TRƯỜNG ĐẠI HỌC KINH TẾ LUẬT

BÁO CÁO CUỐI KỲ MÔN HỌC

TÊN MÔN HỌC: PHÂN TÍCH DỮ LIỆU LỚN TRONG TÀI CHÍNH

MÃ HỌC PHẦN: 241CN1302

ỨNG DỤNG PYSPARK TRONG PHÂN TÍCH KỸ THUẬT VÀ DỰ ĐOÁN GIÁ CỔ PHIẾU

Giảng viên hướng dẫn: TS. Nguyễn Thôn Dã

Danh sách thành viên nhóm 3:

1. K214141338, Võ Hưng Thạnh
2. K214140951, Trần Bích Quân
3. K214140934, Trần Minh Châu
4. K214142106, Nguyễn Thị Mai Vàng
5. K214141337, Nguyễn Đình Ngân Sơn

Thành phố Hồ Chí Minh, 2024

Bảng tự đánh giá thành viên nhóm

STT	Họ tên	MSSV	Điểm tự đánh giá (thang điểm 10)
1	Võ Hưng Thạnh	K214141338	10/10
2	Trần Bích Quân	K214140951	10/10
3	Trần Minh Châu	K214140934	10/10
4	Nguyễn Thị Mai Vàng	K214142106	10/10
5	Nguyễn Đình Ngân Sơn	K214141337	10/10

Lời cảm ơn của nhóm

Lời đầu tiên, chúng tôi xin chân thành cảm ơn sự hướng dẫn tận tình và góp ý quý báu của TS. Nguyễn Thôn Dã trong suốt quá trình tìm hiểu môn Phân tích dữ liệu lớn trong Tài chính. Với sự giúp đỡ tận tình của giảng viên, nhóm chúng tôi đã có thể hoàn thành tốt đề tài "Ứng dụng Pyspark trong phân tích kỹ thuật và dự đoán giá cổ phiếu", đồng thời nâng cao được kiến thức và kỹ năng về việc ứng dụng công nghệ phân tích dữ liệu lớn vào lĩnh vực tài chính. Trong quá trình làm bài chắc chắn khó tránh khỏi những thiếu sót. Do đó, chúng tôi kính mong nhận được những lời góp ý của quý báu thầy để đồ án của nhóm ngày càng hoàn thiện hơn.

Chúng tôi xin chân thành cảm ơn!

Tập thể thành viên nhóm

Lời cam kết

Chúng tôi cam đoan kết quả nghiên cứu này là của riêng chúng tôi, chúng tôi khẳng định không sao chép kết quả nghiên cứu của những cá nhân hoặc nhóm nghiên cứu nào khác.

Thành phố Hồ Chí Minh, ngày 14 tháng 11 năm 2024

Tập thể thành viên nhóm

Mục lục

Danh mục hình ảnh

Danh mục bảng

Danh mục biểu đồ

Danh mục từ viết tắt

Từ viết tắt	Từ đầy đủ	Mô tả
LSTM	Long Short Term Memory	Mạng bộ nhớ dài ngắn
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
CNN	Convolutional Neural Network	Mạng neural tích chập
GRU	Gated Recurrent Unit	Nút hồi tiếp có cổng
DNN	Deep Neural Networks	Mạng nơ-ron sâu
HOSE	Hochiminh Stock Exchange	Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh
JVM	Java Virtual Machine	Máy ảo Java
MACD	Moving Average Convergence Divergence	Trung bình động hội tụ phân kỳ
RSI	Relative Strength Index	Chỉ số sức mạnh tương đối
ANN	Artificial Neural Networks	Mạng thần kinh nhân tạo
FFNN	Feedforward Neural Network	Mạng thần kinh truyền thẳng
MSE	Mean Squared Error	Sai số toàn phương trung bình
RMSE	Root Mean Square Error	Độ lệch bình phương trung bình gốc
MAE	Mean Absolute Error	Sai số tuyệt đối trung bình
MAPE	Mean Absolute Percentage Error	Sai số phần trăm tuyệt đối trung bình
R^2	R-Squared	Hệ số xác định
DAX	Deutscher Aktien Index	Chỉ số thị trường chứng khoán đại diện cho 30 công ty lớn nhất tại Đức
DOW 30	Dow Jones Utility Average	Chỉ số trung bình của 30 công ty có giá trị cổ phiếu lớn nhất của Mỹ
S&P 500	Standard & Poor's 500 Stock Index	Chỉ số chứng khoán được dựa trên vốn hóa của 500 công ty đại chúng lớn nhất nước Mỹ
CSI 300	Shanghai Composite 300	Chỉ số thị trường chứng khoán theo vốn hóa của 300 cổ phiếu hàng đầu

		trên Sở giao dịch chứng khoán Thượng Hải và Thâm Quyến
SMA	Simple Moving Average	Đường trung bình động đơn giản
VN30	VN30 Equal Weight Index	Tập hợp 30 cổ phiếu hàng đầu trên thị trường chứng khoán Việt Nam
VNI	VN – Index	Chỉ số thị trường
RDD	Resilient Distributed Dataset	Một cấu trúc dữ liệu trong Apache Spark
ICB	Industry Classification Benchmark	Điểm chuẩn phân loại ngành
MA50	Moving Average 50 days	Đường trung bình động 50 ngày
MA100	Moving Average 50 days	Đường trung bình động 100 ngày
NIFTY 50	National Stock Exchange FIFTY	Tập hợp 50 công ty lớn nhất Ấn Độ

Tóm tắt báo cáo

Đồ án nghiên cứu ứng dụng Pyspark trong phân tích kỹ thuật và dự đoán giá cổ phiếu tại thị trường chứng khoán Việt Nam. Đồ án đặt ra mục tiêu phát triển phương pháp dự báo giá cổ phiếu hiệu quả bằng cách kết hợp công nghệ xử lý dữ liệu lớn cùng các mô hình học máy như LSTM, GRU, CNN và DNN. Nghiên cứu chỉ ra rằng thị trường chứng khoán Việt Nam đang có sự phát triển mạnh mẽ nhưng cũng gặp nhiều thách thức trong việc dự đoán giá cổ phiếu do sự biến động lớn. Việc ứng dụng Pyspark, một thư viện của Apache Spark, giúp xử lý và phân tích dữ liệu lớn một cách nhanh chóng, từ đó cải thiện độ chính xác trong các mô hình dự báo. Nội dung chính của đồ án bao gồm tổng quan lý thuyết về Pyspark, phân tích kỹ thuật, mô phỏng Monte Carlo và các mô hình học máy. Đồ án cũng tập trung vào việc thu thập và xử lý dữ liệu giá cổ phiếu từ 20 mã chứng khoán lớn tại Sở Giao dịch Chứng khoán TP.HCM trong khoảng thời gian 4 năm (2020-2024). Cuối cùng, đồ án kết luận rằng việc ứng dụng Pyspark và các mô hình học sâu có thể nâng cao khả năng dự đoán giá cổ phiếu, hỗ trợ các nhà đầu tư trong việc ra quyết định hiệu quả hơn.

1. Giới thiệu vấn đề nghiên cứu

1.1 Bối cảnh thị trường và lý do chọn đề tài

Trong những năm gần đây, thị trường chứng khoán Việt Nam đã phát triển mạnh mẽ, thu hút sự quan tâm của các nhà đầu tư trong và ngoài nước. Tuy nhiên, với sự biến động lớn về giá cổ phiếu, việc dự báo chính xác giá cổ phiếu trở thành một thách thức đối với các nhà đầu tư và tổ chức tài chính. Theo nghiên cứu của ông Nguyễn An Phú (7/2024) được đăng trên tạp chí Tài chính, thị trường chứng khoán Việt Nam chịu ảnh hưởng đáng kể từ các yếu tố kinh tế vĩ mô và tâm lý nhà đầu tư, dẫn đến những biến động giá khó lường.

Cùng với sự phát triển của khoa học dữ liệu và các công cụ xử lý dữ liệu lớn, việc ứng dụng công nghệ vào dự báo giá cổ phiếu trở thành một hướng nghiên cứu tiềm năng nhằm tối ưu hóa quy trình phân tích và giảm thiểu rủi ro đầu tư. Nghiên cứu của Cang, Quyên, và Ngoan (2024) chỉ ra rằng các công cụ xử lý dữ liệu lớn như Apache Spark mang lại lợi ích đáng kể trong việc xử lý tập dữ liệu lớn với tốc độ nhanh chóng, hỗ trợ các mô hình dự báo chính xác hơn và hiệu quả hơn. Trong bối cảnh đó, Pyspark – một thư viện của Apache Spark – là lựa chọn phù hợp để tối ưu hóa quá trình xử lý dữ liệu giá cổ phiếu lịch sử và phân tích mô hình dự báo, hứa hẹn tiềm năng cải thiện hiệu quả dự báo trên thị trường chứng khoán Việt Nam.

Một số nghiên cứu đã cho thấy hiệu quả của các mô hình học sâu và công cụ xử lý dữ liệu lớn trong dự báo giá cổ phiếu, đặc biệt trong bối cảnh dữ liệu phức tạp và khối lượng lớn. Nghiên cứu của Oanh và Châu (2024) đã chứng minh rằng mô hình LSTM có thể mang lại kết quả tích cực trong dự báo giá cổ phiếu tại Việt Nam khi được huấn luyện trên dữ liệu lịch sử từ các mã cổ phiếu lớn. Điều này cho thấy mô hình học sâu có thể là giải pháp phù hợp cho các chuỗi thời gian phức tạp của thị trường chứng khoán Việt Nam.

Ở bối cảnh quốc tế, Dey và cộng sự (2021) đã áp dụng mô hình kết hợp RNN, LSTM và GRU để dự báo giá cổ phiếu trên thị trường của công ty mô tô Honda và tập đoàn Orcale, cho thấy cách kết hợp này có thể khai thác cả đặc tính chuỗi thời gian và đặc điểm hình ảnh dữ liệu, từ đó cải thiện độ chính xác dự báo. Đây là một cách tiếp cận tiên tiến có thể là gợi ý cho thị trường Việt Nam nhằm tăng độ chính xác và giảm rủi ro khi dự báo giá cổ phiếu.

