



BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM 2024

**A HYBRID APPROACH COMBINING CLUSTERING AND
DEEP LEARNING FOR STOCK PRICE PREDICTION**

Chuyên ngành: Hệ thống thông tin

Nhóm nghiên cứu:

TT	Họ tên	MSSV	Đơn vị	Nhiệm vụ	Điện thoại	Email
1.	Nguyễn Khánh Linh	K214140942	Khoa Tài chính - Ngân hàng	Nhóm trưởng	0822925446	linhknk21414c@st.uel.edu.vn
2.	Võ Hưng Thanh	K214141338	Khoa Tài chính - Ngân hàng	Tham gia	0363206520	thanhvh21414c@st.uel.edu.vn
3.	Nguyễn Thủy Tiên	K214142087	Khoa Tài chính - Ngân hàng	Tham gia	0862842148	tiennt21414@st.uel.edu.vn
4.	Phạm Công Nguyễn Khôi	K214160990	Khoa Hệ thống thông tin	Tham gia	0858833863	khoipcn21416c@st.uel.edu.vn
5.	Trần Thị Vân Anh	K214061734	Khoa Hệ thống thông tin	Tham gia	0798079138	anhhtt21406@st.uel.edu.vn

ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT



BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN 2024

**A HYBRID APPROACH COMBINING CLUSTERING
AND DEEP LEARNING FOR STOCK PRICE
PREDICTION**

Đại diện nhóm nghiên cứu
(Ký, họ tên)

Nguyễn Khánh Linh

Giảng viên hướng dẫn
(Ký, họ tên)

Nguyễn Thôn Dã

Chủ tịch hội đồng
(Ký, họ tên)

Lãnh đạo Khoa/ Bộ môn/ Trung tâm
(Ký, họ tên)

Thành phố Hồ Chí Minh, tháng 03 năm 2024

LỜI CẢM ƠN

Nhóm nghiên cứu xin gửi lời cảm ơn sâu sắc đến Giảng viên hướng dẫn trực tiếp là TS. Nguyễn Thôn Dã đã tận tình hướng dẫn, truyền đạt kiến thức và kinh nghiệm cho nhóm trong suốt quá trình thực hiện bài nghiên cứu này. Nhờ sự chỉ dẫn cùng lượng kiến thức quý báu thầy đã truyền đạt mà chúng tôi có thể vượt qua khó khăn trong quá trình nghiên cứu và hoàn thành bài luận.

Nhóm nghiên cứu cũng muốn gửi lời cảm ơn chân thành đến các thầy cô Khoa Hệ thống thông tin, Trường Đại học Kinh tế - Luật, ĐHQG-HCM đã chia sẻ kiến thức và giúp đỡ chúng tôi trong quá trình nghiên cứu.

Do về mặt kiến thức còn hạn chế, bài nghiên cứu vẫn còn nhiều khiếm khuyết dù đã cố gắng hết sức. Nhóm nghiên cứu kính mong Quý thầy cô, các chuyên gia, những người quan tâm đến đề tài tiếp tục có những ý kiến đóng góp để đề tài được hoàn thiện hơn. Một lần nữa, chúng tôi xin chân thành cảm ơn sự giúp đỡ và hỗ trợ của Quý thầy cô

Xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, ngày 22 tháng 3 năm 2024

Nhóm nghiên cứu

LỜI CAM KẾT

Nhóm nghiên cứu xin cam đoan đề tài “*A hybrid approach combining clustering and deep learning for stock price prediction*” là công trình do nhóm tự nghiên cứu, đọc, dịch các tài liệu liên quan, tổng hợp và thực hiện dưới sự hướng dẫn của TS. Nguyễn Thôn Dã.

Ngoại trừ các kết quả tham khảo, các trích dẫn từ các công trình nghiên cứu khác đã được ghi rõ nguồn gốc thì nghiên cứu này của chúng tôi chưa từng được công bố trong bất kỳ một công trình nào trước đây. Nhóm nghiên cứu cam đoan rằng chúng tôi đã thực hiện đúng nguyên tắc đạo đức trong nghiên cứu khoa học và tuân thủ đầy đủ các quy định đã được đặt ra trong nguyên tắc này.

