Working Paper 2024.3.3.12
- Vol. 3, No. 3

ỨNG DỤNG MÔ HÌNH NƠ RON NHÂN TẠO LSTM DỰ ĐOÁN GIÁ CỔ PHIẾU VIỆT NAM

Hoàng Nữ Thanh Tuyền[1]

Sinh viên K61 Kế toán Kiểm toán

Trường Đại học Ngoại thương Cơ sở II tại TP. Hồ Chí Minh

Đoàn Nguyễn Bảo An

Sinh viên K61 Kinh tế Đối ngoại

Trường Đại học Ngoại thương Cơ sở II tại TP. Hồ Chí Minh

Vương Thị Thảo Bình

Giảng viên Bộ môn Khoa học cơ bản

Trường Đại học Ngoại thương Cơ sở II tại TP. Hồ Chí Minh

Tóm tắt

Việc sử dụng học sâu để dự đoán giá và xu hướng thị trường chứng khoán đang ngày càng trở nên phổ biến và đóng vai trò quan trọng trong kỷ nguyên dữ liệu lớn ngày nay. Dự đoán giá cổ phiếu có thể cung cấp cho các nhà đầu tư quan tâm đến thị trường chứng khoán những thông tin có giá trị để có thể đưa ra quyết định mang lại lợi nhuận và hạn chế rủi ro cho nhà đầu tư. Do đó, dự đoán giá cổ phiếu đòi hỏi sự hiểu biết sâu sắc về thị trường, khả năng phân tích dữ liệu và sự linh hoạt để thích ứng với những biến động không lường trước được. Các nhà đầu tư và nhà phân tích thị trường phải kết hợp nhiều phương pháp và công cụ khác nhau, từ phân tích các chỉ số tài chính và phân tích kỹ thuật cơ bản đến sử dụng các mô hình phức tạp như mô hình học máy để giúp cải thiện khả năng dự báo. Nghiên cứu này áp dụng các mô hình Feature Expansion (FE), Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) và Long-Short Memory (LSTM RNN) để dự đoán giá cổ phiếu. Kiểm thử trong thời gian ngắn cho kết quả với độ chính xác tương đối cao lần lượt là : 83.82%, 79.75%, 83.44%, 82,32%, 81,12%. Ngoài ra, phương pháp còn tập trung vào xử lý dữ liệu từ dữ liệu thứ cấp, tiền xử lý để tính toán các chỉ số giá tài chính phục vụ cho việc đưa vào đào tạo mô hình. Mô hình mang

[1] Tác giả liên hệ, Email: k61.2211825025@ftu.edu.vn

lại hiệu quả và ý nghĩa cho các nhà đầu tư và quản trị viên, đóng góp vào cộng đồng nghiên cứu và phân tích chứng khoán trong lĩnh vực tài chính.

Từ khóa: Dự đoán giá cổ phiếu; Mạng nơ-ron; Mở rộng tính năng; Loại bỏ tính năng đệ quy; Phân tích thành phần chính;

APPLYING MACHINE LEARNING METHOD TO PREDICT VIETNAM STOCK PRICES

Abstract

The use of deep learning to predict stock market prices and trends is becoming increasingly popular and plays an important role in today's big data era. Stock price prediction can provide investors interested in the stock market with valuable information to be able to make decisions that bring profits and limit risks for investors. Therefore, predicting stock prices requires a deep understanding of the market, the ability to analyze data and the flexibility to adapt to unforeseen fluctuations. Investors and market analysts must combine a variety of methods and tools, from analyzing financial indicators and basic technical analysis to using complex models such as learning models. machine to help improve forecasting capabilities. This study applies Feature Expansion (FE), Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and Long-Short Memory (LSTM RNN) models to predict stock prices. Short-term testing gives results with relatively high accuracy in training sessions: 83.82%, 79.75%, 83.44%, 82,32%, 81,12%. In addition, the method also focuses on processing data from secondary data, preprocessing to calculate financial price indexes for inclusion in model training. The model brings efficiency and meaning to investors and administrators, contributing to the research and analysis community of securities in the financial sector.

Keywords: Stock price prediction; Neural network; Feature expansion; Recursive feature elimination; Principal component analysis;

1. Giới thiệu

Hiện nay, thị trường Chứng khoán được đánh giá đang dần trở thành một kênh huy động và phân bổ vốn quan trọng cho phát triển kinh tế. Đây không chỉ là nơi để các nhà đầu tư tham gia, mà còn là bộ chỉ số của hoạt động kinh tế toàn cầu. Sự biến động của giá cổ phiếu không chỉ thể hiện tình hình cụ thể của doanh nghiệp mà còn phản ánh xu hướng tổng thể của thị trường. Ở Việt Nam, ngành Chứng khoán trải qua hơn 23 năm hình thành, phát triển với những thành tựu nhất định cùng với những chuyển mình ngày càng lớn mạnh của nền kinh tế đất nước. Trong một thị trường có số lượng lớn các nhà đầu tư cá nhân trẻ tuổi, việc dự đoán giá cổ phiếu trong ngắn hạn là vô cùng cần thiết.

Thị trường chứng khoán thường được mô tả như một hệ thống hỗn loạn phi tham số, phi tuyến tính và xác định, với sự xuất hiện của các yếu tố nhiễu (Ahangar và cộng sự, 2010).Do đó, các nhà đầu tư và nhà phân tích thị trường phải kết hợp nhiều phương pháp và công cụ khác nhau, từ phân tích các chỉ số tài chính, phân tích kỹ thuật cơ bản đến sử dụng các mô hình dự đoán phức tạp như học máy để giúp cải thiện khả năng dự báo và giảm rủi ro trong đầu tư chứng khoán. Việc áp dụng học máy vào lĩnh vực tài chính không chỉ giúp dự báo thị trường một cách chính xác hơn mà còn tạo ra cơ hội đầu tư lợi nhuận cao. Điều quan trọng không chỉ là việc dự

đoán, mà còn là cách thức thực hiện chiến lược giao dịch một cách thông minh và hiệu quả. Với sự phổ biến và giá trị cao, lĩnh vực dự đoán thị trường cổ phiếu không ngừng phát triển và trở thành nguồn kiến thức quý giá cho nhà đầu tư và các chuyên gia tài chính. Không thể phủ nhận rằng việc hiểu biết sâu sắc về thị trường cùng khả năng phân tích dữ liệu đóng vai trò quan trọng giúp đưa ra dự báo chính xác về giá cổ phiếu, Chủ đề này đã được nhiều nhà nghiên cứu trên thế giới quan tâm và đưa ra nhiều giải pháp. Mỗi giải pháp đều có ưu, nhược điểm khác nhau, tuy nhiên sử dụng học máy là giải pháp mang lại kết quả tốt và được nhiều nhà đầu tư tin tưởng. Học máy hay còn gọi là Machine Learning là một lĩnh vực của trí tuệ nhân tạo mà máy tính được lập trình để tự học và cải thiện dữ liệu mà nó thu thập được. Chính vì lý do này làm cho việc sử dụng máy học để dự đoán giá cổ phiếu trở nên thực tế và có ý nghĩa cao.