DNN có khả năng học các đặc trưng phức tạp trong dữ liệu và đã được chứng minh là hiệu quả trong dự báo tài chính, vượt trội hơn các mô hình tuyến tính nhờ khả năng học sâu (Kanchanamala, Karnati, Bhaskar Reddy, Practice, & Experience, 2023). Trong nghiên cứu này, DNN kết hợp với Pyspark để xử lý dữ liệu lớn nhằm phát hiện các xu hướng giá cổ phiếu, giúp cải thiện độ chính xác trong dự báo và hỗ trợ đầu tư.

Dựa trên những nghiên cứu này, đề tài ứng dụng Pyspark và các mô hình học sâu như LSTM, GRU, CNN và DNN nhằm xây dựng một quy trình xử lý dữ liệu tối ưu và hệ thống dự báo hiệu quả cho thị trường chứng khoán Việt Nam, góp phần vào việc phát triển các công cụ hỗ trợ ra quyết định đầu tư.

1.2 Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là phát triển một phương pháp dự báo giá cổ phiếu hiệu quả trên thị trường chứng khoán Việt Nam bằng cách kết hợp công nghệ xử lý dữ liệu lớn và các mô hình học sâu. Nghiên cứu hướng đến việc khai thác sức mạnh của Pyspark trong xử lý dữ liệu khối lượng lớn, đồng thời so sánh độ chính xác của các mô hình dự báo tiên tiến như LSTM, GRU, CNN và DNN. Kết quả nghiên cứu sẽ cung cấp nền tảng cho các ứng dụng thực tiễn trong đầu tư tài chính, giúp các nhà đầu tư tối ưu hóa quyết định dựa trên dự báo chính xác và nhanh chóng.

1.3 Phạm vi và đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài này là 20 mã cổ phiếu lớn được niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh, với dữ liệu giá lịch sử và thông tin về ngành của các mã cổ phiếu được thu thập từ Vnstock. Phạm vi nghiên cứu là dữ liệu lịch sử của các mã cổ phiếu trong vòng 4 năm, từ 2020 đến 2024 (tính đến thời điểm lấy dữ liệu).

2. Mô tả và tổng quan tình hình nghiên cứu

2.1 Cơ sở lý luận

Để phát triển các mô hình dự đoán giá cổ phiếu, cần phải thấu hiểu sự phức tạp và tính bất định vốn có của thị trường chứng khoán. Trong khi giả thuyết thị trường hiệu quả Fama (1970)

cho rằng giá tài sản luôn phản ánh mọi thông tin sẵn có ngay lập tức, và giả thuyết bước đi ngẫu nhiên Burton (2018) nhấn mạnh sự độc lập của giá cổ phiếu đối với lịch sử của nó, thì vẫn có những quan điểm cho rằng việc dự đoán giá cổ phiếu là khả thi ở một mức độ nào đó. Nghiên cứu của Lo và MacKinlay (2011) chỉ ra rằng giá cổ phiếu có thể được mô hình hóa và dự đoán thông qua các phương pháp phân tích dữ liệu tiên tiến.

Trong bối cảnh thị trường chứng khoán phức tạp, việc ứng dụng PySpark kết hợp với các mô hình học sâu như LSTM, GRU, DNN và CNN đang mở ra khả năng khai thác dữ liệu lớn và dự đoán giá cổ phiếu hiệu quả hơn. PySpark hỗ trợ xử lý dữ liệu song song trên quy mô lớn, giúp tăng tốc độ phân tích và huấn luyện mô hình. Các mô hình LSTM và GRU có thể lưu trữ thông tin dài hạn, giúp dự đoán giá dựa trên chuỗi thời gian, trong khi CNN và DNN có thể khai thác các mẫu phức tạp trong dữ liệu chuỗi thời gian khi được chuyển đổi thành dạng ảnh. Nghiên cứu này nhằm tối ưu hóa quá trình dự đoán giá cổ phiếu, đồng thời cung cấp góc nhìn khoa học và toàn diện về tác động của các yếu tố thị trường đến xu hướng giá.

2.1.1 Pyspark

Sự kết hợp giữa Python và Apache Spark thông qua PySpark tạo điều kiện thuận lợi cho việc xử lý và phân tích dữ liệu, đặc biệt hiệu quả với các tập dữ liệu lớn. PySpark hỗ trợ đa dạng tính năng của Apache Spark, bao gồm thư viện học máy (MLlib), DataFrames và SparkSQL. Sử dụng PySpark, người dùng có thể dễ dàng chuyển đổi giữa Apache Spark và Pandas, thực hiện xử lý dữ liệu theo luồng, tính toán phát trực tuyến và tương tác với các đối tượng JVM. Khả năng tương thích với các thư viện bên ngoài, như GraphFrames để phân tích biểu đồ và PySparkSQL cho phép xử lý khối lượng dữ liệu khổng lồ một cách linh hoạt, giúp PySpark trở thành công cụ lý tưởng cho những nhu cầu phân tích dữ liệu hiện đại.

PySpark là công cụ mạnh mẽ cho phép khai thác tốc độ xử lý của Apache Spark khi thao tác trên dữ liệu lớn. Nhờ trình báo PySpark, người dùng có thể tương tác và phân tích dữ liệu với hiệu suất nhanh hơn đáng kể so với khi dùng Python đơn thuần. PySpark cung cấp các tính năng nổi bật như tính toán trong bộ nhớ, khả năng chịu lỗi, và xử lý phân tán, hỗ trợ các trình quản lý cụm phổ biến như Yarn, Spark, và Mesos. Những tính năng này giúp PySpark xử lý và phân tích dữ liệu lớn, đáp ứng nhu cầu phức tạp của các hệ thống dữ liệu hiện đại.

2.1.2 Phân tích kỹ thuật

Phân tích kỹ thuật là một phương pháp phổ biến trong dự báo và mô hình hóa thị trường chứng khoán, dựa trên dữ liệu lịch sử về giá và khối lượng giao dịch. Phương pháp này dựa trên một số giả định chính: (1) giá cả được quyết định hoàn toàn bởi quan hệ cung cầu; (2) giá thay đổi theo các xu hướng; (3) những thay đổi về cung cầu sẽ dẫn đến sự đảo chiều của xu hướng; (4) các biến động về cung cầu có thể nhận diện trên biểu đồ; và (5) các mẫu hình giá trên biểu đồ có xu hướng lặp lại (Dahlquist & Kirkpatrick II, 2010). Điều này có nghĩa là phân tích kỹ thuật không xem xét các yếu tố ngoại vi như chính trị, xã hội hoặc yếu tố kinh tế vĩ mô.

Biondo, Pluchino, Rapisarda, và Helbing (2013) nhận thấy rằng các chiến lược giao dịch ngắn hạn dựa trên các chỉ báo phân tích kỹ thuật có thể mang lại hiệu quả cao hơn so với một số phương pháp truyền thống, chẳng hạn như MACD và RSI. Các tín hiệu mua hoặc bán được tạo ra từ phân tích kỹ thuật giúp dự đoán xu hướng thị trường thông qua việc phân tích các mức giá cụ thể, hỗ trợ nhà đầu tư trong việc ra quyết định giao dịch.

Bằng chứng thực nghiệm cũng ủng hộ phân tích kỹ thuật như một công cụ dự báo. Brock, Lakonishok, và LeBaron (1992) đã chỉ ra rằng các quy tắc giao dịch đơn giản dựa trên chuyển động của đường trung bình ngắn và dài hạn có sức mạnh dự báo đáng kể khi áp dụng trên dữ liệu hàng ngày của chỉ số công nghiệp Dow Jones trong hơn một thế kỷ. Nghiên cứu của Fifield, Power, và Donald Sinclair (2005) cũng tiếp tục kiểm nghiệm sức mạnh dự báo của quy tắc bộ lọc và quy tắc dao động trung bình động trên 11 thị trường chứng khoán châu Âu trong giai đoạn 1991-2000. Kết quả cho thấy, các thị trường mới nổi như Hy Lạp, Hungary, Bồ Đào Nha và Thổ Nhĩ Kỳ kém hiệu quả hơn so với các thị trường tiên tiến khác. Tuy nhiên, các kết quả này đôi khi bị phê bình về tính thiên lệch trong dữ liệu, Brock và cộng sự (1992) gợi ý rằng cần tiếp tục các nghiên cứu nhằm kiểm chứng thêm sức mạnh dự đoán của phân tích kỹ thuật.

2.1.3 Mô phỏng Monte Carlo

Phương pháp Monte Carlo là một kỹ thuật trong phân tích tài chính, đặc biệt được áp dụng để đánh giá hiệu quả của danh mục đầu tư nhằm tối đa hóa lợi nhuận và giảm thiểu rủi ro trong các điều kiện thị trường khác nhau. Phương pháp này dựa trên việc tạo ra hàng loạt kịch

bản ngẫu nhiên về giá tài sản và mức lợi nhuận, cho phép các ngân hàng và tổ chức tài chính có được những nhận định chi tiết về kết quả tiềm năng của danh mục và đưa ra các chiến lược quản lý rủi ro. Điều này giúp xây dựng các quyết định dựa trên dữ liệu nhằm đạt được sự cân bằng tối ưu giữa rủi ro và phần thưởng (Boyle, 1977).

