

DỰ BÁO THỊ TRƯỜNG CHỨNG KHOÁN VỚI DEEP LEARNING: ĐÁNH GIÁ TỔNG QUAN GIAI ĐOẠN 2020-2024

Nguyễn Thị Hương Giang^{1*}, Lê Mạnh Thanh², Nguyễn Đình Hoa Cường³

¹Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế

²Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế

³Khoa Công nghệ và Kinh doanh, Trường Đại học Phú Xuân Huế

* Email: nthgiang@hueuni.edu.vn

Ngày nhận bài: 6/12/2024; ngày hoàn thành phản biện: 19/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

Bài báo này đánh giá các tài liệu về sử dụng mô hình học sâu trong lĩnh vực dự báo giá cổ phiếu từ năm 2020 đến năm 2024, thảo luận về các mô hình phổ biến sử dụng trong dự báo, các chỉ số đầu vào, các chỉ số đánh giá hiệu suất của mô hình và bộ dữ liệu thường dùng trong mô hình. Bài báo phân tích ưu điểm, nhược điểm của các mô hình và đề xuất một số hướng nghiên cứu tiềm năng trong tương lai.

Từ khoá: dự đoán, dự báo, giá cổ phiếu, mô hình học sâu, thị trường chứng khoán.

1. MỞ ĐẦU

Dữ liệu tài chính nói chung và dữ liệu chứng khoán nói riêng luôn đóng vai trò quan trọng trong đời sống kinh tế xã hội. Đặc biệt, dữ liệu chứng khoán luôn được xem là hàn thử biểu cho hoạt động của các nền kinh tế trên thế giới. Việc nắm bắt xu hướng biến động của dữ liệu chứng khoán đóng vai trò quan trọng không chỉ đối với các hoạt động kinh tế vi mô mà còn có ý nghĩa quan trọng đối với kinh tế vĩ mô.

Dự báo thị trường chứng khoán luôn được coi là một trong những bài toán then chốt và đầy thách thức trong lĩnh vực tài chính và kinh tế số, với vai trò quan trọng trong việc tối ưu hóa chiến lược đầu tư, quản lý rủi ro cũng như hỗ trợ ra quyết định trong môi trường đầy biến động cho cả nhà đầu tư cá nhân lẫn tổ chức. Dự báo thị trường chứng khoán chính xác và tin cậy sẽ góp phần vào việc ra quyết định đúng đắn, cải thiện kết quả đầu tư và hiệu quả chung của thị trường tài chính. [1-4]. Tuy nhiên, một trong những công việc quan trọng nhất cũng như khó khăn nhất trong nghiên cứu tài chính là dự báo thị trường chứng khoán [3].

Trước đây, các nghiên cứu về dự báo thị trường chứng khoán chủ yếu sử dụng các phương pháp thống kê truyền thống, dựa trên việc phân tích dữ liệu lịch sử để dự đoán giá

cổ phiếu và lợi nhuận [1],[3]. Giai đoạn từ năm 2010, các mô hình kinh tế lượng và thống kê truyền thống như trung bình động tích hợp hồi quy tự động ARIMA (AutoRegressive Integrated Moving Average), GARCH (Generalized Autoregressive Conditional Heteroskedasticity) và các mô hình hồi quy tuyến tính được áp dụng rộng rãi để dự báo giá cổ phiếu [4, 5]. Các phương pháp này dựa trên giả định rằng giá cổ phiếu là tuyến tính hoặc dừng, bộc lộ nhược điểm không nắm bắt được sự phức tạp của thị trường tài chính và tính không dừng của giá cổ phiếu. Giá cổ phiếu là một chuỗi thời gian không dừng vì chúng có xu hướng thay đổi theo thời gian, tăng hoặc giảm phụ thuộc vào nhiều yếu tố như tình hình kinh tế, tin tức, sự kiện, và tâm lý nhà đầu tư..., không có điểm cân bằng cụ thể và thường có xu hướng tăng hoặc giảm theo thời gian dài. Tuy nhiên, các mô hình ARIMA hay GARCH yêu cầu chuỗi thời gian phải có tính dừng để đảm bảo độ chính xác cho các dự báo. Chúng dựa trên các giá trị quá khứ của giá cổ phiếu và giả định rằng mối quan hệ giữa các biến là tuyến tính. Mặc dù có phiên bản mở rộng như EGARCH hoặc TGARCH, GARCH nhưng các mô hình này vẫn gặp khó khăn trong việc xử lý các mối quan hệ đa biến phức tạp trên dữ liệu chứng khoán và các yếu tố tương tác giữa các cổ phiếu hoặc giữa các thị trường khác nhau. Sau đó, các phương pháp học máy như SVM (Support Vector Machines) và Cây quyết định (Decision Trees) đã phần nào khắc phục được một số nhược điểm của các mô hình dự báo thống kê truyền thống như dự đoán các xu hướng giá cổ phiếu ngắn hạn dựa trên các chỉ báo kỹ thuật và các đặc trưng trong dữ liệu quá khứ, được áp dụng dưới dạng SVM hồi quy (SVR - Support Vector Regression), giúp dự đoán giá trị liên tục, phù hợp với dữ liệu chuỗi thời gian. Tuy nhiên, bản chất của thị trường chứng khoán là phi tuyến, phức tạp và dễ biến động, cùng với nhiều yếu tố ảnh hưởng đến biến động giá như giá lịch sử, khối lượng giao dịch, sự kiện chính trị, tin tức, báo cáo tài chính của công ty, tâm lý thị trường, dữ liệu kinh tế vĩ mô..., khiến việc dự báo chính xác giá cổ phiếu gặp nhiều thách thức đáng kể [1, 2].