2. Cơ sở lý thuyết

2.1. Dự đoán giá cổ phiếu bằng học máy

Trong vài năm qua, phương pháp học máy đã được ứng dụng trong nhiều lĩnh vực tài chính và kinh tế. Đó cũng là nền tảng để nhiều nhà nghiên cứu sử dụng các thuật toán học máy để tạo ra các công cụ phân tích và dự đoán từ những dữ liệu tài chính trong quá khứ và các thông tin liên quan khác để hỗ trợ việc ra quyết định trong đầu tư. Nghiên cứu của Mehtabhorn Obthong (2020) khẳng định, thành công lớn trong ML trong vài năm qua đã thay đổi cách các nhà đầu tư sử dụng thông tin và cung cấp cơ hội phân tích tối ưu cho tất cả các loại hình đầu tư. Vì vậy, ML là một công cụ quan trọng để giúp đầu tư tài chính. Việc sử dụng dữ liệu tài chính lịch sử hoặc chuỗi thời gian, lựa chọn cẩn thận các mô hình, dữ liệu và tính năng phù hợp đều rất cần thiết để tạo ra kết quả chính xác. Kết quả chính xác phụ thuộc vào cơ sở hạ tầng hiệu quả, việc thu thập thông tin liên quan và thuật toán được sử dụng (Alpaydin, 2014).

Bảng 1. Các kỹ thuật học máy được sử dụng để phân tích và dự báo các loại công cụ tài chính

Phương pháp	**Loại công cụ tài chính**				
	Cổ phiếu	Trái phiếu	Công cụ phái sinh	Ngoại hối	Hàng hóa
RF	√	√			√
SVM	√	√			
MLP	√		√	√	√
LSTM	√				
RNN	√	√	√	√	√
GAs	√		√		
KNN	√	√	√	√	
SVR	√	√	√	√	
MCS	√	√	√	√	

Phương pháp	Loại công cụ tài chính				
	Cổ phiếu	Trái phiếu	Công cụ phái sinh	Ngoại hối	Hàng hóa
ANNs	√	√	√		
CART	√	√			
GP	√		√		
BSM	√		√		
GRNN	√				√
RBF			√		
BPNN	√	√	√		
LR	√		√		
HMM	√	√	√		

Nguồn: Mehtabhorn Obthong, 2020

2.2. Tổng quan tình hình nghiên cứu

Trong những năm gần đây, chủ đề sử dụng các phương pháp học máy để dự đoán giá cổ phiếu đã trở nên phổ biến. Thị trường chứng khoán được đặc trưng bởi cả sự không chắc chắn và tính biến động, khiến việc dự đoán chính xác xu hướng thị trường trở nên khó khăn. Có nhiều phương pháp học máy đã được sử dụng để dự báo giá cổ phiếu như hồi quy tuyến tính, máy vector hỗ trợ hay mạng nơ ron nhân tạo,… Giá cổ phiếu được dự đoán bằng thuật toán hồi quy tuyến tính trong học máy và triển khai dữ liệu bằng các công cụ và thư viện python như Scikit-learn, Numpy (Sonali Antad và cộng sự, 2023).

Lee và cộng sự (2009) đã sử dụng máy vectơ hỗ trợ (SVM) cùng với phương pháp lựa chọn đặc trưng lai để thực hiện dự đoán xu hướng chứng khoán. Tập dữ liệu trong nghiên cứu này là tập dữ liệu con của Chỉ số NASDAQ trong Cơ sở dữ liệu Tạp chí Kinh tế Đài Loan (TEJD) năm 2008. Cùng năm đó, tác giả (Tsai, C.-F. and Wang, S.-P, 2009) đã chọn các chỉ số cơ bản, chỉ số kỹ thuật, chỉ số kinh tế vĩ mô và tập dữ liệu được lấy từ cơ sở dữ liệu TEJ để dự báo giá cổ phiếu trong ngành điện tử ở Đài Loan. Mạng thần kinh nhân tạo (ANN) có thể mang lại hiệu quả tương đối tốt trong việc dự báo giá cổ phiếu nhưng nó không thể giải thích rõ ràng các quy tắc dự báo. Mặt khác, mô hình cây quyết định (DT) có thể tạo ra một số quy tắc để mô tả các quyết định dự báo. Do đó, tác giả tập trung vào việc kết hợp ANN và cây quyết định để tạo ra mô hình dự báo giá cổ phiếu. Kết quả thực nghiệm cho thấy mô hình DT+ANN kết hợp có độ chính xác 77%, cao hơn so với mô hình ANN và DT đơn lẻ trong ngành điện tử.

Amin Hedayati Moghaddama và cộng sự (2016) đã nghiên cứu khả năng sự báo tỷ giá hối đoái chứng khoán NASDAQ hằng ngày bằng Mạng thần kinh nhân tạo (ANN). Tỷ giá hối đoái hàng ngày của NASDAQ từ ngày 28 tháng 1 năm 2015 đến ngày 18 tháng 6 năm 2015 được sử dụng để phát triển một mô hình mạnh mẽ. 70 ngày đầu tiên (28 tháng 1 đến 7 tháng 3) được

chọn làm tập dữ liệu huấn luyện và 29 ngày cuối cùng được sử dụng để kiểm tra khả năng dự đoán của mô hình.