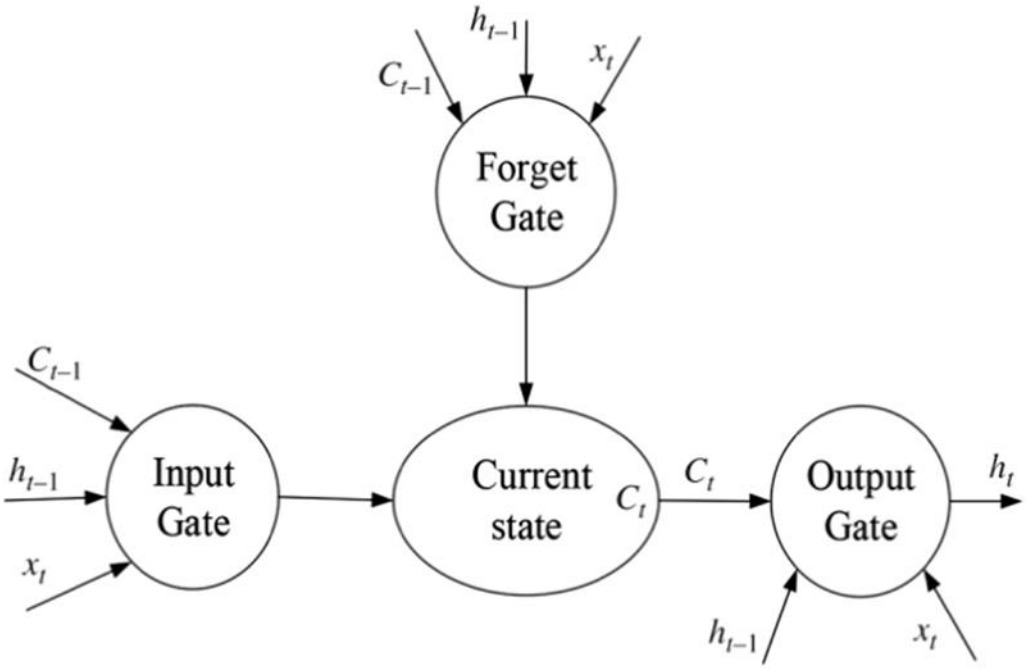
Monte Carlo được xem là một phần của phân tích số - lĩnh vực toán học chuyên về các phương pháp xấp xỉ để giải quyết các bài toán phức tạp không thể áp dụng các phương pháp xác định một cách hiệu quả. Nguyên lý chủ đạo của phương pháp là tìm kiếm các giải pháp gần đúng có độ chính xác cao bằng cách sử dụng trung bình cộng của nhiều mẫu ngẫu nhiên từ tập hợp các vấn đề hiện có (Carlo, 2001). Điều này đặc biệt hữu ích trong những hệ thống phức tạp với nhiều yếu tố không chắc chắn và các bộ phận tương tác, khi mà phương pháp Monte Carlo có thể cung cấp cái nhìn toàn diện về cách hệ thống hoạt động trong thực tế.

Phương pháp này được xây dựng trên cơ sở của định luật số lớn trong lý thuyết xác suất, cho rằng khi số lượng biến ngẫu nhiên tăng lên, trung bình cộng của chúng sẽ tiến gần đến giá trị kỳ vọng nếu giá trị này tồn tại (Chan & Kroese, 2011). Phương pháp Monte Carlo vì thế trở thành một công cụ không thể thiếu trong phân tích tài chính hiện đại, cho phép các tổ chức dự báo và tối ưu hóa trong bối cảnh nhiều yếu tố biến động và không chắc chắn.

2.1.4 Mô hình LSTM

Mạng Nơ-ron nhân tạo là một lớp mô hình học máy tiên tiến được phát triển nhằm khắc phục các hạn chế của các thuật toán truyền thống, vốn dựa trên các quy tắc lập trình cố định (LeCun, Bengio, & Hinton, 2015). ANN có thể phân loại thành hai dạng chính: mạng nơ-ron truyền thẳng và mạng nơ-ron hồi quy. Trong đó, FFNN thường được sử dụng rộng rãi trong phân loại dữ liệu, nhận dạng đối tượng và xử lý hình ảnh, nhưng lại gặp khó khăn trong việc xử lý các phụ thuộc theo thời gian. RNN, một dạng ANN khác do Elman (1990) đề xuất, giải quyết vấn đề này bằng cách cho phép thông tin từ các trạng thái trước tác động đến các trạng thái sau nhờ vào cấu trúc kết nối đệ quy, cho phép lưu trữ thông tin lịch sử để phân tích các chuỗi dữ liệu dài hạn. Tuy nhiên, trong quá trình huấn luyện, RNN phải đối mặt với hiện tượng "gradient biến mất" và "gradient bùng nổ", gây khó khăn trong việc duy trì và xử lý các chuỗi dữ liệu dài (Hochreiter, 1997).

Để khắc phục những hạn chế trên, mạng LSTM được Hochreiter (1997) đề xuất, với một kiến trúc đặc biệt gồm các ô nhớ (memory cell) và các cổng kiểm soát như cổng đầu vào, cổng quên và cổng đầu ra. Kiến trúc dựa trên cổng của LSTM cho phép thông tin được lưu trữ và truyền đi có chọn lọc, từ đó giúp mạng duy trì và xử lý hiệu quả các phụ thuộc dài hạn trong dữ liệu chuỗi, tránh được vấn đề gradient biến mất. Mô hình LSTM đã đạt được nhiều kết quả đáng chú ý trong các lĩnh vực như xử lý ngôn ngữ tự nhiên và nhận dạng chữ viết tay, nhờ khả năng duy trì thông tin cần thiết qua nhiều bước thời gian (Graves & Graves, 2012).



Hình 1: Mô hình LSTM

Các cổng kết nối trong LSTM đóng vai trò quan trọng trong việc điều chỉnh lượng thông tin cần lưu trữ hoặc bỏ qua tại mỗi bước thời gian. Ba cổng này, bao gồm cổng đầu vào, cổng quên và cổng đầu ra, đảm bảo thông tin đầu vào chỉ được chuyển tiếp khi đáp ứng các tiêu chí của mạng, trong khi những thông tin không cần thiết sẽ bị loại bỏ thông qua cổng quên (Graves & Schmidhuber, 2008). Các nghiên cứu sau này đã thử nghiệm một số biến thể của LSTM, nhưng nhìn chung, không có biến thể nào đạt được cải tiến đáng kể so với kiến trúc ban đầu của Hochreiter và Schmidhuber (Greff và cộng sự 2016). Kết cấu LSTM cho phép mô hình khắc phục những hạn chế của RNN truyền thống và trở thành công cụ phổ biến trong dự báo chuỗi thời gian phức tạp.

2.1.5 Mô hình GRU

Mô hình GRU (Gated Recurrent Unit) hay còn gọi là mạng nơ-ron hồi tiếp có nút cổng, được Chung, Gulcehre, Cho, và Bengio (2014) giới thiệu như một biến thể đơn giản hóa của LSTM. Khác với LSTM, GRU chỉ sử dụng hai cổng chính: cổng khởi tạo (reset gate) và cổng cập nhật (update gate). Thiết kế tối giản của GRU giúp giảm thiểu số lượng tham số cần thiết, từ đó nâng cao hiệu suất tính toán, giúp mô hình nhanh hơn trong quá trình huấn luyện và dự báo (Chung và cộng sự 2014). GRU thường được áp dụng hiệu quả trên các tập dữ liệu nhỏ và yêu cầu thời gian huấn luyện ngắn, đồng thời giúp giảm nguy cơ quá tải mô hình, một yếu tố quan trọng trong việc tránh hiện tượng quá khớp.

Nguyên lý hoạt động của GRU bao gồm hai bước chính. Đầu tiên, cổng khởi tạo r_t điều chỉnh mức độ tích hợp thông tin của trạng thái hiện tại với thông tin từ trạng thái trước đó, cho phép chọn lọc thông tin lưu trữ hoặc bỏ qua. Thứ hai, cổng cập nhật z_t (kết hợp vai trò của cổng đầu vào và cổng quên trong LSTM) quyết định xem có nên giữ lại thông tin từ trạng thái trước đó tại thời điểm hiện tại hay không. Phương trình toán học thể hiện quá trình tính toán của các đơn vị ẩn trong GRU như sau:

$$r_t = \sigma\left(\mathcal{W}_r\left[h_{t-1}, x_t\right] + b_r\right)$$

$$z_t = \sigma\left(\mathcal{W}_z\left[h_{t-1}, x_t\right] + b_z\right)$$

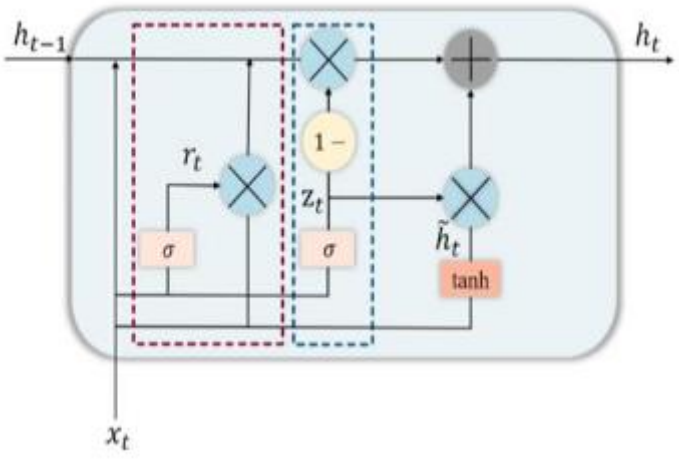
$$\check{h}_t = \varphi\left(\mathcal{W}_{\check{h}}\left[r_t * h_{t-1}, x_t\right] + b_{\check{h}}\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \check{h}_t$$

Trong đó $\mathcal{W}_r, \mathcal{W}_z, \mathcal{W}_{\check{h}}$ lần lượt là ma trận trọng lượng của $r_t, z_t, \check{h}_t, b_r, b_z, b_{\check{h}}$ biểu thị các vecto thiên vị của $r_t, z_t, \check{h}_t$ tương ứng. Phương trình cuối cùng cho thấy trạng thái ẩn hiện tại h_t được xác định bởi sự kết hợp giữa đầu ra của trạng thái trước đó h_{t-1} và trạng thái ứng viên hiện tại $\check{h}_t$, tạo thành một cấu trúc linh hoạt giúp mô hình có khả năng duy trì thông tin cần thiết.