Trong bối cảnh đó, sự ra đời và phát triển nhanh chóng các kỹ thuật học sâu đã tạo ra những ảnh hưởng đáng kể trong lĩnh vực này. Các mô hình LSTM (Long Short Term Memory – Bộ nhớ dài ngắn hạn), CNN (Convolutional Neural Network – Mạng nơ-ron tích chập) và RNN (Recurrent Neural Network – Mạng nơ-ron hồi tiếp) đã cho thấy sự vượt trội trong khả năng nắm bắt các mối quan hệ phụ thuộc về thời gian và không gian trong dữ liệu chuỗi thời gian tài chính so với các phương pháp thống kê và học máy truyền thống [6-13]. Các mô hình dự báo dựa trên học sâu có thể tự động học và thích ứng với các mô hình phức tạp, có khả năng xử lý khối lượng dữ liệu lớn và phân biệt các kết nối phi tuyến phức tạp giữa các thuộc tính đầu vào và các mục tiêu dự đoán, khiến chúng có hiệu quả vượt bậc so với các mô hình tuyến tính và học máy truyền thống, đặc biệt trong lĩnh vực dự đoán thị trường chứng khoán. Một số phương pháp để tăng tính chính xác dự báo và hiệu suất mô hình đã tích hợp mô hình thống kê với học sâu, các kỹ thuật tối ưu hay cơ chế chú ý hoặc kết hợp (lai) các mô hình với nhau [6-13].

Nghiên cứu "Dự báo thị trường chứng khoán với Deep Learning: Đánh giá tổng quan giai đoạn 2020-2024" thực hiện việc tổng hợp, phân tích và thảo luận các nghiên

cứu từ năm 2020 đến năm 2024 về áp dụng các mô hình học sâu tiên tiến cho dự báo thị trường chứng khoán. Mục tiêu của nghiên cứu này là cung cấp một đánh giá tổng quan về các tiến bộ trong lĩnh vực này, đồng thời phân tích ưu điểm, nhược điểm của các mô hình và đề xuất một số hướng nghiên cứu cho tương lai.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Dự báo giá trong thị trường tài chính là chủ đề nghiên cứu ngày càng thu hút và được phát triển mạnh mẽ bởi những lợi ích kinh tế mà nó mang lại. Có thể nói rằng, trong thời gian gần đây, các mô hình học sâu đã cách mạng hóa lĩnh vực dự báo tài chính nhờ khả năng xử lý dữ liệu phi tuyến và đa chiều.

Tuy nhiên, thị trường tài chính là một hệ thống động, nơi giá cổ phiếu không chỉ đơn giản phụ thuộc vào dữ liệu lịch sử mà còn chịu ảnh hưởng sâu sắc từ nhiều yếu tố. Tính phi tuyến của dữ liệu chứng khoán, sự phức tạp và khả năng biến động mạnh của thị trường tài chính; cùng với các sự kiện chính trị, các kênh tin tức và nền tảng truyền thông, các yếu tố kinh tế vĩ mô.... tất cả những điều này đều tác động đến tâm lý công chúng, ảnh hưởng sâu rộng đến toàn bộ thị trường chứng khoán, dẫn tới những thách thức đáng kể đối với dự báo chính xác thị trường chứng khoán [1-10]. Nhiều phương pháp khác nhau đã được đề xuất để khắc phục các vấn đề trên, trong đó việc kết hợp các mô hình học sâu với nhau, hoặc kết hợp các mô hình học sâu với các mô hình thống kê hay xây dựng các mô hình tích hợp với kỹ thuật tối ưu hoá được sử dụng rộng rãi [6-13].

Nghiên cứu này nhằm mục đích cung cấp một phân tích toàn diện về ứng dụng của các kỹ thuật học sâu trong dự báo giá thị trường chứng khoán với nội dung xác định các xu hướng xây dựng mô hình học sâu hiện tại, các kỹ thuật đánh giá hiệu suất mô hình phổ biến và khám phá các mô hình lai mới trong giai đoạn 2020-2024, từ đó đề xuất các hướng nghiên cứu trong tương lai. Để làm rõ các nội dung trên, các câu hỏi nghiên cứu được đề xuất: 1. Các mô hình học sâu tiên tiến nào được sử dụng để dự đoán giá thị trường chứng khoán? 2. Các chỉ số phổ biến được sử dụng để phân tích trong các mô hình dự đoán? 3. Các chỉ số phổ biến được sử dụng để đánh giá hiệu suất của các mô hình dự đoán? 4. Bộ dữ liệu được sử dụng phổ biến trong các nghiên cứu?

Trong nghiên cứu, chúng tôi sử dụng phương pháp thu thập, phân tích và đánh giá các tài liệu liên quan nhằm giải quyết các câu hỏi nghiên cứu đặt ra. Các nghiên cứu được thu thập từ nguồn cơ sở dữ liệu học thuật của Google Scholar, các bài báo được lấy chủ yếu từ cơ sở dữ liệu IEEE Xplore, Springer Nature và Science Direct, với từ khoá được sử dụng gồm: "stock market", "financial market", "predict", "forecasting", "deep learning" kết hợp với các toán tử Boolean AND, OR. Từ các kết quả ban đầu, chúng tôi đã lọc tiêu đề, tóm tắt, tiêu chí từ khoá và thực hiện quy trình so sánh với các tiêu chí nghiên cứu, loại bỏ các bài báo trùng lặp và không liên quan, sau đó phân tích đánh giá các bài báo được nghiên cứu trong

giai đoạn 2020 đến 2024, nhằm đảm bảo bài đánh giá phản ánh được những tiến bộ mới nhất và xu hướng hiện tại trong lĩnh vực phân tích thị trường chứng khoán và mô hình dự báo giá cổ phiếu.