Bên cạnh đó vào năm 2019, Dharmaraja Selvamuthu và cộng sự đánh giá các kỹ thuật phổ biến nhất được sử dụng trong dự báo chuỗi thời gian tài chính là Máy vectơ hỗ trợ (SVM), Hồi quy vectơ hỗ trợ (SVR) và Mạng thần kinh lan truyền ngược (BPNN). Cụ thể, tác giả sử dụng mạng thần kinh dựa trên ba thuật toán học khác nhau, tức là Levenberg-Marquardt, Độ dốc liên hợp theo tỷ lệ và Chính quy Bayes để dự đoán thị trường chứng khoán dựa trên dữ liệu đánh dấu cũng như dữ liệu 15 phút của một công ty Ấn Độ và kết quả của chúng được so sánh. Cả ba thuật toán đều cung cấp độ chính xác 99,9% khi sử dụng dữ liệu đánh dấu. Độ chính xác trên tập dữ liệu 15 phút giảm xuống lần lượt là 96,2%, 97,0% và 98,9% .

Vaishnavi Gururaj và cộng sự (2019) về phương pháp đánh giá các mô hình, chỉ cần chứng minh rằng mô hình dự đoán phù hợp với dữ liệu nhất có thể là đủ để đánh giá các kỹ thuật học máy. Thay vì đưa ra dự đoán thực tế, tập dữ liệu thử nghiệm được cung cấp để so sánh mức độ phù hợp với kết quả dự đoán của phương pháp học máy từ dữ liệu sử dụng để huấn luyện.

Nghiên cứu của Yixin Guo (2020) đã sử dụng chỉ số S&P500 thông qua yahoo Finance, dữ liệu giao dịch cho mỗi ngày từ 26/09/2001 đến 24/09/2021 có tất cả 5000 quan sát. Lựa chọn mô hình chuỗi thời gian truyền thống và mô hình mạng nơ-ron LSTM để xây dựng mô hình giá cổ phiếu và đưa ra dự đoán. Mô hình LSTM được tận dụng vào dự báo chuỗi thời gian tài chính truyền thống và mô hình dự báo chứng khoán dựa trên mạng thần kinh bộ nhớ ngắn hạn (LSTM) được thiết lập. Sai số tuyệt đối và hệ số xác định đã được đánh giá và thu được hiệu quả dự đoán tốt hơn. Chứng minh tính khả thi của của các mô hình Nơ ron nhân tạo trong dự báo chuỗi thời gian tài chính, có thể hướng dẫn hành vi đầu tư của các tổ chức và cá nhân ở một mức độ nhất định và cung cấp những ý tưởng mới cho nghiên cứu dự báo chứng khoán.

Ở Việt Nam, việc nghiên cứu về dự đoán giá cổ phiếu bằng phương pháp học máy cũng đã thu hút được sự quan tâm của nhiều nhà nghiên cứu, nhà đầu tư. Họ đều hi vọng rằng thông qua việc sử dụng các mô hình học máy này, họ có thể giúp cho những người quan tâm có thể đưa ra những quyết định đầu tư thông minh hơn và mang lại hiệu quả tốt hơn trên thị trường tài chính.

Tiêu biểu có các nghiên cứu như đề tài của Phạm Hữu Lê Quốc Phục (2010). Trong nghiên cứu này, tác giả đã ứng dụng mạng Nơ-ron sâu (Deep Neural Networks) để giải quyết bài toán dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam. Mô hình này đề xuất các mô hình mạng Nơ-ron sâu để dự đoán xu hướng giá cổ phiếu thông qua dữ liệu lịch sử giá cổ phiếu và các chỉ số tài chính.

Hay gần đây nhất, nhóm sinh viên Trần Mẫn Quân, Đặng Nguyễn Phước An,Nguyễn Minh Nhựt Đại học Quốc gia Thành phố Hồ Chí Minh tác giả của bài Community Detection for Personalized Learning Pathway Recommendations on IT E-Learning System được đăng tại Hội nghị Quốc tế FDSE 2023. Bằng cách sử dụng các dữ liệu kỹ thuật và dữ liệu cơ bản, nghiên cứu đã đề xuất các mô hình dự đoán hiệu quả cho cho các cổ phiếu khác nhau nhờ sử dụng kết hợp các kỹ thuật học máy như vector hỗ trợ (SVM) và mạng nơ-ron nhân tạo (ANN).

Nghiên cứu của Bùi Thanh Khoa và cộng sự (2022) cũng là một trong những đề tài tiêu biểu về sử dụng phương pháp học máy để dự dự đoán giá cổ phiếu. Nghiên cứu đã áp dụng mô

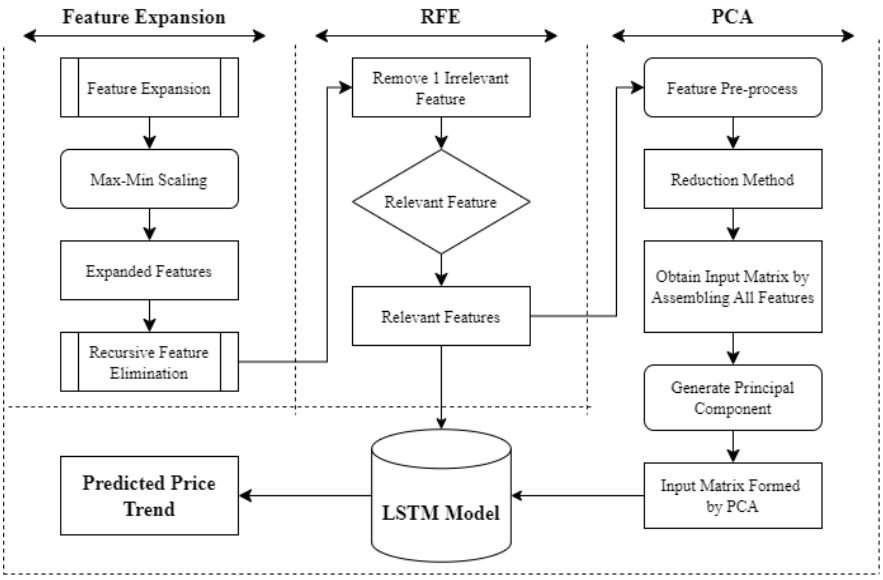
hình Support Vector Regression (SVR) dựa trên tiền đề là mô hình Capital Asset Pricing Model (CAPM) để dự báo tỷ suất sinh lợi của từng cổ phiếu và xác định các nhân tố ảnh hưởng đến sai số trong dự báo. Bằng việc sử dụng lượng dữ liệu thu thập được từ các công ty niêm yết trên thị trường chứng khoán TP. Hồ Chí Minh từ 12/2012 đến tháng 9/2020, nghiên cứu cũng đã chỉ ra rằng mô hình SVR hiệu quả hơn so với CAPM. Ngoài ra, một số nghiên cứu khác áp dụng các mô hình học máy như học sâu Deep Learning hay kết hợp giữa hai phương pháp học máy như học sâu và học máy tập trung để dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam.