Nhờ số lượng tham số ít hơn, GRU giảm thiểu nguy cơ quá tải trong quá trình huấn luyện, cho phép thực hiện dự báo nhanh chóng và hiệu quả. GRU cũng được chứng minh là có hiệu suất tốt tương đương hoặc thậm chí vượt trội so với LSTM trong một số tác vụ dự báo chuỗi

thời gian và xử lý ngôn ngữ tự nhiên (Chung và cộng sự 2014; Jozefowicz, Zaremba, & Sutskever, 2015)



Hình 2: Mô hình GRU

2.1.6 Mô hình CNN

CNN được giới thiệu lần đầu bởi LeCun và cộng sự (2015) như một phương pháp tiên tiến cho các tác vụ xử lý hình ảnh và nhận dạng mẫu. Cấu trúc CNN bao gồm năm thành phần chính: lớp đầu vào, lớp tích chập, lớp gộp (pooling layer), lớp kết nối đầy đủ (fully connected layer) và lớp đầu ra. Trong đó, lớp tích chập và lớp gộp là các lớp quan trọng nhất của mô hình, đóng vai trò chính trong việc trích xuất và giảm chiều dữ liệu đầu vào để tạo ra các đặc trưng phi tuyến tính. Lớp tích chập thực hiện phép lọc để trích xuất các tính năng quan trọng, trong khi lớp gộp giảm thiểu độ phức tạp của dữ liệu bằng cách nén các đặc trưng, giúp tăng khả năng khái quát hóa và giảm thiểu hiện tượng quá khớp (overfitting).

Với khả năng vượt trội trong việc trích xuất và nhận dạng đặc trưng từ dữ liệu, CNN đã được áp dụng rộng rãi và thành công trong các tác vụ như phân loại hình ảnh và xử lý chuỗi thời gian. Trong nghiên cứu này, CNN được sử dụng để trích xuất đặc trưng phi tuyến tính cục bộ từ dữ liệu chứng khoán thông qua các lớp tích chập, đồng thời nén và tối ưu hóa đặc trưng qua các lớp gộp. Điều này giúp mô hình tạo ra các thông tin đặc trưng quan trọng, giúp nâng cao khả năng dự đoán và phân tích xu hướng của dữ liệu chứng khoán.

2.1.7 Mô hình DNN

DNN là một kỹ thuật học máy được phát triển để cho phép máy tính thực hiện các tác vụ phức tạp thông qua việc huấn luyện trên các bộ dữ liệu lớn, thay vì sử dụng các quy tắc lập trình thủ công (LeCun và cộng sự 2015). Lấy cảm hứng từ hoạt động của não bộ con người, các mạng nơ-ron nhân tạo này được thiết kế để có khả năng học từ các lần lặp lại và kinh nghiệm, tương tự như cách con người dự đoán và đưa ra quyết định. DNN được xây dựng trên cấu trúc mạng với nhiều lớp ẩn, cho phép mô hình giữ lại và xử lý nhiều thông tin từ dữ liệu, qua đó giảm sự phụ thuộc vào kỹ thuật lựa chọn đặc trưng thủ công. Trong DNN, các lớp neuron liền kề được kết nối hoàn toàn. Ban đầu, trọng số của các kết nối này được thiết lập thông qua một quá trình huấn luyện trước không giám sát, và quá trình học chính thức bao gồm hai giai đoạn chính: lan truyền tiến (forward propagation) và lan truyền ngược (backpropagation).

Trong quá trình lan truyền tiến, đầu ra của mỗi neuron ở một lớp sẽ trở thành đầu vào cho lớp tiếp theo. Mối quan hệ giữa các neuron trong hai lớp liên kết có thể biểu diễn qua phương trình:

$$z = \sum_{i=1}^{m} w_i x_i + b$$

Trong đó m là số lượng neuron đầu vào, x_ii là đầu ra của neuron ở lớp trước, $w_i$là trọng số, và b là hệ số dịch. Để ngăn chặn giá trị đầu ra tăng vô hạn, các hàm kích hoạt như tanh, softmax hoặc ReLU thường được áp dụng. Khi mô hình lan truyền qua nhiều lớp, đầu ra cuối cùng sẽ được tính toán và đưa ra dự đoán dựa trên các đặc trưng phức tạp từ dữ liệu đầu vào. Phương trình này được mở rộng như sau: giả sử có m neutron trong lớp l – 1, đầu ra của neutron thứ j trong lớp l được biểu diễn bởi:

$$a_{ij} = \sigma(z_{ij}) = \sigma\left(\sum_{k=1}^{m} w_{ij} a_i^{l-1} + b_{ij}\right)$$

Nếu lớp l có n neutron, đầu ra của lớp l có thể được biểu diễn dưới dạng ma trận như sau:

$$a_i = \sigma(z_i) = \sigma(W_i a^{l-1} + b_{ij}$$

Trong đó W_l là ma trận hệ số trọng số kích thước $n\ x\ m$ trong lớp l, và b_l là vecto dịch có kích thước n x 1.

Thuật toán lan truyền ngược sau đó được áp dụng để tối ưu trọng số của mạng. Phương pháp này sử dụng một hàm mất mát để đo lường mức độ sai lệch giữa dự đoán của mô hình và dữ liệu thực tế, với mục tiêu giảm thiểu giá trị của hàm mất mát thông qua các lần lặp của thuật toán giảm độ dốc (gradient descent). Kết quả là, các giá trị tối ưu cho ma trận trọng số W và vector dịch b được tìm thấy, giúp nâng cao hiệu suất dự đoán của mô hình. Nhờ cấu trúc mạng phức tạp và khả năng xử lý dữ liệu lớn, DNN đã đạt được nhiều thành công trong các lĩnh vực như nhận diện hình ảnh, phân loại văn bản và phân tích chuỗi thời gian. Mô hình DNN giúp giảm thiểu các hạn chế của các mô hình truyền thống trong việc lựa chọn đặc trưng và cho phép xử lý thông tin phi tuyến tính hiệu quả hơn.

2.1.8 Các chỉ số đánh giá

Các chỉ số đánh giá được sử dụng trong báo cáo này bao gồm hệ số R^2 hệ số giải thích phương sai ev_score, nhằm đánh giá mức độ phù hợp của mô hình trong bối cảnh học máy (Géron, 2019). Các giá trị $r2score$ và ev_scoree đều nằm trong khoảng từ 0 đến 1; giá trị càng gần 1 thể hiện khả năng dự báo càng chính xác của mô hình, minh chứng cho sự hiệu quả trong các phân tích hồi quy (Nguyen, Lin, Huang, & Processing, 2023).

Hệ số R^2 được tính theo công thức:

$$r2score = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|^2}{\sum_{i=1}^{n} |y_i - \bar{y}|^2}$$

Trong đó, y_i là giá trị thực, $\hat{y}_i$ là giá trị dự báo, và $\bar{y}$ là giá trị trung bình của các quan sát thực tế.

Ngoài ra, báo cáo cũng sử dụng các chỉ số khác để đánh giá sai số trong dự, các chỉ số này cho thấy hiệu suất của các mô hình dự đoán:

Sai số bình phương trung bình gốc hay root mean squared error: Chỉ số này đo lường mức độ sai lệch trung bình giữa giá dự đoán và giá thực tế. Giá trị RMSE càng thấp, nghĩa là mô hình dự đoán càng chính xác.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2}$$

Sai số tuyệt đối trung bình hay mean absolute error: Chỉ số này đo lường sai lệch tuyệt đối trung bình giữa giá dự đoán và giá thực tế. Giá trị MAE càng thấp, nghĩa là mô hình dự đoán càng chính xác.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Sai số phần trăm tuyệt đối trung bình hay mean absolute percentage error: Chỉ số này đo lường sai lệch tương đối trung bình giữa giá dự đoán và giá thực tế. Giá trị MAPE càng thấp, nghĩa là mô hình dự đoán càng chính xác.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100$$

2.2 Thực trạng vấn đề nghiên cứu

2.2.1 Trên thế giới

Nghiên cứu của Song và Choi (2023) đã đề xuất ba mô hình lai dựa trên mạng nơ-ron hồi quy nhằm dự đoán giá đóng cửa một bước thời gian và nhiều bước thời gian của các chỉ số thị trường chứng khoán DAX, DOW và S&P500. Bao gồm CNN-LSTM, GRU-CNN, và mô hình tổng hợp kết hợp ba mô hình RNN, LSTM và GRU để khai thác tối đa khả năng biểu diễn của từng loại mô hình. Kết quả thực nghiệm cho thấy các mô hình lai này cho hiệu quả dự đoán tốt hơn so với các mô hình học máy truyền thống, trong đó mô hình tổng hợp có kết quả ấn tượng trong dự báo một bước thời gian. Thêm vào đó, việc sử dụng đặc trưng mới "trung bình" cũng như đặc trưng giá trị trung bình và khối lượng giao dịch giúp cải thiện hiệu suất của mô hình, đồng thời giảm số lượng đặc trưng nhằm hạn chế hiện tượng quá khớp. Nghiên cứu cũng cho thấy rằng, khung mô hình này có thể áp dụng vào các lĩnh vực khác như dự báo tiêu thụ năng lượng, giá dầu, nồng độ khí, chất lượng không khí và lưu lượng sông. Khả năng cải thiện hiệu suất dự báo trong tương lai có thể đạt được thông qua việc kết hợp

các loại mô hình dựa trên RNN khác nhau, hỗ trợ các nghiên cứu về xây dựng danh mục đầu tư dựa trên giá trị dự đoán của thị trường chứng khoán.

Trong bài nghiên cứu của Zhang, Ye, và Lai (2023) giới thiệu một mô hình dự đoán giá cổ phiếu dựa trên mạng nơ-ron sâu có tên là CNN-BiLSTM-Attention. Mô hình được thử nghiệm trên chỉ số CSI 300 của Trung Quốc và 12 chỉ số thị trường chứng khoán khác, cả trong và ngoài nước, và kết quả cho thấy nó hoạt động tốt hơn so với các mô hình dự đoán giá cổ phiếu khác như LSTM, CNN-LSTM và CNN-LSTM-Attention. Kết luận cho thấy rằng mô hình lai CNN-BiLSTM-Attention là một phương pháp đầy hứa hẹn để dự đoán giá cổ phiếu, và cung cấp những gợi ý cho nghiên cứu tiếp theo.

Trong khi đó, nghiên cứu của Mehtab và Sen (2020) đề xuất hai mô hình hồi quy được xây dựng trên mạng CNN và ba mô hình dự đoán dựa trên LSTM. Với mục đích dự báo các giá trị mở của các hồ sơ chỉ số NIFTY 50 bằng cách áp dụng một kỹ thuật dự đoán nhiều bước với xác thực tiến về phía trước. Với bộ dữ liệu, kết quả cho thấy mỗi mô hình riêng biệt sẽ có các ưu điểm khác nhau với LSTM tích chập mã hóa - giải mã đơn biến với dữ liệu của hai tuần trước làm đầu vào là mô hình chính xác nhất. Mặt khác, một mô hình CNN đơn biến với dữ liệu của một tuần trước làm đầu vào được phát hiện là mô hình nhanh nhất về tốc độ thực hiện của nó.