Để đánh giá chất lượng nguồn tài liệu, chúng tôi sử dụng bảng PRISMA chọn lọc nghiên cứu như sau:

Giai đoạn	Số tài liệu
<i>Nhận dạng:</i> - Các cơ sở dữ liệu IEEE Xplore, Springer Nature và Science Direct - Từ khóa: “stock market”, “financial market”, “predict”, “forecasting”, “deep learning” kết hợp với các toán tử Boolean AND, OR	245
<i>Lọc dữ liệu:</i> - Loại bỏ trùng lặp	178
- Lọc theo tóm tắt và toàn văn để đánh giá sự phù hợp với mục tiêu và các câu hỏi nghiên cứu	69
<i>Đánh giá toàn văn:</i> - Lý do loại bỏ: thiếu dữ liệu thực nghiệm, không có hiệu suất đánh giá, không có phương pháp luận, không so sánh với mô hình cơ sở	38
<i>Nghiên cứu đảm bảo chất lượng:</i>	19

19 bài toàn văn được xem xét phải đảm bảo các tiêu chí: chứa dữ liệu thỏa mãn các câu hỏi nghiên cứu, dữ liệu được thực nghiệm phải có thời gian 5 năm trở lên, các chỉ số đánh giá hiệu suất mô hình được sử dụng phổ biến, có so sánh với một mô hình truyền thống (ARIMA, SVM...) và/hoặc một mô hình deep learning khác, sử dụng bộ dữ liệu thị trường chính thức (Yahoo Finance, Bloomberg, Kaggle...).

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Các mô hình học sâu phổ biến trong dự báo thị trường chứng khoán

Theo kết quả phân tích, các mô hình học sâu được sử dụng phổ biến để dự đoán giá thị trường chứng khoán là RNN với các biến thể LSTM, GRU (Gated Recurrent Unit) với khả năng xử lý các phụ thuộc dài hạn trong dữ liệu chuỗi thời gian [6-12]. RNN có cấu trúc “mạch nối vòng” (recurrent connection), nghĩa là đầu ra của bước tính trước được đưa trở lại làm một phần đầu vào của bước hiện tại, cho phép RNN “ghi nhớ” một số thông tin ngắn hạn từ quá khứ khi dự báo bước tiếp theo. Tuy nhiên khi xử lý các chuỗi dữ liệu dài, thông tin từ đầu chuỗi có thể không được truyền lại một cách hiệu quả do các gradient giảm dần theo số lớp, từ đó làm giảm đáng kể hiệu suất dự báo của mô

hình. Các kiến trúc như LSTM và GRU được thiết kế nhằm khắc phục vấn đề này, bằng cách sử dụng các cơ chế cổng để điều chỉnh, điều phối và duy trì thông tin qua các bước thời gian, từ đó phép mô hình duy trì và ghi nhớ các thông tin quan trọng trong khoảng thời gian dài hơn.

Bên cạnh đó, mô hình CNN và Transformer cũng xuất hiện phổ biến trong các tài liệu với vai trò đặc trưng của mình [8,12,14, 19]. CNN là kiến trúc mạng nơ-ron được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc lưới, sử dụng các lớp tích chập (convolutional layers) để trích xuất đặc trưng từ hình ảnh, giúp phát hiện các mẫu như cạnh, góc và kết cấu, có hiệu suất vượt trội trong trích xuất hiệu quả các đặc trưng không gian và thời gian từ dữ liệu chuỗi thời gian.

Trong khi RNN và LSTM gặp khó khăn trong việc nắm bắt các mối quan hệ dài hạn do hiện tượng mất mát gradient (*vanishing gradient*), mô hình Transformer được xem như bước đột phá trong lĩnh vực học sâu, sử dụng cơ chế chú ý (*attention*)- tập trung vào các phần quan trọng của dữ liệu đầu vào khi đưa ra dự đoán, cho phép xử lý toàn bộ chuỗi dữ liệu cùng một lúc so với cơ chế xử lý dữ liệu tuần tự của CNN và RNN, cải thiện nhược điểm nắm bắt phụ thuộc dài hạn và mất mát thông tin.

Ngoài ra, để cải thiện hiệu suất và độ chính xác của mô hình, nhiều nghiên cứu đã kết hợp (lai) các mô hình với nhau [8-12] hoặc kết hợp các kỹ thuật tối ưu như thuật toán đàn ong ABC (Artificial Bee Colony), tối ưu thụ phấn hoa FPA (Flower Pollination Algorithm) hay tối ưu hoá bầy hạt PSO (Particle Swarm Optimization) [6, 7, 10] và cơ chế chú ý [11]. Thuật toán tối ưu PSO [6] tìm ra bộ thông số tối ưu cho LSTM để bắt xu hướng chính và bỏ qua biến động nhiễu ngắn hạn của giá cổ phiếu, giúp mô hình dự đoán chính xác hơn. Nghiên cứu [7] sử dụng thuật toán ABC như một “đàn ong thông minh” tự động tìm ra công thức tối ưu để cân bằng ảnh hưởng dữ liệu cảm xúc từ mạng xã hội và dữ liệu giá cổ phiếu, vừa bám sát dữ liệu giá lịch sử, vừa phản ứng nhanh với tin tức, giúp nhà đầu tư dự đoán sớm biến động. [10] kết hợp RNN-LSTM và hai thuật toán tối ưu PSO và FPA xử lý dữ liệu giá và khối lượng giao dịch trong ngày, kết hợp các chỉ số kỹ thuật và tối ưu các tham số mô hình để đạt được dự báo chính xác nhất. [11] đề xuất mô hình CNN-BiLSTM-Attention (CLATT) để dự báo tương quan cổ phiếu ngắn hạn trong đó CNN lọc nhiễu, trích xuất đặc trưng quan trọng từ dữ liệu giá; cơ chế chú ý đánh dấu điểm quan trọng, phân bổ trọng số cho các đặc trưng mà CNN đã trích xuất và cuối cùng, BiLSTM phân tích hai chiều để dự đoán giá cổ phiếu ngày tiếp theo với độ chính xác cao.