Nhìn chung, các phương pháp học máy đã đem lại một cách tiếp cận mới cho việc dự đoán giá cổ phiếu trong thị trường chứng khoán nước ngoài. Sự kết hợp giữa lý thuyết học máy cùng với các dữ liệu lịch sử và chỉ số tài chính là một cách tiếp cận mới, có sự toàn diện hơn so với các cách tiếp cận độc lập riêng học máy hay độc lập về khía cạnh tài chính. Phương pháp này đã làm tư tăng độ chính xác hay nâng cao hiệu suất của việc dự đoán giá cổ phiếu trong ngắn hạn. Các nghiên cứu mà nhóm tác giả liệt kê cũng như tiếp cận ở trên là minh chứng rõ ràng nhất cho việc áp dụng các thuật toán như hồi quy tuyến tính, mạng nơ-ron nhân tạo,...hoàn toàn có thể mang lại kết quả ấn tượng và tương đối chính xác trọng việc dự báo xu hướng giá cổ phiếu.

3. Phương pháp nghiên cứu

3.1. Quy trình nghiên cứu

Trong nghiên cứu này, dữ liệu được nhóm tác giả thu thập từ thư viện Vnstock trong Python. Sau đó, tiến hành gán nhãn dữ liệu thô, đồng thời tính các chỉ số dựa trên giá đóng cửa trong bộ dữ liệu. Sau khi xử lý dữ liệu thô, bộ dữ liệu sẽ được tiến hành tiền xử lý. Quy trình tiền xử lý dữ liệu là quy trình kết hợp ba kỹ thuật: Kỹ thuật mở rộng tính năng; Kỹ thuật loại bỏ tính năng đệ quy; Kỹ thuật phân tích thành phần chính. Nhóm tác giả đã sử dụng các thư viện matplotlib, pandas, numpy trong Python trong bước này. Tiếp theo, bộ dữ liệu được đưa vào mô hình mạng Nơ ron nhân tạo để huấn luyện và dự đoán kết quả. Cuối cùng, để đánh giá mô hình, sử dụng thư viện Sklearn và Mathplotlib để tính toán các chỉ số độ đo như độ chính xác, độ phân loại và độ lỗi.



Hình 1. Sơ đồ thuật toán quy trình nghiên cứu

Nguồn: Nhóm tác giả tự tổng hợp

3.1. Thu thập dữ liệu

Dữ liệu được sử dụng là bộ dữ liệu giá cổ phiếu của Dữ liệu bao gồm 1 file excel là thông tin giao dịch từng ngày của 5 công ty: Ngân hàng Thương mại Cổ phần Á Châu (ACB); Công ty Cổ phần Văn hóa Phương Nam (PNC); Công ty Cổ phần Xuất nhập khẩu Thủy sản Bến Tre (ABT); Công ty Cổ phần FPT (FPT) và Công ty Cổ phần Cao su Hòa Bình (HRC). Bộ dữ liệu gồm có 21085 mẫu giá cổ phiếu theo ngày giao dịch của 5 công ty từ ngày 02/04/2007 đến 01/03/2024. Trước tiên, dữ liệu sẽ được gán nhãn 0 hoặc 1, gán nhãn 0 đối với các ngày giao dịch có giá đóng cửa giữ nguyên hoặc giảm so với ngày đóng cửa trước đó và gán 1 đối với các ngày giao dịch có giá đóng cửa tăng so với ngày trước đó. Sau đó, nhóm tác giả tiến hành tính giá trần, giá sàn của các ngân hàng theo quy định của sàn niêm yết. Và cuối cùng, sử dụng các trường dữ liệu đó để tính các chỉ số SMA10, BIAS20, CCI24, MTM10, ROC10, RSI14, WNR9, SlowK3, SlowD3. Dữ liệu đầu ra của quá trình này là file excel bao gồm 20 trường dữ liệu bao gồm 8 trường dữ liệu ban đầu, 1 trường là nhãn được gán, 2 trường là giá trần, giá sàn và 9 chỉ số. Trong các bước tiếp theo, các trường dữ liệu này sẽ được gọi bằng thuật ngữ "tính năng".

```
df =  stock_historical_data(symbol='ACB',
                            start_date="2007-04-01",
                            end_date='2024-03-01',
                            resolution='1',
                            type='stock',
                            beautify=True,
                            decor=False,
                            source='DNSE')
```

Hình 2. Đoạn code lấy dữ liệu giao dịch của Ngân hàng ACB

Nguồn: Nhóm tác giả tự tổng hợp

3.2. Kỹ thuật tiền xử lý dữ liệu

3.2.1. Áp dụng kỹ thuật mở rộng tính năng

Trong quá trình này, nhóm tác giả sử dụng 2 phương pháp: Chia tỷ lệ tối đa - tối thiểu và tính phần trăm dao động để mở rộng từ bộ dữ liệu bao gồm 20 tính năng sau khi trải qua bước xử lý dữ liệu thô ở mục 3.2.3 trở thành bộ dữ liệu mới bao gồm 33 tính năng sau khi kết thúc quá trình. Không phải tất cả các tính năng đều áp dụng được cả 2 phương pháp nêu trên, do đó nhóm tác giả chỉ sử dụng các phương pháp đối với các tính năng nhất định. Chi tiết được mô tả ở bảng 2.

Bảng 2. Lựa chọn phương pháp mở rộng tính năng đối với các tính năng

Tính năng	Chia tỷ lệ tối đa - tối thiểu	Phần trăm dao động
Volume	√	
Amount	√	
SMA10	√	√
CCI24		
MTM10		√
ROC10		√
RSI14	√	√
WNR9	√	
SlowK3	√	√
SlowD3	√	√
BIAS20		

Nguồn: Nhóm tác giả tự tổng hợp

(1) Áp dụng phương pháp Chia tỷ lệ tối đa tối thiểu:

Để áp dụng phương pháp này, nhóm tác giả thực hiện đã sử dụng thư viện Sklearn với MaxMinScaler có sẵn trong ngôn ngữ lập trình Python, bộ scaler MinMaxScaler sẽ đưa các biến về miền giá trị [0, 1]. Các bước như sau:

- Tập dữ liệu huấn luyện được sử dụng để thích hợp biến scaler. Dữ liệu huấn luyện cần được xác định giá trị tối đa và tối thiểu để chuẩn hóa, sử dụng hàm fit() để thực hiện điều này.