Kết quả từ nghiên cứu của Shah, Campbell, và Zulkernine (2018) cho thấy khả năng phân tích xu hướng thị trường chứng khoán có thể vô cùng có giá trị đối với các nhà đầu tư và nhà nghiên cứu. Nghiên cứu trình bày một nghiên cứu so sánh về hai mô hình mạng nơ-ron nhân tạo rất hứa hẹn, cụ thể là RNN, LSTM, DNN trong việc dự báo các biến động hàng ngày và hàng tuần của chỉ số BSE Sensex của Ấn Độ. Với cả hai mạng, các biện pháp đã được thực hiện để giảm tình trạng quá khớp. Dự đoán hàng ngày về giá cổ phiếu Tech Mahindra đã được thực hiện để kiểm tra khả năng khái quát hóa của các mô hình. Cả hai mạng đều thực hiện tốt trong việc đưa ra dự đoán hàng ngày và cả hai đều tổng quát hóa tốt để đưa ra dự đoán hàng ngày về dữ liệu Tech Mahindra. LSTM và RNN vượt trội hơn DNN về mặt dự đoán hàng tuần và do đó, hứa hẹn hơn trong việc đưa ra dự đoán dài hạn.

Như đã biết, dự đoán luôn là một nhiệm vụ đầy thách thức trong mọi khía cạnh. Mục tiêu của bất kỳ phương pháp dự đoán cổ phiếu nào là phát triển một phương pháp mạnh mẽ để dự đoán giá giao dịch có thể được sử dụng để cải thiện các quyết định đầu tư và các mô hình

chính xác. Nghiên cứu Sisodia và Boghey (2024) đề xuất một mô hình lai kết hợp các điểm mạnh của các mô hình học sâu CNN và DNN, và phát triển một phương pháp toàn diện để dự đoán giá cổ phiếu tại Ấn Độ. Trong bối cảnh dự đoán giá cổ phiếu, các lớp CNN có thể được sử dụng để trích xuất các tính năng từ dữ liệu đầu liên quan đến giá trị tương lai ước tính. Lớp DNN có thể được sử dụng để kết hợp các tính năng học được từ các lớp CNN. Hiệu suất của mô hình sẽ được đánh giá bằng nhiều số liệu khác nhau bao gồm MSE, RMSE, MAE, R^2 . Phương pháp này cũng phân tích tác động của nhiều yếu tố khác nhau lên giá cổ phiếu, bao gồm biến động thị trường, chỉ số kinh tế và các sự kiện địa chính trị. Độ chính xác đạt được là 97,48% cho thấy mô hình đã thành công trong việc dự đoán chính xác giá cổ phiếu của 50 công ty lớn nhất tại Ấn Độ. Phương pháp được đề xuất dự kiến sẽ cung cấp cho các nhà đầu tư và nhà phân tích tài chính một công cụ có giá trị để đưa ra quyết định đầu tư sáng suốt.

2.2.2 Tại Việt Nam

Nghiên cứu của Phuoc, Anh, Tam, và Nguyen (2024) dự báo xu hướng giá cổ phiếu trên thị trường chứng khoán tại một nền kinh tế mới nổi là Việt Nam với LSTM và các chỉ báo phân tích kỹ thuật tương ứng cho từng mã cổ phiếu bao gồm SMA, MACD, RSI và dữ liệu thứ cấp từ các cổ phiếu VN-Index và VN-30, kết quả nghiên cứu cho thấy mô hình dự báo có độ chính xác cao 93% đối với hầu hết dữ liệu cổ phiếu được sử dụng, chứng minh tính phù hợp của mô hình LSTM và bộ dữ liệu kiểm tra được sử dụng để đánh giá hiệu suất của mô hình. Đồng thời, bài nghiên cứu Trịnh, Trần, Hà, Trần, và Đỗ (2022), Oanh và Châu (2024) cũng đưa ra kết quả tương tự rằng LSTM là tốt nhất trong các mô hình.

Nghiên cứu với tiêu đề "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" của Tuyên (2024) với mô hình mạng nơ-ron LSTM và GRU và các phức hợp được thiết kế bằng ngôn ngữ lập trình Python với các gói phụ trợ có sẵn, cho thấy kết quả dự báo với độ chính xác cao, hiệu suất của mô hình LSTM-GRU Hybrid cho kết quả tốt nhất. Thông qua mô hình LSTM-GRU Hybrid, nghiên cứu dự báo xu hướng biến động chỉ số VNIndex 100 ngày tiếp theo cho kết quả chỉ số VNIndex có xu hướng tăng. Điều đó gián tiếp chỉ ra rằng thị trường chứng khoán Việt Nam có dấu hiệu khởi sắc trở lại cùng với các chính sách mới của Chính phủ. Tuy nhiên, nghiên cứu cũng nhận thấy rằng việc chuẩn bị dữ liệu, lựa chọn siêu tham số và kiểm tra hiệu suất mô hình vẫn còn là những thách thức, cần có sự

cân nhắc và kiểm tra kỹ lưỡng để đảm bảo tính chính xác và tin cậy của kết quả dự báo. Bên cạnh đó, một số vấn đề cần phải nghiên cứu thêm như: học máy trong việc nghiên cứu tâm lý đám đông, các chính sách của Chính phủ, các vấn đề khác trên thế giới ảnh hưởng đến thị trường chứng khoán của Việt Nam hiện nay. Song song, mô hình GRU cho kết quả tương đối tốt, nhưng gặp phải hiện tượng quá khớp và nghiên cứu cũng chưa đề xuất các giải pháp khắc phục phù hợp nhất.

2.2.3 Phát triển giả thuyết nghiên cứu

Những nghiên cứu tại Việt Nam đã bước đầu áp dụng học máy với các mô hình học sâu như LSTM, GRU, DNN, CNN và lai các mô hình để tìm ra mô hình tốt nhất. Tuy nhiên với mỗi mô hình với các tập dữ liệu khác lại đưa ra kết quả chưa đồng nhất, vậy nên trong bài này nhóm nghiên cứu sẽ thực hiện thêm phép chọn, sau khi phân tích các mã cổ phiếu, sẽ hiển thị mô hình nào là tốt nhất, từ đó nhà đầu tư có thể đưa ra quyết định hợp lý. Bên cạnh đó trong kết quả các bài nghiên cứu trước đưa ra rằng khi chạy mô hình GRU gặp phải trường hợp quá khớp. Để tránh tình trạng đó, nhóm nghiên cứu đã áp dụng thêm Dropout và Early Stopping trong quá trình huấn luyện mô hình giúp cải thiện độ chính xác và tăng tính ổn định của các mô hình dự báo.

Ngoài ra, việc sử dụng PySpark trong xử lý dữ liệu lớn kết hợp với các mô hình học sâu như LSTM, CNN, DNN và GRU đang được quan tâm trong cộng đồng nghiên cứu và thực tiễn tại Việt Nam. Tuy nhiên, các nghiên cứu cụ thể về việc áp dụng PySpark trong dự báo giá cổ phiếu tại Việt Nam hiện còn hạn chế. Vì thế trong bài sử dụng PySpark để xử lý dữ liệu lớn trong dự báo giá cổ phiếu giúp tăng tốc độ tính toán, cải thiện hiệu suất và tính mở rộng của mô hình khi xử lý các tập dữ liệu lớn.

3. Phương pháp nghiên cứu

3.1 Quy trình và phương pháp nghiên cứu

Nghiên cứu này được tiến hành theo một quy trình gồm các bước: thu thập dữ liệu, xử lý và phân tích dữ liệu bằng Pyspark, xây dựng và huấn luyện các mô hình học sâu (LSTM, GRU, CNN và DNN) để dự báo và cuối cùng là đánh giá hiệu quả của các mô hình dự báo

bằng các chỉ số RMSE, MAE, R^2 và MAPE để xác định mô hình tối ưu cho dự báo giá cổ phiếu trên thị trường Việt Nam.

3.2 Thu thập dữ liệu

Dữ liệu trong bài nghiên cứu là dữ liệu giá cổ phiếu lịch sử, thông tin về các mã cổ phiếu và các chỉ số liên quan được lấy từ nền tảng Vnstock, được thu thập trong thời gian 4 năm (2020 - 2024). Mỗi tập dữ liệu từ Vnstock chứa các thông tin: Ngày, giá mở cửa, giá cao, giá thấp, giá đóng cửa và khối lượng giao dịch. Việc thu thập dữ liệu từ Vnstock đảm bảo tính cập nhật và độ tin cậy, phù hợp với mục tiêu nghiên cứu về dự báo giá cổ phiếu tại Việt Nam.

3.3 Xử lý dữ liệu bằng Pyspark

Pyspark có cấu trúc xử lý dữ liệu lớn hiệu quả, cho phép nghiên cứu này xử lý một lượng lớn dữ liệu chứng khoán với tốc độ cao và khả năng chịu lỗi. Đặc biệt, nhóm nghiên cứu sử dụng RDD là một cấu trúc dữ liệu cơ bản của Spark để thực hiện các bước tiền xử lý dữ liệu bao gồm thu thập thông tin mã chứng khoán, xử lý dữ liệu giá lịch sử và thực hiện mô phỏng Monte Carlo.

3.3.1 Dữ liệu giá lịch sử

Quá trình thu thập và tiền xử lý dữ liệu giá lịch sử của các mã cổ phiếu từ Vnstock được thực hiện qua 03 bước chính. Đầu tiên, dữ liệu giá cổ phiếu được tải về và lưu trữ dưới dạng RDD trong Pyspark, cho phép phân phối và xử lý dữ liệu trên nhiều nút trong cụm máy tính để tăng hiệu quả.