Các mô hình lai như CNN-GRU [8], LSTM-GRU [9], BiCuDNNLSTM-1dCNN [11, 12] cũng đã kết hợp khả năng của các mô hình như GRU để xử lý thông tin ngắn hạn trong chuỗi thời gian (biến động giá trong 5-10 phút); LSTM bổ sung "bộ nhớ dài hạn" so với GRU, phát hiện xu hướng tuần/ngày cũng như chủ động loại bỏ thông tin không liên quan; Transformer có khả năng tính toán trọng số ảnh hưởng giữa tất cả các cặp điểm dữ liệu trong chuỗi thời gian và đặc trưng, cơ chế đa chú ý phân tích song song

nhiều mối quan hệ khác nhau giữa các bước thời gian, cho biết yếu tố nào ảnh hưởng đến giá cổ phiếu, khắc phục được hạn chế xử lý tuần tự của RNN hay LSTM, tăng độ chính xác trong dự báo dài hạn.

Đặc điểm của từng nhóm mô hình được biểu diễn trong bảng sau:

Nhóm mô hình	Công trình tiêu biểu	Đầu vào	Độ sâu mô hình	Ưu thế về hiệu suất	Hạn chế thực tế
LSTM	[16], [19]	-Giá OHLCV -Dữ liệu tweet của người dân về cổ phiếu	-Gồm mô hình LSTM với cấu trúc 4 lớp	-Độ chính xác dự báo $\geq 93\%$	- Dự báo gặp khó khăn trong trường hợp giá cổ phiếu biến động mạnh - Yêu cầu nhiều tài nguyên và thời gian để huấn luyện, gặp vấn đề với biến mất/bùng nổ gradient
CNN	[11]	-Giá OHLCV	-Gồm năm thành phần chính: lớp đầu vào, module CNN, lớp BiLSTM, lớp self-attention và lớp đầu ra	-tăng độ chính xác dự đoán (giảm MAE, MSE và tăng R2) (~4.0-7.1%)	- Khó nắm bắt dữ liệu dài hạn - Dễ bị quá khớp dữ liệu (overfitting) khi có nhiều lớp tích chập
Transformer	[12]	dữ liệu lịch sử của các chỉ số thị trường chứng khoán, được cấu trúc như	Gồm bộ mã hoá và bộ giải mã	- giảm các chỉ số lỗi MAE (0.7-4.3%) và RMSE (1.9-3.4%) so với các mô hình cơ sở khác trong hầu hết các trường hợp và trên nhiều	- Kém hiệu quả khi dữ liệu ít/chuỗi dữ liệu quá ngắn -Nhu cầu tài nguyên (RAM/GPU) rất lớn

		một chuỗi thời gian		bộ dữ liệu khác nhau.	
Mô hình lai – Kết hợp các mô hình deep learning	[8], [9]	-Giá OHLCV - Giá OHLCV + chỉ báo kỹ thuật (SMA, EMA, RSI, MACD và ADX)	-Kết hợp các mô hình gồm 1D-CNN và GRU. 1D- CNN làm nhiệm vụ trích xuất đặc trưng và GRU hồi quy thời gian - Kết hợp mô hình GRU và LSTM để xử lý dữ liệu OHLC và các chỉ báo kỹ thuật để tối ưu hoá mô hình.	CNN-GRU cho thấy hiệu suất vượt trội hơn so với các mô hình cơ sở CNN-LSTM và CNN-RNN trên cả sáu bộ dữ liệu được thử nghiệm, với chỉ số RMSE luôn thấp nhất và R2 luôn cao nhất (RMSE giảm ~1.62-45.66 %, R2 tăng 0.02- 6.85%) - MAPE của mô hình đề xuất/mô hình hiện giảm 60-67%, cho thấy sự vượt trội về độ chính xác dự đoán.	- Cấu trúc phức tạp, phụ thuộc chất lượng dữ liệu đầu vào
Mô hình lai – kết hợp kỹ thuật tối ưu hoá (PSO, ABC, FPA)	[6], [7], [10]	- Giá OHLVC - Các bài đánh giá trên phương tiện truyền thông của các quỹ đầu tư	-Mô hình LSTM với hai lớp ẩn được tối ưu hóa bởi các thuật toán tối ưu	-Tốt hơn LSTM 52,63% (RMSE), 56,56% (MAE) và 9,28% (R²)	- tăng thời gian huấn luyện - cần điều chỉnh tham số trong thuật toán hội tụ