- Gọi hàm transform() để tăng kích thước dữ liệu.

- Áp dụng lại bộ scaler để sử dụng cho việc dự đoán về sau.

```
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['Volume'] = scaler.fit_transform(df_2['Volume'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['Amount'] = scaler.fit_transform(df_2['Amount'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['SMA10'] = scaler.fit_transform(df_2['SMA10'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['RSI14'] = scaler.fit_transform(df_2['RSI14'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['WNR9'] = scaler.fit_transform(df_2['WNR9'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['SlowK3'] = scaler.fit_transform(df_2['SlowK3'].values.reshape(-1, 1))
scaler = MinMaxScaler(feature_range=(0, 1))
df_2['SlowD3'] = scaler.fit_transform(df_2['SlowD3'].values.reshape(-1, 1))
```

Hình 3. Đoạn code áp dụng phương pháp chia tỷ lệ tối đa – tối thiểu đối với 7 tính năng được chọn

Nguồn: Nhóm tác giả tự tổng hợp

(2) Tính phần trăm dao động

Đây là một phép toán đơn giản, do đó nhóm tác giả đã sử dụng excel để tính toán đối với các tính năng được chọn để áp dụng phương pháp này.

3.2.2. Áp dụng kỹ thuật loại bỏ tính năng đệ quy

Nhóm tác giả sử dụng thư viện RFE từ sklearn.feature nhằm giảm chiều dữ liệu (dimensionality reduction) bằng cách loại bỏ đặc trưng không quan trọng hoặc không cần thiết, giúp cải thiện hiệu suất mô hình và giảm thời gian huấn luyện. Các bước như sau:

- Sử dụng một Gradient Boosting Classifier từ sklearn.ensemble để huấn luyện mô hình trên toàn bộ tập dữ liệu huấn luyện và đánh giá hiệu suất của mô hình sử dụng F1-score.

- Dùng Recursive Feature Elimination (RFE) để lựa chọn số lượng đặc trưng quan trọng nhất.

- Trực quan hóa kết quả: Kết quả F1-score được trực quan hóa bằng biểu đồ cột, với trục x biểu diễn số lượng đặc trưng được chọn và trục y biểu diễn F1-score tương ứng.

```
fig, ax = plt.subplots()

x = np.arange(1, 14)
y = f1_score_list

ax.bar(x, y, width=0.2)
ax.set_xlabel('Number of features selected using mutual information')
ax.set_ylabel('F1-Score (weighted)')
ax.set_ylim(0, 1.2)
ax.set_xticks(np.arange(1, 14))
ax.set_xticklabels(np.arange(1, 14), fontsize=12)

for i, v in enumerate(y):
    plt.text(x=i+1, y=v+0.05, s=str(v), ha='center')

plt.tight_layout()
```

Hình 4. Đoạn code tạo biểu đồ trực quan hóa kết quả F1-score.

Nguồn: Nhóm tác giả tự tổng hợp

3.2.3. Áp dụng kỹ thuật phân tích thành phần chính

Bước đầu tiên trước khi tận dụng PCA là tiền xử lý tính năng. Bởi vì một số tính năng sau RFE là dữ liệu phần trăm, trong khi những tính năng khác là số lượng rất lớn, tức là đầu ra từ RFE có các đơn vị khác nhau. Nó sẽ ảnh hưởng đến kết quả tách thành phần chính. Vì vậy, trước khi đưa dữ liệu vào thuật toán PCA, cần phải xử lý trước tính năng. Sau khi thực hiện tiền xử lý đối tượng, bước tiếp theo là cung cấp dữ liệu đã xử lý với các đối tượng **i** đã chọn vào thuật toán PCA để giảm thang đo ma trận đối tượng thành các đối tượng **j**. Bước này nhằm giữ lại càng nhiều tính năng hiệu quả càng tốt và đồng thời loại bỏ độ phức tạp tính toán của việc huấn luyện mô hình. Nghiên cứu này cũng đánh giá sự kết hợp tốt nhất giữa **i** và **j**, có độ chính xác dự đoán tương đối tốt hơn, đồng thời giảm được độ phức tạp của thuật toán. Kết quả của bước PCA là ma trận mới với **j** cột.

```
data_scaled.set_index('id', inplace=True)
pca = PCA().fit(data_scaled)
explained_variance_ratio = pca.explained_variance_ratio_
cumulative_explained_variance = np.cumsum(explained_variance_ratio)
optimal_k = 6
sns.set(rc={'axes.facecolor': '#fcf0dc'}, style='darkgrid')
plt.figure(figsize=(20, 10))
barplot = sns.barplot(x=list(range(1, len(cumulative_explained_variance) + 1)),
                      y=explained_variance_ratio,
                      color='#fcc36d',
                      alpha=0.8)
lineplot, = plt.plot(range(0, len(cumulative_explained_variance)), cumulative_explained_variance,
                     marker='o', linestyle='--', color='#ff6200', linewidth=2)
optimal_k_line = plt.axvline(optimal_k - 1, color='red', linestyle='--', label=f'Optimal k value = {optimal_k}')
plt.xlabel('Number of Components', fontsize=14)
plt.ylabel('Explained Variance', fontsize=14)
plt.title('Cumulative Variance vs. Number of Components', fontsize=18)
plt.xticks(range(0, len(cumulative_explained_variance)))
plt.legend(handles=[barplot.patches[0], lineplot, optimal_k_line],
           labels=['Explained Variance of Each Component', 'Cumulative Explained Variance', f'Optimal k value = {optimal_k}'],
           loc=(0.62, 0.1),
           frameon=True,
           framealpha=1.0,
           edgecolor='#ff6200')
x_offset = -0.3
y_offset = 0.01
for i, (ev_ratio, cum_ev_ratio) in enumerate(zip(explained_variance_ratio, cumulative_explained_variance)):
    plt.text(i, ev_ratio, f"{ev_ratio:.2f}", ha="center", va="bottom", fontsize=10)
    if i > 0:
        plt.text(i + x_offset, cum_ev_ratio + y_offset, f"{cum_ev_ratio:.2f}", ha="center", va="bottom", fontsize=10)

plt.grid(axis='both')
plt.show()
```

Hình 5. Đoạn code tạo ra một biểu đồ thể hiện tỷ lệ phương sai.