Cuối cùng, dữ liệu được chuẩn hóa và chuyển đổi thành các định dạng phù hợp với mô hình dự báo, đảm bảo rằng các giá trị được chuẩn hóa đồng đều, sẵn sàng cho bước phân tích và huấn luyện mô hình tiếp theo. Quy trình này giúp đảm bảo dữ liệu giá cổ phiếu lớn có độ tin cậy cao, tối ưu cho việc xử lý và dự báo với các mô hình học máy.

3.3.2 Thông tin mã chứng khoán

Ngoài dữ liệu giá lịch sử, nghiên cứu còn thu thập thông tin cơ bản về mã chứng khoán như tên công ty, ngành nghề, và các chỉ số tài chính bổ sung. Các dữ liệu này được xử lý thông qua RDD để tạo thành một tập thông tin tổng hợp về mã cổ phiếu, hỗ trợ cho quá trình

phân tích và mô phỏng. Bằng cách ánh xạ các thuộc tính của mã chứng khoán vào từng RDD tương ứng, dữ liệu được tổ chức thành từng phần riêng biệt, giúp giảm thời gian xử lý khi kết hợp với các tập dữ liệu giá lịch sử.

3.3.3 Mô phỏng Monte Carlo

Mô phỏng Monte Carlo là một phần quan trọng trong quy trình xử lý dữ liệu, cho phép dự báo biến động giá cổ phiếu trong tương lai thông qua các kịch bản giả định được tạo dựa trên xác suất. Đầu tiên, dữ liệu giá đóng cửa của từng mã cổ phiếu được trích xuất từ DataFrame Spark và chuyển thành RDD để chuẩn bị cho mô phỏng. Với mỗi mã cổ phiếu, mô phỏng sẽ thiết lập các biến ngẫu nhiên bằng cách khởi tạo mức giá cuối cùng của chuỗi dữ liệu giá lịch sử và tạo ra các biến động ngẫu nhiên trong một khoảng thời gian giả định. Các biến động này được lấy từ phân phối chuẩn, giúp mô phỏng được các thay đổi nhỏ trong giá cổ phiếu, dựa trên một giả định đơn giản về phân phối biến động giá theo thời gian.

Quy trình mô phỏng giá cổ phiếu gồm 1.000 lần thử nghiệm, mỗi thử nghiệm chạy trong 10 ngày tương lai. Từng kịch bản mô phỏng bao gồm một chuỗi giá được xây dựng dựa trên giá cuối cùng hiện có, sau đó kết hợp với các biến ngẫu nhiên để tạo ra mức giá mới cho mỗi ngày. Nhờ sử dụng RDD với các phép toán *map* và *groupByKey*, quá trình mô phỏng được thực hiện đồng thời trên nhiều mã cổ phiếu, giúp tăng hiệu quả tính toán nhờ khả năng phân tán của Spark.

4. Kết quả và đánh giá

4.1 Kết quả sau xử lý dữ liệu

Sau khi phân tích và xử lý dữ liệu lịch sử giá cổ phiếu của 20 mã chứng khoán được thu thập trong khoảng thời gian 4 năm, dữ liệu được chuyển đổi thành Spark DataFrame để dễ dàng thao tác và xử lý. Bảng dữ liệu được bổ sung thêm thông tin về sàn giao dịch (exchange), tên viết tắt công ty (organ_short_name), tên đầy đủ công ty (organ_name) và ngành của công ty theo phân loại ICB mức độ 2 (icb_name2). Nhóm nghiên cứu đã thu được các kết quả sau:

Bảng 1: Dữ liệu lịch sử giá cổ phiếu của 20 mã chứng khoán

index	symbol	exchange	organ_short_name	organ_name	icb_name2
1	VCB	HSX	Vietcombank	Ngân hàng Thương mại Cổ phần Ngoại thương Việt Nam	Ngân hàng
2	POW	HSX	Điện lực Dầu khí Việt Nam	Tổng Công ty Điện lực Dầu khí Việt Nam - CTCP	Điện, nước & xăng dầu khí đốt
3	GVR	HSX	Tập đoàn CN Cao su VN	Tập đoàn Công nghiệp Cao su Việt Nam - Công ty Cổ phần	Hóa chất
4	DHG	HSX	Dược Hậu Giang	Công ty Cổ phần Dược Hậu Giang	Y tế
5	DGC	HSX	Hóa chất Đức Giang	Công ty Cổ phần Tập đoàn Hóa chất Đức Giang	Hóa chất
6	CTD	HSX	Xây dựng Coteccons	Công ty Cổ phần Xây dựng Coteccons	Xây dựng và Vật liệu
7	VNM	HSX	VINAMILK	Công ty Cổ phần Sữa Việt Nam	Thực phẩm và đồ uống
8	SSI	HSX	Chứng khoán SSI	Công ty Cổ phần Chứng khoán SSI	Dịch vụ tài chính
9	MSN	HSX	Tập đoàn Masan	Công ty Cổ phần Tập đoàn Masan	Thực phẩm và đồ uống

10	KBC	HSX	TCT Đô thị Kinh Bắc	Tổng Công ty Phát triển Đô thị Kinh Bắc	Bất động sản
11	GAS	HSX	PV Gas	Tổng Công ty Khí Việt Nam - Công ty Cổ phần	Điện, nước & xăng dầu khí đốt
12	FPT	HSX	FPT Corp	Công ty Cổ phần FPT	Công nghệ Thông tin
13	CMG	HSX	Tập đoàn Công nghệ CMC	Công ty Cổ phần Tập đoàn Công nghệ CMC	Công nghệ Thông tin
14	VIC	HSX	VinGroup	Tập đoàn Vingroup - Công ty CP	Bất động sản
15	REE	HSX	Cơ Điện Lạnh REE	Công ty Cổ phần Cơ điện Lạnh	Điện, nước & xăng dầu khí đốt
16	HPG	HSX	Hòa Phát	Công ty Cổ phần Tập đoàn Hòa Phát	Tài nguyên Cơ bản
17	GMD	HSX	Gemadept	Công ty Cổ phần Gemadept	Hàng & Dịch vụ Công nghiệp
18	VRE	HSX	Vincom Retail	Công ty Cổ phần Vincom Retail	Bất động sản
19	VHC	HSX	Thủy sản Vĩnh Hoàn	Công ty Cổ phần Vĩnh Hoàn	Thực phẩm và đồ uống
20	MWG	HSX	Thế giới di động	Công ty Cổ phần Đầu tư Thế Giới Di Động	Bán lẻ

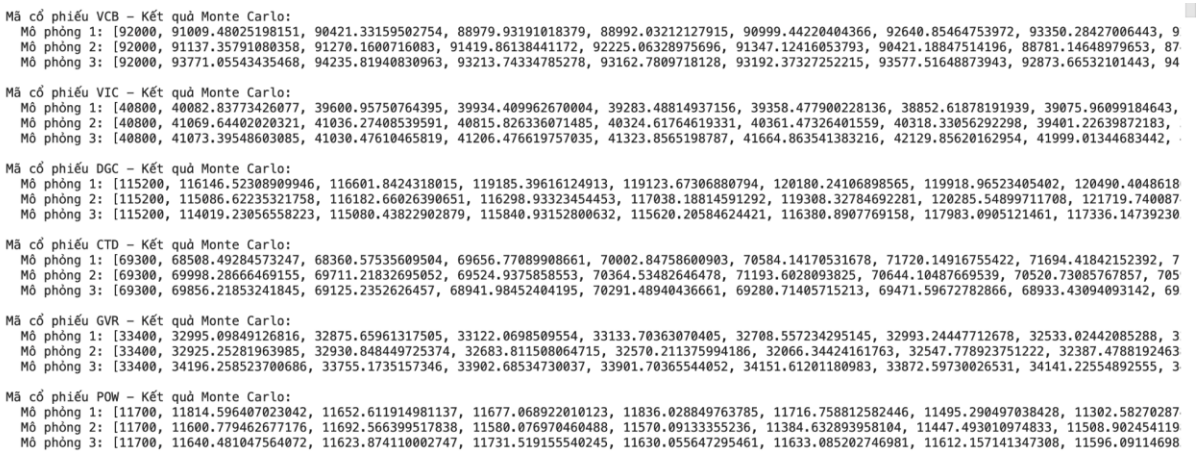
Dữ liệu được thu thập từ API của VNStock. Tiến hành làm sạch và chuẩn hóa dữ liệu, đảm bảo không có giá trị thiếu sót. Bao gồm các cột dữ liệu Date, Open, Close, High, Low, Volume, Ticker.

Bảng 2: Bảng dữ liệu trích từ API và được làm sạch

	date	open	close	high	low	volume	ticker
0	2020-11-13	56330	56530	56530	56070	417840	VCB
1	2020-11-16	56720	55810	57110	55290	1250340	VCB
2	2020-11-17	55810	56980	55980	55620	548250	VCB
3	2020-11-18	56980	56980	57050	56590	906810	VCB
4	2020-11-19	57050	58680	58680	56850	2584440	VCB
…	…	…	…	…	…		…
19957	2024-11-05	17600	17750	17850	17600	2633700	VRE
19956	2024-11-06	17900	18050	18050	17800	2971100	VRE
19958	2024-11-07	18200	18050	18050	17800	2971100	VRE
19959	2024-11-08	18150	17950	17950	18150	4487200	VRE
19960	2024-11-11	17800	18000	18000	17600	5983800	VRE

4.2 Kết quả mô phỏng Monte Carlo:

Mô phỏng Monte Carlo là một phương pháp thống kê được sử dụng để ước lượng các kết quả của một hệ thống phức tạp thông qua việc sử dụng các mẫu ngẫu nhiên. Mô phỏng Monte Carlo được thực hiện với RDD để dự đoán giá cổ phiếu trong 10 ngày tiếp theo. Kết quả mô phỏng bao gồm danh sách các mức giá có thể xảy ra cho mỗi mã cổ phiếu trong 10 ngày tới.