3.2. Các chỉ số phân tích trong dự đoán thị trường chứng khoán

Các yếu tố ảnh hưởng đến diễn biến của thị trường chứng khoán thường được gọi là các yếu tố tác động thị trường [2]. Những yếu tố này có thể được định lượng, và một hoặc nhiều yếu tố trong số đó có thể góp phần vào sự biến động của giá cổ phiếu. Nhiều nghiên cứu đã được thực hiện trong các môi trường kinh tế khác nhau để xác định các yếu tố chính ảnh hưởng đến hiệu suất của thị trường chứng khoán. Mặc dù có nhiều yếu tố có thể ảnh hưởng đến giá cổ phiếu, nhưng chúng thường là kết quả của sự kết hợp của nhiều yếu tố thay vì chỉ một nguyên nhân duy nhất [18]. Dựa trên kết quả phân tích các tài liệu thu thập, các chỉ số chính được sử dụng trong dự đoán thị trường chứng khoán gồm có 4 loại:

- *Chỉ số cơ bản*: giúp các nhà đầu tư phân tích tình hình tài chính, hoạt động và định giá của doanh nghiệp nhằm ước tính giá trị nội tại của cổ phiếu, cung cấp cái nhìn toàn diện về giá trị thực của cổ phiếu, phù hợp với các đầu tư dài hạn. Một số dữ liệu được sử dụng trong phân tích chỉ số cơ bản gồm giá thị trường của cổ phiếu: giá OHLVC – giá mở cửa, đóng cửa, cao nhất, thấp nhất; giá đóng cửa điều chỉnh, khối lượng giao dịch và doanh thu, cung cấp dữ liệu cơ bản để xác định khả năng giá trị cổ phiếu thay đổi so với giá thị trường hiện tại [8-17].

- *Chỉ số kỹ thuật*: công cụ quan trọng được các nhà giao dịch và nhà đầu tư sử dụng để dự đoán các biến động thị trường chứng khoán trong tương lai dựa trên dữ liệu về giá và khối lượng giao dịch trong quá khứ và hiện tại. Một số chỉ số phổ biến được sử dụng: SMA (Simple Moving Average)-đường trung bình động đơn giản: tính trung bình của giá đóng cửa (hoặc giá khác) trong một khoảng thời gian xác định; EMA (Exponential Moving Average)-đường trung bình động hàm mũ: nhận diện nhanh các thay đổi trong xu hướng giá, hỗ trợ xác định các điểm vào và ra giao dịch; MACD (Moving Average Convergence Divergence)-chỉ số trung bình động hội tụ/phân kỳ: giúp nhận diện sớm các dấu hiệu đảo chiều xu hướng; RSI (Relative Strength Index)-Chỉ số sức mạnh tương đối: đo lường tốc độ và sự thay đổi của biến động giá, từ đó dự báo khả năng đảo chiều của xu hướng giá; CCI (Commodity Channel Index)-Chỉ số kênh hàng hóa: có khả năng phát hiện các dao động mạnh mẽ của thị trường, hữu ích trong việc xác định xu hướng và phát hiện tín hiệu quá mua/quá bán [9, 10, 13].

- *Các yếu tố kinh tế vĩ mô* như tỉ giá hối đoái, nợ chính phủ, chỉ số tăng trưởng GDP, tỉ lệ lạm phát, biến động lãi suất và mối tương quan của chúng với hành vi của thị trường chứng khoán [14]: là các chỉ số đầu vào quan trọng trong việc dự báo giá cổ phiếu, cung cấp thông tin quan trọng về tình hình kinh tế, tâm lý thị trường và xu hướng đầu tư, giúp nhà đầu tư đưa ra quyết định chính xác hơn. Sự biến động của các chỉ số kinh tế vĩ mô như lạm phát cao làm giảm sức mua và tăng chi phí đầu vào, dẫn đến giảm lợi nhuận và giá cổ phiếu; lãi suất cao làm tăng chi phí vay, giảm đầu tư và tiêu dùng, ảnh hưởng tiêu cực đến giá cổ phiếu.

[9] đã sử dụng LSTM và GRU để trích xuất những thông tin có ý nghĩa, các đặc trưng từ dữ liệu thị trường chứng khoán như giá OHLCV của các cổ phiếu để dự báo giá trong tương lai. Các chỉ số kỹ thuật SMA, EMA, RSI, MACD được chọn lọc từ hàng loạt các chỉ báo

kỹ thuật và kết hợp với dữ liệu OHLCV nhằm tối ưu hóa mô hình dự báo. [14] đã thu thập dữ liệu về nợ chính phủ trên GDP từ Quỹ Tiền tệ Quốc tế (IMF) và Chỉ số Phát triển Thế giới (WDI) để nghiên cứu mối liên hệ quốc tế giữa nợ chính phủ và nguy cơ sụt giảm giá cổ phiếu. Nghiên cứu này sử dụng dữ liệu từ 86 quốc gia để chứng minh rằng nợ chính phủ cao có xu hướng làm tăng nguy cơ giá cổ phiếu giảm mạnh.