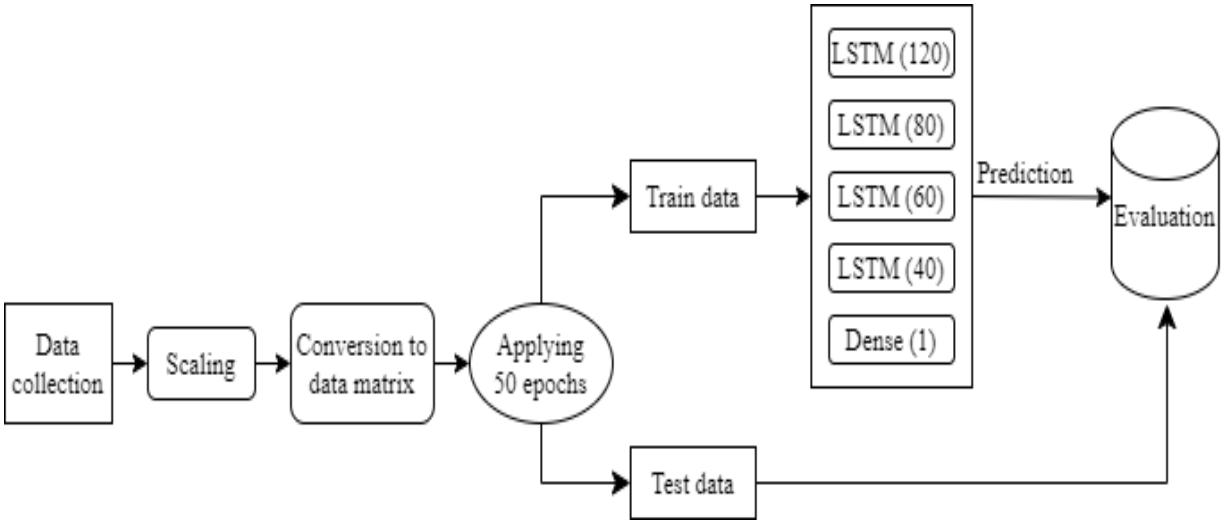
Nguồn: Nhóm tác giả tự tổng hợp.

3.3. Huấn luyện mô hình LSTM dự đoán giá cổ phiếu

Sau khi phân tích thành phần, nhóm tác giả sẽ thu được ma trận rút gọn tỷ lệ. Đầu tiên nhóm tác giả sẽ dùng các lớp Dense, Dropout, LSTM từ thư viện keras.layers và Sequential từ thư viện keras.layers để xây dựng và tinh chỉnh mô hình mạng nơ-ron. Chuỗi dữ liệu được xử lí chia làm đầu vào và dự báo xu hướng.

Đầu tiên cần tiền xử lý dữ liệu trước khi đưa vào mô hình. Lớp đầu vào (Data collection) bao gồm những thuộc tính gốc, thuộc tính mở rộng, và giảm chiều PCA. Lớp thứ hai (Scaling) được yêu cầu xử lý dưới dạng tỉ lệ hóa. Vì LSTM hoạt động dựa trên chuỗi dữ liệu, sử dụng để dự đoán giá trị tiếp theo, do đó cần tạo một ma trận từ tập dữ liệu huấn luyện. Bộ dữ liệu lúc này sẽ được chia làm hai phần, một phần là bộ dữ liệu dùng để huấn luyện mô hình chiếm 80% số mẫu, một phần còn lại được dùng làm bộ kiểm thử. Dữ liệu huấn luyện được đưa vào bao gồm một mảng đa chiều bao gồm các trường hợp của biến phụ thuộc và biến độc lập tương đương. Mô hình đã được thử nghiệm các biến thể của mô hình bao gồm thêm các lớp Dense,Droput. Mô hình LSTM cho nghiên cứu bao gồm 5 lớp, được thể hiện như hình 20, bao gồm 4 lớp LSTM lần lượt với 120, 80, 60, 40 nơ – ron, tiếp theo là lớp Dense với 1 nơ ron.

Cụ thể, mô hình được sử dụng 18100 mẫu để huấn luyện (train data), 2985 mẫu để kiểm thử (test data). Mô hình được huấn luyện 50 lần với độ mất mát trung bình là 0.25.

