí Khoa học Đại học Cần Thơ*, 60(1), 235-249.

### Tiếng Anh

- [6] A. Chatterjee, H. Bhowmick and J. Sen (2021), “Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models”, *IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India*, 2021, pp. 289-296, DOI: 10.1109/MysuruCon52639.2021.9641610.
- [7] Anwar Ul Haq, Adnan Zeb, Zhenfeng Lei, Defu Zhang (2021), “Forecasting daily stock trend using multi-filter feature selection and deep learning”, *Expert Systems with Applications*, vol 168, 114444, ISSN 0957-4174. DOI: 10.1016/j.eswa.2020.114444.

- [8] Chong, Eunsuk & Han, Chulwoo & Park, Frank. (2017), “Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies”, *Expert Systems with Applications*. DOI: 10.1016/j.eswa.2017.04.030.
- [9] Fama, E. F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work”, *The Journal of Finance*, 25(2), 383–417. DOI: 10.2307/2325486
- [10] Fu K, Zhang Y. (2024), “Incorporating Multi-Source Market Sentiment and Price Data for Stock Price Prediction”, *Mathematics*, 12(10):1572. DOI: 10.3390/math12101572
- [11] Htun, H.H., Biehl, M. & Petkov, N (2023), “Survey of feature selection and extraction techniques for stock market prediction”, *Financ Innov*, 9, 26. DOI: 10.1186/s40854-022-00441-7
- [12] Hua, Yuxiu & Zhifeng, Zhao & Li, Rongpeng & Chen, Xianfu & Liu, Zhiming & Zhang, Honggang. (2018), “Deep Learning with Long Short-Term Memory for Time Series Prediction”. arXiv:1810.10161
- [13] Hochreiter, Sepp and Schmidhuber, Jürgen (1997), “Long Short-term Memory.” *Neural Computation*, vol. 9, no. 8, pp. 1735-80. DOI: 10.1162/neco.1997.9.8.1735.
- [14] Huiyu Li and Junhua Hu (2024), “A Hybrid Deep Learning Framework for Stock Price Prediction Considering the Investor Sentiment of Online Forum Enhanced by Popularity”, arXiv:2405.10584 [cs.LG].
- [15] Ince, H., & Trafalis, T. B. (2008), “Short term forecasting with support vector machines and application to stock price prediction.” *International Journal of General Systems*, 37(6), 677–687. DOI: 10.1080/03081070601068595
- [16] Kandadi, Thirupathi and Shankarlingam, G. (2025), “DRAWBACKS OF LSTM ALGORITHM: A CASE STUDY”. DOI: 10.2139/ssrn.5080605

- [17] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar (2020), “Stock Closing Price Prediction using Machine Learning Techniques”, *Procedia Computer Science*, 167, 599–606. DOI: 10.1016/j.procs.2020.03.326.
- [18] Mishra, Ranjan Kumar, Reddy, G. Y. Sandesh and Pathak, Himanshu (2021), “The Understanding of Deep Learning: A Comprehensive Review”, *Mathematical Problems in Engineering*, vol. 2021. DOI:10.1155/2021/5548884
- [19] Phuoc, T., Anh, P.T.K., Tam, P.H. et al. (2024) “Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam”, *Humanit Soc Sci Commun*, 11, 393. DOI: 10.1057/s41599-024-02807-x
- [20] Robin M. Schmidt (2019), “Recurrent Neural Networks (RNNs): A gentle Introduction and Overview”, arXiv:1912.05911 [cs.LG].
- [21] Shen, J., Shafiq, M.O. (2020), “Short-term stock market price trend prediction using a comprehensive deep learning system”, *Journal of Big Data*, 1(7), 66. DOI: 10.1186/s40537-020-00333-6
- [22] X. Wen and W. Li (2023), "Time Series Prediction Based on LSTM-Attention-LSTM Model," *IEEE Access*, vol. 11, 48322-48331. DOI: 10.1109/ACCESS.2023.3276628.

#### **Tham khảo từ Internet**

- [23] <https://www.pluralsight.com/resources/blog/guides/introduction-to-lstm-units-in-rnn>, truy cập ngày 20/12/2023
- [24] <https://www.investopedia.com/terms/s/sma.asp>, truy cập ngày 22/12/2023
- [25] <https://www.steema.com/docs/financialFunctionsRef/expMovingAverageFunction.htm>, truy cập ngày 22/12/2023
- [26] <https://www.strike.money/technical-analysis/bollinger-bands>, truy cập ngày 22/12/2023

- [27] <https://www.investopedia.com/terms/r/rsi.asp>, truy cập ngày 22/12/2023
- [28] <https://www.investopedia.com/terms/o/onbalancevolume.asp>, truy cập ngày 22/12/2023
- [29] <https://towardsdatascience.com/back-to-basics-part-uno-linear-regression-cost-function-and-gradient-descent-590dcb3eee46>, truy cập ngày 24/12/2023
- [30] <https://doanhnhansaigon.vn/thi-truong-chung-khoan-phong-vu-bieu-cua-nen-kinh-te-227399.html>, truy cập ngày 27/12/2023
- [31] <https://www.investopedia.com/terms/f/fundamentalanalysis.asp>, truy cập ngày 27/12/2023
- [32] <https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm>, truy cập ngày 10/4/2025
- [33] <https://www.kaggle.com/code/riddhimanghatak/kriti-final>, truy cập ngày 10/4/2025
- [34] <https://www.kaggle.com/code/mohammedobeidat/us-stocks-price-prediction-using-fundamentals>, truy cập ngày 11/4/2025
- [35] <https://www.kaggle.com/code/ysthehurricane/tesla-stock-price-prediction-using-gru-tutorial>, truy cập ngày 11/4/2025