- *Tâm lý thị trường và dữ liệu xã hội*: dữ liệu trên mạng xã hội Twitter, Forum và Blog của nhà đầu tư; tâm lý tin tức tài chính, hành vi của các nhà đầu tư; các sự kiện địa chính trị: Nội dung của tin tức tài chính có sức ảnh hưởng rất lớn đến hành vi của nhà đầu tư và diễn biến thị trường, thông tin lan truyền trên thị trường, dù đúng hay sai, đều có thể tác động tức thì đến giá cổ phiếu. Mô hình CoATSMP [12] đã sử dụng Transformer để phân tích dữ liệu tin tức về lợi nhuận công ty với biến động giá cổ phiếu giúp nhận diện sớm cơ hội mua/bán, cho kết quả dự báo chính xác hơn nhờ kết hợp đa nguồn, đặc biệt trong các giai đoạn thị trường biến động mạnh (khủng hoảng, tin nóng). Nghiên cứu [16] sử dụng LSTM, XGBoost và kỹ thuật xử lý ngôn ngữ tự nhiên để trích xuất thông tin cảm xúc từ các bài đăng trên mạng xã hội Twitter, phân tích và đánh giá mối tương quan giữa sự biến động giá cổ phiếu và ý kiến của người dùng trên mạng xã hội. Nghiên cứu [17] đề xuất mô hình MEME-AO-LSTM để dự báo giá cổ phiếu với độ chính xác cao, đặc biệt trong bối cảnh thị trường biến động do các yếu tố chính trị (xung đột Nga-Ukraina) và kinh tế vĩ mô, sử dụng mô hình LSTM kết hợp xử lý dữ liệu tài chính truyền thống (giá, khối lượng giao dịch) với yếu tố phi cấu trúc (tin tức chính trị, các bài đăng từ mạng xã hội), chứng minh hiệu quả của việc tích hợp học sâu đa biến và yếu tố chính trị trong dự báo tài chính.

3.3. Các chỉ số phổ biến đánh giá hiệu suất các mô hình dự đoán

Trong lĩnh vực dự báo cổ phiếu, việc lựa chọn các chỉ số đánh giá phụ thuộc chặt chẽ vào mục tiêu cụ thể của nghiên cứu. Theo các nghiên cứu đã được phân tích trong bài đánh giá, các chỉ số đánh giá hiệu suất mô hình được sử dụng phổ biến nhất gồm:

- *Chỉ số lỗi*: đo lường mức độ sai lệch dự đoán: định lượng mức độ sai lệch giữa dự đoán của mô hình và giá thị trường thực tế: gồm các số liệu phổ biến như Sai số tuyệt đối trung bình (MAE), Sai số bình phương trung bình (MSE), Sai số bình phương trung bình căn bậc 2 (RMSE), Lỗi phần trăm tuyệt đối trung bình (MAPE) [6-9], [11], [12], [16, 17].

- *Chỉ số thống kê và tương quan*: đo lường mối tương quan thống kê và mối quan hệ giữa dự đoán của mô hình với dữ liệu thực tế: R^2 , Theil's U [6, 8, 11, 16, 17].

- *Chỉ số phân loại*: đánh giá hiệu suất của mô hình trong nhiệm vụ phân loại: gồm các chỉ số phổ biến như Precision, Recall, Accurarray, F-Score, Hit-ratio, [10] [19].

3.4. Bộ dữ liệu trong dự báo giá cổ phiếu

Trong dự báo giá cổ phiếu, chất lượng của tập dữ liệu là tiêu chí quan trọng để đánh giá hiệu suất mô hình, là nền tảng cho nghiên cứu. Nguồn dữ liệu lịch sử của cổ phiếu gồm giá đóng cửa, mở cửa, cao, thấp và khối lượng trong khung thời gian như hàng năm, hàng tháng, hàng ngày có thể được tải xuống từ Yahoo, <https://finance.yahoo.com>, Bloomberg <https://bloomberg.com>, Kaggle <https://kaggle.com> (quốc tế) hay <http://Investing.com> (Việt Nam). Các công trình được đánh giá đã sử dụng bộ dữ liệu S&P 500, chỉ số DJIA, NASDAQ [7, 9, 12, 16], NYSE [10] của thị trường Mỹ, tiếp theo là các chỉ số SSE [17], CSI100 [8], A-Share [11] của thị trường Trung Quốc, là những thị trường có các chỉ số ổn định. Một số thị trường có giá trị vốn hóa thị trường thấp hơn cũng được xem xét như NSE [9], BOM [10] của Ấn Độ, S&P 500/ASX200 [6] của Úc, LON [10] của Anh, NIKKEI25 [17] của Nhật, KOSPI [17] của Hàn Quốc, FTSE TWSE [13] của Đài Loan, VN-Index [19] của Việt Nam.

KẾT LUẬN

Kết quả của nghiên cứu đã chỉ ra RNN và các biến thể của nó như LSTM và GRU là những mô hình học sâu được sử dụng phổ biến trong dự báo giá của thị trường chứng khoán bởi khả năng xử lý phụ thuộc dài hạn trong dữ liệu chuỗi thời gian vượt trội. Các mô hình kết hợp như CNN với LSTM/GRU cũng được sử dụng rộng rãi với mục đích sử dụng thế mạnh của từng mô hình như CNN để trích xuất đặc trưng từ dữ liệu không gian (ảnh, biểu đồ giá) và LSTM/GRU để xử lý dữ liệu giá cổ phiếu. Các nghiên cứu cũng tập trung vào mô hình lai Transformer+LSTM để khắc phục hạn chế cũng như phát huy thế mạnh của từng mô hình-Transformer nắm bắt tốt các phụ thuộc dài hạn như xu hướng thị trường, còn LSTM bổ sung xử lý tuần tự ngắn hạn (biến động trong ngày).

Hạn chế trong các nghiên cứu hiện tại:

- *Phụ thuộc vào dữ liệu lịch sử*: Hầu hết các mô hình deep learning hiện nay như LSTM, CNN-LSTM, và Transformer, đều khai thác triệt để mối quan hệ trong chuỗi giá quá khứ để dự báo tương lai. Tuy nhiên, trong các thị trường có nhiều biến động, đặc biệt là các thị trường mới nổi như Việt Nam, Indonesia, Ấn Độ, Nga, Brazil... khi có các sự kiện ảnh hưởng như địa chính trị, khủng hoảng tài chính, biến động chính sách vĩ mô... sẽ dẫn tới khả năng dự đoán của mô hình bị giảm nghiêm trọng, tạo rủi ro cho nhà đầu tư và người tham gia thị trường.