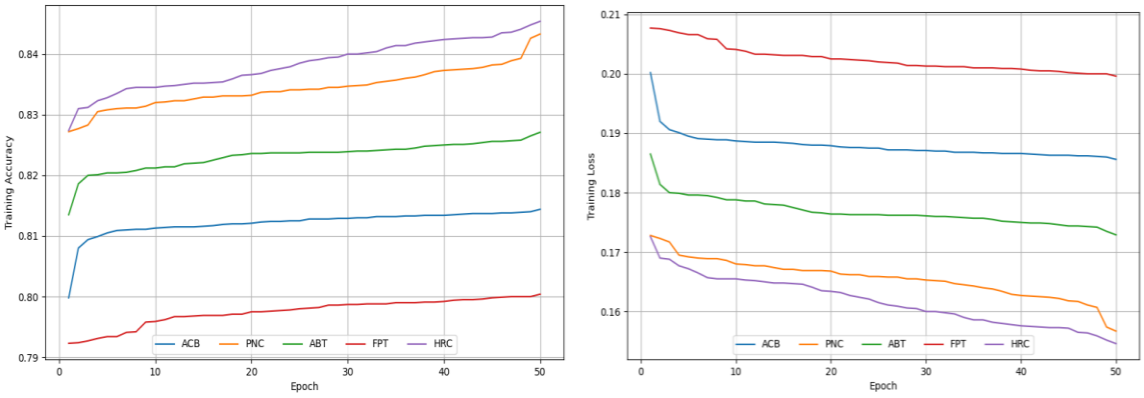


Hình 6. Kiến trúc mô hình LSTM trong việc dự đoán giá

Nguồn: Nhóm tác giả tự tổng hợp

4. Kết quả nghiên cứu huấn luyện mô hình

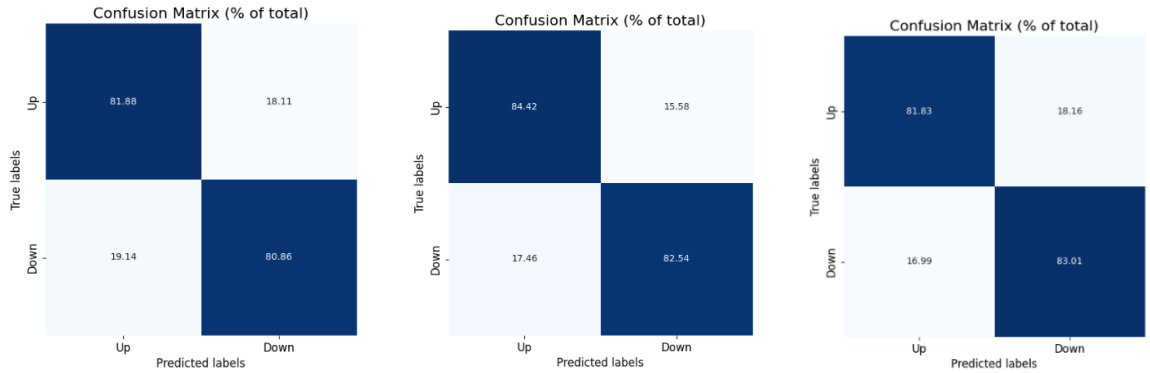
Hình 7 thể hiện xu hướng chính xác và mất mát của mô hình huấn luyện 50 lần đối với 5 công ty. Dựa vào các biểu đồ kết quả, độ chính xác của mô hình với các bộ dữ liệu của các công ty khác nhau đều có xu hướng tăng dần dao động trong khoảng từ 79% đến 85% và độ mất mát giảm dần trong khoảng 21% đến 15%. Kết quả đạt tốt nhất đối với bộ dữ liệu của công ty HRC đạt 84.54% sau 50 lần huấn luyện. Kết quả chứng minh mô hình đang học được các đặc trưng quan trọng trong dữ liệu và có khả năng áp dụng cho các dữ liệu mới.



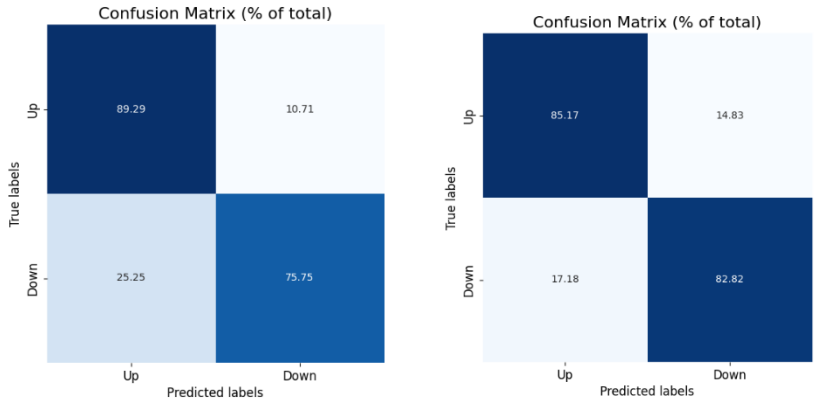
Hình 7: Độ chính xác và độ mất mát của mô hình LSTM đối với các công ty.

Nguồn: Nhóm tác giả tự vẽ bằng Python

So sánh các ma trận sai lầm khi huấn luyện mô hình LSTM đối với bộ dữ liệu của các công ty, rõ ràng nhận thấy rằng mô hình LSTM đặc biệt dự đoán tốt xu hướng tăng của giá cổ phiếu so với xu hướng giảm đặc biệt đối với bộ dữ liệu của FPT. Trong khi mô hình dự đoán đúng xu hướng tăng của giá cổ phiếu đạt tới 89.29% thì mô hình dự đoán đúng xu hướng giảm đạt 75.75%.



ACB PNC ABT



FPT HRC

Hình 9. Ma trận sai lầm mô hình dự đoán giá cổ phiếu các công ty

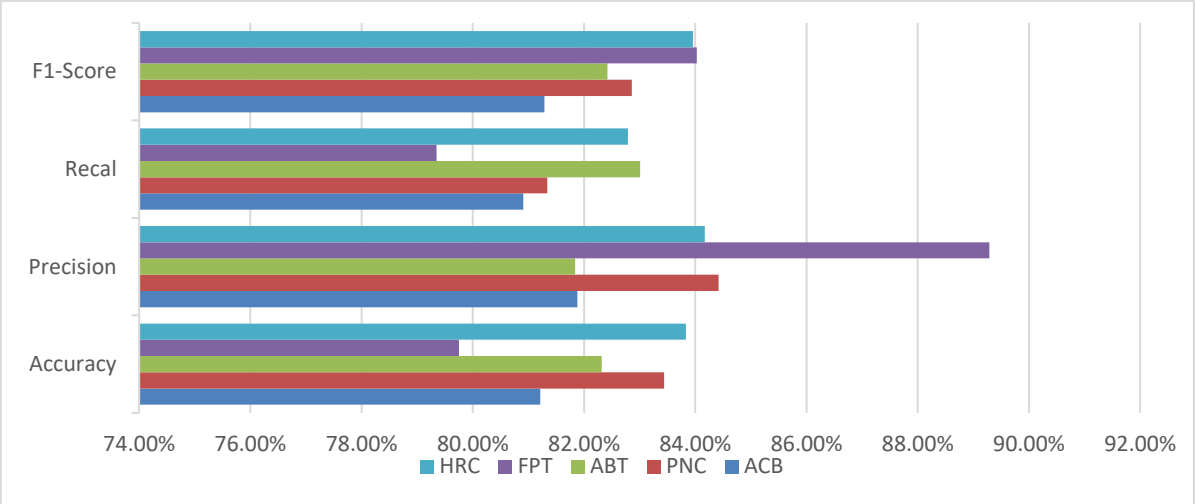
Nguồn: Nhóm tác giả tự vẽ bằng Python

Mô hình LSTM dự đoán giá cổ phiếu có kết quả tương đối tốt, với độ chính xác trên tập kiểm tra khoảng 80%. Độ chính xác này cao nhất khi huấn luyện với bộ dữ liệu của HRC, trong khi độ phủ đạt tốt nhất là 83% đối với bộ dữ liệu của ABT, độ chính xác và F1-score đạt cao nhất 89.29% và 84.03% khi huấn luyện mô hình với bộ dữ liệu của FPT. Nghĩa là mô hình LSTM tương đối phù hợp với các tập dữ liệu huấn luyện.

Bảng 3. Bảng so sánh kết quả của 5 công ty.