Thành phố Hồ Chí Minh, ngày 22 tháng 3 năm 2024

Nhóm nghiên cứu

MỤC LỤC

DANH MỤC BẢNG	1
DANH MỤC HÌNH ẢNH	2
CHƯƠNG 1: GIỚI THIỆU VẤN ĐỀ NGHIÊN CỨU	3
CHƯƠNG 2: TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU.....	5
2.1. Bối cảnh nghiên cứu	5
2.2. Cơ sở lý luận.....	6
2.2.1. K-means clustering	6
2.2.2. Các chỉ số tài chính	9
2.2.3. Deep Learning.....	13
2.3. Thực trạng vấn đề nghiên cứu: Khái quát các kết quả nghiên cứu đã đạt được .	23
2.3.1. Trên thế giới	23
2.3.2. Tại Việt Nam	27
CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU.....	29
3.1. Mẫu nghiên cứu	29
3.2. Quy trình nghiên cứu	29
3.3. Phương pháp xử lý thông tin	31
3.3.1. Phương pháp thu thập dữ liệu	31
3.3.2. Phương pháp phân tích dữ liệu	31
CHƯƠNG 4. KẾT QUẢ VÀ ĐÁNH GIÁ	32
4.1. Thu thập và tính toán dữ liệu	32
4.2. Phân cụm	32
4.3. Đánh giá mô hình.....	36
4.4. Dự đoán	39
CHƯƠNG 5. KẾT LUẬN VÀ ĐỀ XUẤT	41
5.1. Kết luận	41
5.2. Hạn chế và đề xuất hướng phát triển	41
5.2.1. Hạn chế	41
5.2.2. Hướng phát triển	42
TÀI LIỆU THAM KHẢO.....	44

DANH MỤC BẢNG

Bảng 1. Ưu và nhược điểm của K-Means Clustering.....	9
Bảng 2. Danh sách 20 nhóm ngành được lựa chọn nghiên cứu	29
Bảng 3. Kết quả tính toán	32
Bảng 4. Phân tích và dự đoán cụm cho mỗi công ty	34
Bảng 5. Chi tiết 5 cụm K- Means	36
Bảng 6. Dữ liệu được lấy từ thư viện Vnstock của công ty GMC	37
Bảng 7. Bảng chỉ số hiệu suất đánh giá mô hình	38

DANH MỤC HÌNH ẢNH

Hình 1. Sơ đồ của K-Means	8
Hình 2. Nguyên lý cơ bản của Autoencoder	16
Hình 3. Cấu trúc cơ bản của Restricted Boltzmann Machine	17
Hình 4. Cấu trúc của Boosted Deep Belief Network được tăng cường bao gồm nhiều Deep belief networks. Chỉ những cái trong hộp chấm sẽ được tinh chỉnh chung.	18
Hình 5. Mô hình sử dụng VGG-16 làm mạng cơ sở. Các lớp đối tượng bổ sung được thêm vào mạng cơ sở để có thể sử dụng các lớp đối tượng để đưa ra dự đoán phát hiện của mọi bản đồ đối tượng.	21
Hình 6. Mô hình cây của Deep Neural Decision Forests bao gồm một số cây quyết định. Các nút quyết định quyết định cách các mẫu đi qua cây. Các nút dự đoán tạo ra xác suất quyết định sau khi các mẫu đến các nút cuối cùng tương ứng.	21
Hình 7. Kiến trúc mạng bộ nhớ ngắn hạn dài 2D. Ảnh đầu vào I_k được chia thành một số cửa sổ. Mỗi cửa sổ được đưa vào bốn khối bộ nhớ LSTM riêng biệt. Đầu ra của mỗi khối LSTM được chuyển đến lớp tiếp liệu. Ở lớp cuối cùng, đầu ra của các khối LSTM cuối cùng được tổng hợp và gửi đến lớp softmax. Cuối cùng, các mạng đưa ra xác suất của lớp cho từng cửa sổ đầu vào	21
Hình 8. Bốn giai đoạn của phương pháp nghiên cứu	30
Hình 9. Kết quả phân cụm	33
Hình 10. Sơ đồ 3D biểu thị 5 cụm K-Means	34
Hình 11. Cụm 0	39
Hình 12. Cụm 1	39
Hình 13. Cụm 2	40
Hình 14. Cụm 3	40
Hình 15. Cụm 4	40

CHƯƠNG 1: GIỚI THIỆU VẤN ĐỀ NGHIÊN CỨU

Trong thị trường chứng khoán, giá cổ phiếu là một chỉ số quan trọng phản ánh giá trị của doanh nghiệp và kỳ vọng của nhà đầu tư. Việc dự đoán giá cổ phiếu là một bài toán hấp dẫn và đầy thách thức, vì nó liên quan đến nhiều yếu tố bên trong và bên ngoài doanh nghiệp, cũng như sự biến động và phi tuyến tính của dữ liệu. Nhiều nghiên cứu đã được tiến hành để áp dụng các phương pháp thống kê, toán học và máy học để giải quyết bài toán này.