- *Thiếu tính khả chuyển sang thị trường mới nổi*: Phần lớn các nghiên cứu hiện tại tập trung vào các chỉ số lớn của thị trường phát triển như NYSE, S&P 500 hay CSI 300, chỉ một số khảo sát đã thử nghiệm trên các thị trường mới nổi như Việt Nam, Ấn Độ, Indonesia và Brazil. Hơn nữa, hiện nay chưa có nhiều điều chỉnh để thích nghi với đặc thù của các thị trường này, bao gồm tính biến động cao, tính thanh khoản thấp, chi phí

giao dịch tương đối lớn, và rủi ro chính sách khó đoán. Điều này hạn chế khả năng áp dụng các mô hình dự báo hiệu quả trong bối cảnh toàn cầu.

- *Khó khăn trong việc xác minh và kiểm tra lại các kết quả nghiên cứu/mô hình*: nhiều nghiên cứu không công bố mã nguồn của mô hình, làm giảm tính khả thi của việc kiểm chứng và phát triển mô hình; quy trình tiền xử lý dữ liệu, chất lượng dữ liệu cũng như các thông tin về tham số mô hình không được công bố và kiểm soát đầy đủ làm giảm tính tin cậy của các nghiên cứu, ngăn cản sự phát triển, cải tiến trong lĩnh vực nghiên cứu cũng như giảm sự tin tưởng của các nhà đầu tư.

- *Dữ liệu đầu vào còn hạn chế*: Hầu hết các mô hình chỉ sử dụng dữ liệu giá OPHLCV và một số chỉ số kỹ thuật cơ bản, trong khi bỏ qua các tác động từ tin tức, tâm lý nhà đầu tư, và chi phí giao dịch thực tế. Việc tích hợp thông tin tài chính vĩ mô (như lãi suất, tỷ giá), phân tích cảm tính hoặc mô hình mô phỏng hành vi nhà đầu tư vẫn còn rất hạn chế. Những yếu tố này có thể cải thiện đáng kể độ chính xác dự báo, nhưng vẫn chưa được khai thác triệt để trong các nghiên cứu hiện tại. Các mô hình kết hợp nhiều tham số đầu vào cả tuyến tính và phi tuyến, có khả năng học theo mô hình các kỹ thuật AI như học liên tục, học từ tri thức và tích hợp tri thức đa ngành là những hướng mở trong tương lai.

Thị trường chứng khoán Việt Nam đã có những bước tiến đáng kể trong việc thu thập và lưu trữ dữ liệu. Các sàn giao dịch như HOSE và HNX cung cấp một lượng lớn dữ liệu giá cả, khối lượng giao dịch và các chỉ số tài chính, song song với dữ liệu kinh tế vĩ mô (GDP, CPI, tỷ giá) do Tổng cục Thống kê và Ngân hàng Nhà nước công bố, cùng dữ liệu cảm xúc từ các mạng xã hội (Zalo, Facebook, Twitter) đều có thể tích hợp để tăng chiều sâu cho mô hình. Tuy nhiên chất lượng và tính đầy đủ của dữ liệu vẫn là yếu tố then chốt ảnh hưởng đến độ tin cậy của các mô hình dự báo. Bên cạnh đó, để đáp ứng nhu cầu thực tế của xã hội, nhiều trường đại học và cơ quan nghiên cứu đã đào tạo hàng trăm kỹ sư, chuyên gia Data Science, Machine Learning; các hoạt động chuyên môn của cộng đồng AI/ML/DL trên Kaggle, Github... cũng rất hiệu quả. Các doanh nghiệp và viện nghiên cứu đã đầu tư trang bị các máy chủ GPU (NVIDIA Tesla, A100) hoặc tận dụng dịch vụ đám mây (AWS Singapore, Google Cloud, Viettel IDC) có hỗ trợ GPU/TPU; áp dụng ML/DL vào các hệ khuyến nghị đầu tư. Các công cụ và nền tảng Deep Learning như TensorFlow, PyTorch..., và các dịch vụ điện toán đám mây hiện nay ngày càng dễ tiếp cận hơn, tạo điều kiện thuận lợi cho cộng đồng lập trình viên và nhà nghiên cứu trong việc phát triển và triển khai các mô hình phức tạp, cho thấy Việt Nam tuy là thị trường mới nổi nhưng đã sẵn sàng và đầy tiềm năng trong việc triển khai deep learning vào dự báo.