Mã cổ phiếu	Testing Accuracy	Testing Precision	Testing recall	Testing F1 Score
ACB	81.21%	81.88%	80.91%	81.29%
PNC	83.44%	84.42%	81.34%	82.86%
ABT	82.32%	81.84%	**83.01%**	82.42%
FPT	79.75%	**89.29%**	79.35%	**84.03%**
HRC	**83.83%**	84.17%	82.79%	83.96%

Nguồn: Nhóm tác giả tự tổng hợp



Hình 10. So sánh kết quả huấn luyện của các công ty

Nguồn: Nhóm tác giả tự tổng hợp

Tổng kết lại, huấn luyện mô hình Nơ ron nhân tạo LSTM để dự đoán giá cổ phiếu cho kết quả tốt đối với các bộ dữ liệu, với độ chính xác dự đoán trên tập kiểm tra, độ phủ và F1-score lên đến 80%. khả năng tổng quát hóa tốt nhất đối với các tập dữ liệu mới.

5. Conclusion

Trong đề tài này, nhóm tác giả đã thực hiện tìm hiểu tổng quan cũng như nghiên cứu về mạng nơ-ron cũng như ứng dụng mô hình mạng LSTM vào bài toán dự đoán giá cổ phiếu trong tương lai. Đề tài tập trung vào việc nghiên cứu và xây dựng mô hình tổng quát cho việc áp dụng mạng nơ-ron cho bài toán dự đoán xu hướng cổ phiếu và phân loại dữ liệu đầu vào.

Nhóm tác giả đã bắt đầu bằng việc thu thập và tiền xử lý dữ liệu để chuẩn bị dữ liệu đầu vào cho quá trình huấn luyện mô hình. Các bước này bao gồm mở rộng tính năng và bỏ dữ liệu

gây nhiễu và tạo ra các tính năng mới. Sau đó nhóm tác giả tiến hành xây dựng mô hình mạng LSTM với kiến trúc phù hợp với bài toán dự đoán giá cổ phiếu. Mô hình được huấn luyện dựa trên dữ liệu lịch sử và sau đó được đánh giá dựa trên dữ liệu kiểm tra để đánh giá hiệu suất của nó. Nhóm tác giả đã tiến hành huấn luyện 5 lần với 5 bộ dữ liệu của 5 công ty khác nhau. Kết quả của 5 lần huấn luyện đã cho thấy mô hình đã đạt được hiệu suất tương đối tốt với độ chính xác trung bình trên 80% trong việc dự đoán xu hướng giá cổ phiếu.

Mục tiêu của nghiên cứu là cung cấp cho những nhà giao dịch một công cụ mới trong việc phân tích và dự đoán giá cổ phiếu. Đây là một nguồn thông tin cho các nhà đầu tư tham khảo từ đó họ có thể đưa ra các quyết định tài chính, hạn chế phần nào rủi ro trong việc đầu tư của họ. Ngoài ra, nghiên cứu còn là một công cụ hỗ trợ đắc lực cho các nhà quản trị doanh nghiệp và các công ty niêm yết. Các nhà quản trị có thể sử dụng thông tin từ nghiên cứu để có thể phân tích và dự đoán biến động giá cổ phiếu của họ để có thể đưa ra các định hướng và các quyết định về số lượng cổ phiếu lưu hành của đơn vị, các quyết định phát hành thêm hay thực hiện các chính sách mua lại nhằm có thể tối ưu hóa lợi ích cho công ty cũng như các cổ đông.

Tóm lại, nghiên cứu này không chỉ đơn thuần là một công cụ phân tích và dự đoán giá cổ phiếu mà còn là một nguồn thông tin chiến lược, giúp các nhà đầu tư và các nhà quản trị doanh nghiệp thấu hiểu sâu hơn về thị trường và đưa ra các quyết định thông minh và hiệu quả.

REFERENCES

Amin, H. M. (2016), "Stock market index prediction using artificial neural network," *Journal of Economics, Finance and Administrative Science*, Vol.21, No.41, pp. 89-93.

Dharmaraja Selvamuthu, Vineet Kumar, Abhishek Mishra (2019), "Indian stock market prediction using artificial neural networks on tick data," *Financial Innovation*, Springer; Southwestern University of Finance and Economics, Vol.5, No.1, pp. 1-12.

Khoa, B. & Huynh, T. (2022), "Forecasting stock price movement direction by machine learning algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, Vol.12, No.6, pp. 6625-6634.

Lee MC, "Using support vector machine with a hybrid feature selection method to the stock trend prediction," *Expert Systems with Applications*, Vol.36, No.8, pp. 10896-10904.

Phục, P.H.L.Q. (2010), "Nghiên cứu ứng dụng mạng Nơ-Ron nhân tạo giải quyết lớp bài toán dự đoán và phân loại," *Luận văn thạc sĩ*, Đại học Đà Nẵng.

Reza Gharoie Ahangar et al. (2010), "The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange," *International Journal of Computer Science and Information Security*, Vol.7, No.2, pp. 1-6.

Sonali Antad et al. (2023), "A New Way for Face Sketch Construction and Detection Using Deep CNN," *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol.11, No.10, pp. 472–480.

S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon & K. P. Soman (2017), "Stock price prediction using LSTM, RNN and CNN-sliding window model," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, pp. 1643-1647.

Thuan, N.D., Quan, T.M., An, D.N.P. & Nhut, N.M. (2023), "Community Detection for Personalized Learning Pathway Recommendations on IT E-Learning System," *Future Data and Security Engineering: Big Data, Security and Privacy, Smart City and Industry 4.0 Applications, Communications in Computer and Information Science*, Vol.1925, pp. 499-512.

Tsai, C.-F. and Wang, S.-P (2019), "Stock Price Forecasting by Hybrid Machine Learning Techniques," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009*, Vol.I, IMECS 2009.

Vaishnavi Gururaj et al. (2019), "Stock market prediction using linear regression and support vector machines," *International Journal of Applied Engineering Research*, Vol.14, No.8, pp. 1931-1934.

Yixin Guo (2022), "Stock Price Prediction Using Machine Learning," *Economics Spring 2022; Södertörn University*.

Zou, J., Zhao, Q., Jiao, Y., Cao, H., Liu, Y., Yan, Q., Abbasnejad, E., Liu, L. & Shi, J. (2022), "Stock Market Prediction via Deep Learning Techniques: A Survey."