Hình 3: Kết quả Monte Carlo sau khi xử lý với RDD

Mỗi mô phỏng cho thấy khả năng giá cổ phiếu tăng hoặc giảm sau 10 ngày. Dựa trên kết quả mô phỏng, xác suất giá tăng (increase_prob) được tính toán và khuyến nghị mua, bán, hoặc giữ (recommendation) được đưa ra cho từng mã cổ phiếu. Vì thế có thể dựa vào nhiều lần mô phỏng để đánh giá xem có nên mua, bán hay giữ cổ phiếu

Bảng 3: Kết quả mô phỏng Monte Carlo và khuyến nghị

ticker	avg_close	increase_prob	recommendation
CMG	34579.55	0.482	Hold
CTD	52893.71	0.505	Hold
DGC	72121.04	0.49	Hold
DHG	94291.39	0.482	Hold
FPT	70041.18	0.505	Hold
GAS	72216.43	0.477	Hold
GMD	44556.35	0.489	Hold
GVR	25766.42	0.481	Hold
HPG	25532.89	0.512	Hold
KBC	29562.41	0.498	Hold
MSN	89319.94	0.481	Hold
MWG	53357.55	0.503	Hold
POW	12818.13	0.477	Hold
REE	50008.6	0.498	Hold

SSI	22116.38	0.49	Hold
VCB	75369.1	0.466	Hold
VHC	55506.76	0.512	Hold
VIC	68439.24	0.503	Hold
VNM	70793.44	0.466	Hold
VRE	27551.2	0.489	Hold

Xác suất tăng giá dao động giữa 0.46 và 0.51 cho thấy các cổ phiếu như VCB và FPT mặc dù có giá trung bình cao nhưng xác suất tăng giá không quá cao, cho thấy sự ổn định nhưng không có nhiều cơ hội tăng trưởng lớn trong ngắn hạn. Mô phỏng Monte Carlo cung cấp cái nhìn sâu sắc về khả năng tăng trưởng của từng mã cổ phiếu, với một số mã có xác suất tăng giá vượt quá 70%, được khuyến nghị mua. Các khuyến nghị đầu tư được đưa ra dựa trên xác suất tăng giá giúp các nhà đầu tư có cơ sở để ra quyết định.

4.3 Biểu đồ các chỉ báo kỹ thuật

Sau khi hoàn tất, sẽ có các biểu đồ kỹ thuật thể hiện các chỉ báo mà người dùng cần.Các chỉ báo kỹ thuật như MA50, MA100, Bollinger Bands, MACD, và RSI được tính toán dựa trên dữ liệu giá cổ phiếu.

Biểu đồ 1: Biểu đồ phân tích kỹ thuật của VCB (1)

Biểu đồ 2: Biểu đồ phân tích kỹ thuật của VCB (2)

4.4 Kết quả dự đoán từ các mô hình LSTM, GRU, CNN và DNN

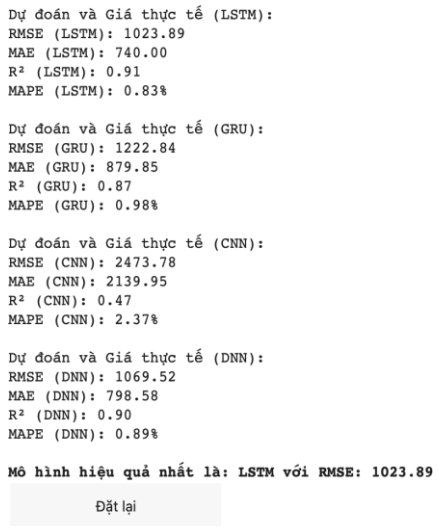
Kết quả thu được của từng mô hình tương đối sát với giá thực tế, đây là một thành công của nhóm khi tất cả mô hình được huấn luyện đầy đủ và đủ nhiều để có thể đề xuất ra giá dự đoán. Mô hình LSTM, GRU và CNN đều cho kết quả dự đoán khả quan, với các chỉ số RMSE, MAE, R^2 và MAPE đạt mức chấp nhận được.

Biểu đồ 3: Biểu đồ dự đoán giá cổ phiếu VCB

Qua việc so sánh các mô hình dự đoán, ta nhận thấy rằng mô hình LSTM có độ chính xác ổn định và tiệm cận nhất với giá thực tế. Đường dự báo của LSTM gần như song song với đường giá thực tế, cho thấy khả năng dự đoán tương đối tốt. Trong khi đó, mô hình GRU cũng có độ chính xác khá cao, chỉ kém LSTM một chút.

Ngược lại, mô hình CNN lại có độ chính xác thấp hơn, thể hiện qua sự biến động mạnh và xa rời so với diễn biến thực tế của giá cổ phiếu. Mô hình DNN cũng không ổn định bằng LSTM và GRU, mặc dù vẫn nằm trong mức chấp nhận được.

Xét về diễn biến giá cổ phiếu VCB, ta nhận thấy xu hướng tăng khá rõ ràng, đặc biệt từ tháng 9/2024 trở về sau. Tuy nhiên, thị trường vẫn chịu nhiều biến động mạnh, thể hiện qua những đỉnh và đáy liên tiếp. Điều này cho thấy việc dự đoán giá cổ phiếu vẫn còn nhiều thách thức, đòi hỏi các nhà đầu tư phải thận trọng.

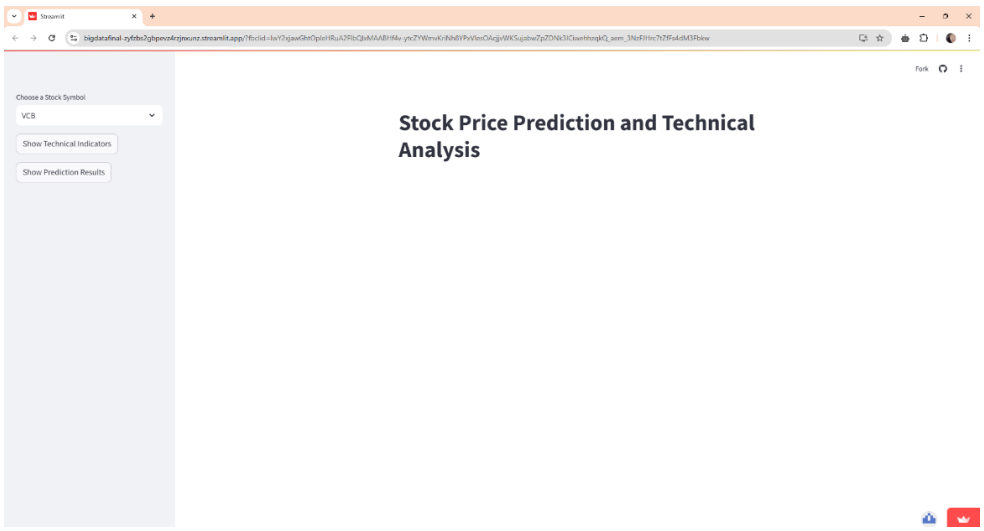


Hình 4: Kết quả dự đoán giá

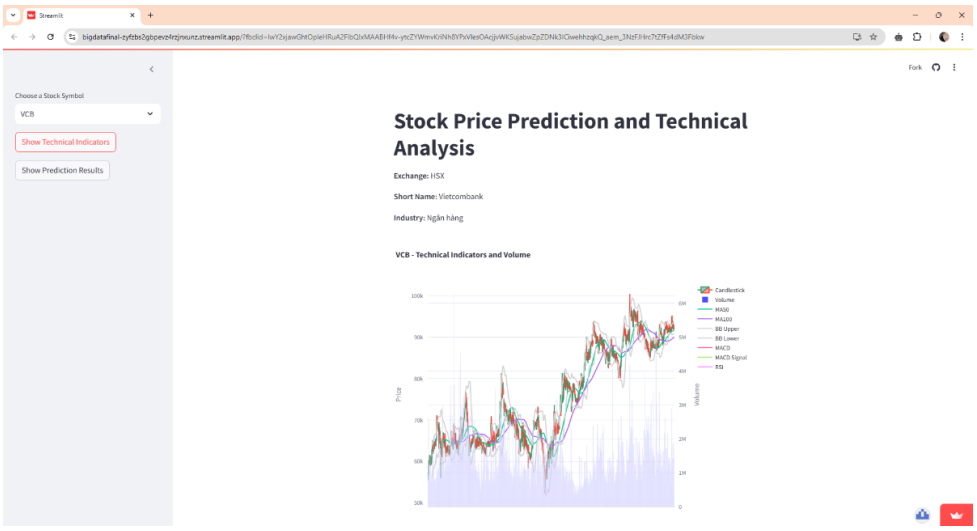
Dựa trên kết quả phân tích, có thể khuyến nghị việc kết hợp sử dụng các mô hình, đặc biệt là LSTM và GRU, để có cái nhìn toàn diện hơn về diễn biến giá cổ phiếu VCB trong tương lai. Điều này sẽ giúp các nhà đầu tư đưa ra quyết định đầu tư hiệu quả hơn. Mô hình có RMSE, MAE và MAPE thấp, cùng với R^2 cao sẽ được coi là mô hình có hiệu suất tốt nhất.