TÀI LIỆU THAM KHẢO

- [1]. W. Bao, Y. Cao, Y. Yang, H. Che, J. Huang, S. Wen (2024), *"Data-driven stock forecasting models based on neural networks: A review"*, Information Fusion, Published by Elsevier B.V.
- [2]. R. Atoosa, A. Iheb, U. Amjad, (2025), *"Towards Economic Sustainability: A Comprehensive Review of Artificial Intelligence and Machine Learning Techniques in Improving the Accuracy of Stock Market Movements"*, International Journal of Financial Studies, 13(1).
- [3]. P. I Made, Kusrini, S. Arief, U. Ema (2024), *"Deep Learning in Financial Markets: A Systematic Literature Review of Methods and Future Direction for Price Prediction"*, International Conference on Cybernetics and Intelligent System (ICORIS), IEEE Xplore.
- [4]. G. Hyndman, R. and Athanasopoulos, (2018), *"Chapter 8. ARIMA Models, in: Forecasting: Principles and Practice (2nd ed.)"*, Otexts.
- [5]. C. Francq, J. Zakoćian, (2010), *"GARCH Models: Structure, Statistical Inference and Financial Applications"*, John Wiley and Sons, Ltd., Publication.
- [6]. Y. Ji, A. W. C. Liew, and L. Yang, (2021), *"A Novel Improved Particle Swarm Optimization with Long-Short Term Memory Hybrid Model for Stock Indices Forecast"*, IEEE Access, vol. 9, pp. 23660–23671.
- [7]. Kumar, R., Kumar, P., & Kumar, Y. (2022), *"Integrating big data driven sentiments polarity and ABC-optimized LSTM for time series forecasting"*, Multimedia Tools and Applications.
- [8]. Jaiswal, R., & Singh, B. (2022), *"A hybrid convolutional recurrent (CNN-GRU) model for stock price prediction"*, IEEE 11th International Conference on Communication Systems and Network Technologies.
- [9]. E. T. Sivadasan, N. Mohana Sundaram, and R. Santhosh, (2024), *"Stock market forecasting using deep learning with long short-term memory and gated recurrent unit"*, Soft computing, vol. 28, no. 4, pp. 3267–3282, Feb. 2024, doi: 10.1007/s00500-023-09606-7
- [10]. K. Kumar, M.T.U. Haider (2021), *"Enhanced Prediction of Intra-day Stock Market Using Metaheuristic Optimization on RNN–LSTM Network"*, New Generation Computing 39(10):1-42.
- [11]. A. Luo, L. Zhong, J. Wang, Y. Wang, S. Li, and W. Tai, (2024), *"Short-Term Stock Correlation Forecasting Based on CNN-BiLSTM Enhanced by Attention Mechanism"*, IEEE Access, vol. 12.
- [12]. Z. Tao, W. Wu, and J. Wang, (2024), *"Series decomposition Transformer with period-correlation for stock market index prediction"*, Expert Syst Appl, vol. 237, Mar. 2024.
- [13]. Wang, H.-C., Hsiao, W.-C., & Liou, R.-S. (2023). *"Integrating technical indicators, chip factors and stock news for enhanced stock price predictions: A multi-kernel approach"*. Asia Pacific Management Review.
- [14]. Ben-Nasr, H., & Boubaker, S. (2024), *"Government debt and stock price crash risk: International evidence"*, Journal of Financial Stability.
- [15]. Zhang, Q., Zhang, Y., Bao, F., Liu, Y., Zhang, C., & Liu, P. (2024), *"Incorporating stock prices and text for stock movement prediction based on information fusion"*, Engineering Applications of Artificial Intelligence.
- [16]. Ouf S., Hawary M., Aboutabl. A, Adel. S, (2024), *"A Deep Learning-Based LSTM for Stock Price Prediction Using Twitter Sentiment Analysis"*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 12, 2024.

- [17]. G. Quing, 2025, “Enhancing stock market Forecasting: A hybrid model for accurate prediction of S&P 500 and CSI 300 future prices”, Expert Systems with Applications.
- [18]. Liu, J., Li, H., Hai, M., & Zhang, Y., 2023, “A study of factors influencing financial stock prices based on causal inference”, Procedia Computer Science, 221, 861–869.
- [19]. Phuoc T., Anh P., Tam P, Chien V. Nguyen, 2024, “Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam”, Humanities and Social Sciences Communications.

DEEP LEARNING IN STOCK MARKET PREDICTIONS: A SYSTEMATIC REVIEW (2020-2024)

Nguyen Thi Huong Giang^{1*}, Le Manh Thanh², Nguyen Dinh Hoa Cuong³

¹ Faculty of Informatics, University of Education, Hue University

² Faculty of Information Technology, University of Sciences, Hue University

³ Faculty of Business Technology, Phu Xuan University

* Email: nthgiang@hueuni.edu.vn

ABSTRACT

This paper reviews the literature on the use of deep learning models in the field of stock price forecasting from 2020 to 2024, discussing popular models used in forecasting, variables in stock market predictions, performance metrics, and datasets in the models. The paper analyzes the advantages and disadvantages of the models and proposes some potential future research directions.

Keywords: predict, forecasting, stock market, stock price, deep learning.



Nguyễn Thị Hương Giang sinh năm 1980 tại Thành phố Huế. Bà tốt nghiệp Cử nhân ngành Công nghệ thông tin năm 2002 và Thạc sĩ ngành Khoa học máy tính năm 2005 tại Trường Đại học Khoa học, Đại học Huế. Bà hiện đang là giảng viên Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế.

Lĩnh vực nghiên cứu: cơ sở dữ liệu, ontology, học máy và học sâu, hệ ra quyết định.



Lê Mạnh Thạnh sinh năm 1953 tại Quảng Trị. Ông nhận học vị Tiến sĩ ngành khoa học máy tính tại Đại học Budapest (ELTE), Hungary vào năm 1994 và trở thành Phó giáo sư tại trường Đại học Khoa học, Đại học Huế vào năm 2004. Ông hiện đang là giảng viên Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: cơ sở dữ liệu, cơ sở tri thức và lập trình logic.



Nguyễn Đình Hoa Cương nhận bằng Thạc sĩ Khoa học và Tiến sĩ chuyên ngành Khoa học Máy tính tại Trường Đại học Khoa học, Đại học Huế và tại Đại học Khon Kaen, Thái Lan lần lượt vào các năm 2005 và 2016. Hiện nay, ông là giảng viên tại Trường Đại học Phú Xuân, Huế.

Lĩnh vực nghiên cứu: khai phá dữ liệu, khai phá văn bản, quản lý tri thức và hệ thống hỗ trợ ra quyết định