4.5 Ứng dụng web

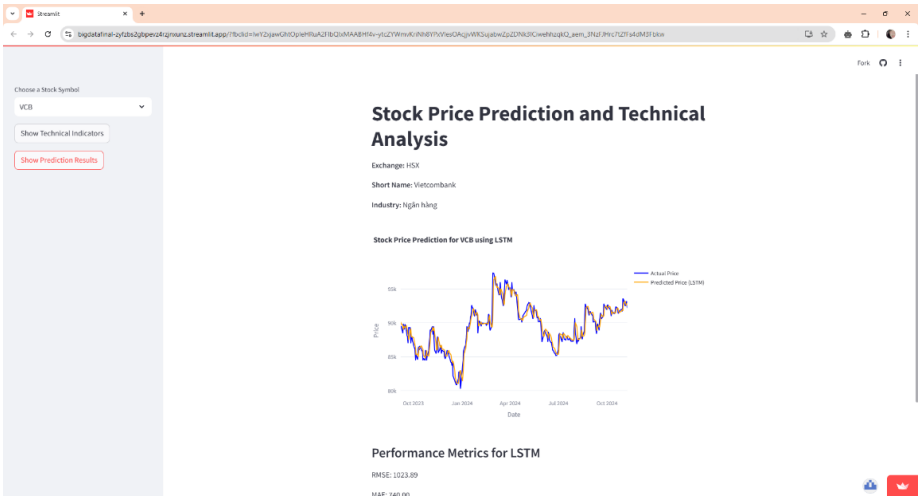
Ứng dụng web được xây dựng bằng Streamlit để hiển thị kết quả dự đoán và chỉ báo kỹ thuật một cách trực quan và thân thiện với người dùng. Người dùng có thể chọn mã cổ phiếu, xem biểu đồ kỹ thuật và biểu đồ dự đoán giá cổ phiếu từ các mô hình. Dưới đây là giao diện trang web:



Hình 5: Giao diện trang web ban đầu



Hình 6: Giao diện trang web với phân tích kỹ thuật



Hình 7: Giao diện trang web với mô hình dự đoán giá

5. Kết luận và khuyến nghị

5.1 Kết luận

Các kết quả cho thấy mô hình đã có thể đưa ra các khuyến nghị đầu tư với độ chính xác cao, giúp các nhà đầu tư có cơ sở để quyết định "Mua", "Bán", hoặc "Giữ". Bên cạnh đó, việc áp dụng mô phỏng Monte Carlo đã mở ra những cái nhìn mới mẻ về khả năng tăng trưởng và rủi ro của các mã cổ phiếu, tạo ra một bức tranh toàn diện hơn cho các quyết định đầu tư. Các mô hình cũng đã được xây dựng và phát triển, cung cấp các mô phỏng về giá cho nhà đầu tư một cách trực quan và hiệu quả nhất về thị trường qua giao diện thân thiện và dễ thao tác.

Nghiên cứu đã đạt được những kết quả quan trọng trong việc xây dựng một hệ thống dự đoán giá cổ phiếu hiệu quả dựa trên bốn mô hình học sâu khác nhau: LSTM, GRU, CNN và DNN, cho phép dự đoán xu hướng giá trong tương lai bằng cách sử dụng dữ liệu lịch sử. Nghiên cứu cũng đã cung cấp các chỉ báo kỹ thuật quan trọng như MA50, MA100, Bollinger Bands, MACD và RSI. Những chỉ báo này được trực quan hóa dưới dạng biểu đồ nến Nhật, giúp nhà đầu tư dễ dàng theo dõi và phân tích, từ đó đưa ra quyết định đầu tư hợp lý hơn. Ứng dụng web được xây dựng bằng Streamlit thân thiện với người dùng khi người dùng có thể dễ dàng tương tác với hệ thống, lựa chọn các mã cổ phiếu, và xem kết quả dự đoán cùng với các chỉ báo kỹ thuật, giúp họ có cái nhìn tổng quan và chính xác hơn về tình hình thị trường.

Các mô hình dự đoán hiện tại chủ yếu dựa vào dữ liệu lịch sử về giá cổ phiếu mà chưa xem xét một số yếu tố bên ngoài quan trọng. Tác động của các yếu tố kinh tế vĩ mô, chính trị và tâm lý thị trường cần được đánh giá kỹ lưỡng hơn, vì chúng có thể gây ra những biến động đáng kể trong giá cổ phiếu. Việc không tích hợp những yếu tố này vào các mô hình dự đoán có thể hạn chế khả năng chính xác của chúng. Bên cạnh đó, hiện tại chưa có thông tin về việc tối ưu hóa các siêu tham số của mô hình hay việc áp dụng các phương pháp học tập kết hợp. Những cải tiến này có thể giúp nâng cao hiệu suất của mô hình dự đoán một cách đáng kể. Phạm vi ứng dụng của hệ thống hiện tại vẫn còn hạn chế, chỉ tập trung vào việc dự đoán giá cổ phiếu. Việc mở rộng ứng dụng để bao gồm các chỉ số tài chính khác hoặc phân tích rủi ro có thể mang lại giá trị gia tăng cho người dùng và giúp họ đưa ra quyết định đầu tư tốt hơn.

5.3 Khuyến nghị

Trong tương lai, nghiên cứu có thể mở rộng bằng cách kết hợp các yếu tố vĩ mô, tin tức thị trường và tâm lý của nhà đầu tư thông qua phân tích cảm xúc. Việc tích hợp những yếu tố này vào mô hình dự đoán sẽ giúp nâng cao độ chính xác trong việc dự báo xu hướng giá cổ phiếu. Một hướng nghiên cứu khác có thể là tinh chỉnh các siêu tham số của mô hình, thử nghiệm với các kiến trúc mô hình đa dạng, và áp dụng các phương pháp học tập kết hợp nhằm tăng cường hiệu suất dự đoán. Những điều này có thể tạo ra những cải tiến đáng kể trong khả năng dự đoán giá cổ phiếu. Cuối cùng, nghiên cứu cũng có thể được mở rộng để không chỉ dự đoán giá cổ phiếu mà còn phân tích các chỉ số thị trường khác. Việc này sẽ cung cấp cái nhìn sâu sắc về rủi ro trong danh mục đầu tư và hỗ trợ phát triển các hệ thống giao dịch tự động, giúp các nhà đầu tư đưa ra quyết định chính xác hơn trong môi trường tài chính đầy biến động.

Danh mục tài liệu tham khảo

Biondo, A. E., Pluchino, A., Rapisarda, A., & Helbing, D. J. P. o. (2013). Are random trading strategies more successful than technical ones? , *8*(7), e68344.

Boyle, P. P. J. J. o. f. e. (1977). Options: A monte carlo approach. *4*(3), 323-338.

Brock, W., Lakonishok, J., & LeBaron, B. J. T. J. o. f. (1992). Simple technical trading rules and the stochastic properties of stock returns. *47*(5), 1731-1764.

Burton, N. (2018). *An analysis of Burton G. Malkiel's A random walk down Wall Street*: Macat Library.

Cang, P. T., Quyên, T. T. T., & Ngoan, T. T. J. T. c. K. h. Đ. h. C. T. (2024). Đánh giá các thuật toán lọc hiệu quả trong xử lý dữ liệu lớn. *60*(5), 59-68.

Carlo, Q.-M. (2001). Monte Carlo methods in financial engineering.

Chan, J. C., & Kroese, D. P. J. A. o. O. R. (2011). Rare-event probability estimation with conditional Monte Carlo. *189*, 43-61.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. J. a. p. a. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Dahlquist, J. R., & Kirkpatrick II, C. D. (2010). *Technical analysis: the complete resource for financial market technicians*: FT press.

Dey, P., Hossain, E., Hossain, M. I., Chowdhury, M. A., Alam, M. S., Hossain, M. S., & Andersson, K. J. A. (2021). Comparative analysis of recurrent neural networks in stock price prediction for different frequency domains. *14*(8), 251.

Elman, J. L. J. C. s. (1990). Finding structure in time. *14*(2), 179-211.

Fama, E. F. J. J. o. f. (1970). Efficient capital markets. *25*(2), 383-417.

Fifield, S. G., Power, D. M., & Donald Sinclair, C. J. T. E. J. o. F. (2005). An analysis of trading strategies in eleven European stock markets. *11*(6), 531-548.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised Learning Techniques*: O'Reilly Media, Incorporated.

Graves, A., & Graves, A. J. S. s. l. w. r. n. n. (2012). Long short-term memory. 37-45.

Graves, A., & Schmidhuber, J. J. A. i. n. i. p. s. (2008). Offline handwriting recognition with multidimensional recurrent neural networks. *21*.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., Schmidhuber, J. J. I. t. o. n. n., & systems, l. (2016). LSTM: A search space odyssey. *28*(10), 2222-2232.

Hochreiter, S. J. N. C. M.-P. (1997). Long Short-term Memory.

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). *An empirical exploration of recurrent network architectures.* Paper presented at the International conference on machine learning.

Kanchanamala, P., Karnati, R., Bhaskar Reddy, P. V. J. C., Practice, C., & Experience. (2023). Hybrid optimization enabled deep learning and spark architecture using big data analytics for stock market forecasting. *35*(8), e7618.

LeCun, Y., Bengio, Y., & Hinton, G. J. n. (2015). Deep learning. *521*(7553), 436-444.

Lo, A. W., & MacKinlay, A. C. (2011). *A non-random walk down Wall Street*: Princeton University Press.

Mehtab, S., & Sen, J. (2020). *Stock price prediction using CNN and LSTM-based deep learning models.* Paper presented at the 2020 International Conference on Decision Aid Sciences and Application (DASA).

Nguyen, M. T., Lin, W. W., Huang, J. H. J. C., Systems,, & Processing, S. (2023). Heart sound classification using deep learning techniques based on log-mel spectrogram. *42*(1), 344-360.

Oanh, Đ. L. K., & Châu, N. T. M. (2024). Dự báo chỉ số chứng khoán bằng học máy: Bằng chứng thực nghiệm từ thị trường chứng khoán Việt Nam.

Phú, N. A. (7/2024). Tác động của các yếu tố kinh tế vĩ mô đến thị trường chứng khoán Việt Nam. *Tạp chí Tài chính*.

Phuoc, T., Anh, P. T. K., Tam, P. H., & Nguyen, C. V. (2024). Applying machine learning algorithms to predict the stock price trend in the stock market–The case of Vietnam. *11*(1), 1-18.

Shah, D., Campbell, W., & Zulkernine, F. H. (2018). *A comparative study of LSTM and DNN for stock market forecasting.* Paper presented at the 2018 IEEE international conference on big data (big data).

Sisodia, J., & Boghey, R. J. J. o. A. A. I. (2024). An improved index price/movement prediction by using ensemble cnn and dnn deep learning technique. *5*(1), 41-53.

Song, H., & Choi, H. J. A. S. (2023). Forecasting stock market indices using the recurrent neural network based hybrid models: CNN-LSTM, GRU-CNN, and ensemble models. *13*(7), 4644.

Tuyên, T. Đ. J. T. c. K. h. Đ. h. c. T. (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. *60*(1), 235-249.

Trịnh, N. P., Trần, N. A. K., Hà, V. L., Trần, T. H., & Đỗ, T. H. (2022). Dự đoán giá cổ phiếu sử dụng mô hình chuỗi thời gian với BigDL.

Zhang, J., Ye, L., & Lai, Y. J. M. (2023). Stock price prediction using CNN-BiLSTM-Attention model. *11*(9), 1985.