Các bài nghiên cứu về việc dự đoán giá cổ phiếu đã xuất hiện từ rất lâu và đã đạt được một số kết quả ấn tượng. Có rất nhiều thuật dự đoán giá cổ phiếu là rất phong phú và đa dạng. Có thể phân loại các nghiên cứu này theo các tiêu chí như phương pháp dự đoán, loại dữ liệu đầu vào, thời gian dự đoán, v.v. Một trong những thách thức lớn là tính không ổn định và phức tạp của dữ liệu giá cổ phiếu. Các mô hình dự đoán truyền thống có thể không thể hiện được những biến động và mối quan hệ phức tạp trong dữ liệu này, dẫn đến kết quả dự đoán không chính xác.

Để vượt qua các hạn chế này, trong nghiên cứu này, nhóm nghiên cứu đề xuất một phương pháp kết hợp giữa phân cụm và deep learning, gọi là "A hybrid approach combining clustering and deep learning for stock price prediction." Nhóm nghiên cứu chọn đề tài này là để khai thác tiềm năng của kết hợp phân cụm và deep learning trong bài toán dự đoán giá cổ phiếu, một bài toán có ý nghĩa thực tiễn cao và đòi hỏi sự chính xác và tính hiệu quả. Bằng cách phân cụm các công ty có đặc điểm tài chính tương tự, ta có thể giảm thiểu sự phức tạp và nhiễu của dữ liệu, và tăng khả năng dự đoán các xu hướng và biến động của giá cổ phiếu. Bằng cách sử dụng Deep Learning, ta có thể học được các đặc trưng phức tạp và trừu tượng của dữ liệu, và xấp xỉ được các hàm phi tuyến tính và động lực học của giá cổ phiếu. Đề tài này mong muốn đóng góp vào việc nâng cao chất lượng và hiệu năng của các mô hình dự đoán giá cổ phiếu, và cung cấp cho nhà đầu tư những thông tin hữu ích và đáng tin cậy để ra quyết định đầu tư.

Đối tượng nghiên cứu của đề tài này là 98 công ty lớn nhỏ đến từ 20 ngành khác nhau được niêm yết trên sàn HoSE. Sàn HoSE là sàn giao dịch chứng khoán lớn nhất Việt Nam, có quy mô vốn hóa thị trường và khối lượng giao dịch cao nhất. Các công ty được chọn là những công ty có dữ liệu đầy đủ và liên tục trong vòng 5 năm từ 2019

đến 2023, và có đại diện cho các ngành kinh tế chủ lực của Việt Nam, như ngân hàng, bất động sản, dầu khí, điện lực, thép, dệt may, thực phẩm và đồ uống, v.v.

Nghiên cứu của đề tài này bao gồm hai bước chính: phân cụm và dự đoán. Trong bước phân cụm, nhóm đã tính trung bình các chỉ số ROE, P/E, P/B của các công ty trong vòng 5 năm từ năm 2019-2023, và sử dụng phương pháp phân cụm K-means để phân chia các công ty thành các nhóm có trung bình các chỉ số gần nhau. Trong bước dự đoán, ta sẽ lấy giá cổ phiếu mỗi ngày của các công ty trên thư viện vnstock, và sử dụng một mô hình Deep Learning kết hợp CNN, LSTM và GRU để dự đoán giá cổ phiếu trong tương lai của các công ty được phân cùng nhóm.

Ngoài ra, đề tài này cũng nhằm đáp ứng nhu cầu của thị trường chứng khoán Việt Nam trong bối cảnh hậu đại dịch Covid-19, khi mà giá xăng dầu trên thị trường thế giới biến động không ngừng, ảnh hưởng đến giá cả trong nước. Dịch Covid-19 là một yếu tố bất ngờ và bất lợi, gây ra những ảnh hưởng tiêu cực đến nền kinh tế và thị trường chứng khoán trên toàn cầu. Theo một nghiên cứu của He và cộng sự (2020), dịch Covid-19 đã làm giảm giá trị thị trường chứng khoán của 26 quốc gia trong khu vực châu Á - Thái Bình Dương trong đó có Việt Nam. Dịch Covid-19 cũng làm giảm niềm tin của nhà đầu tư, tăng biến động và rủi ro của thị trường chứng khoán. Do đó, việc dự đoán giá cổ phiếu trở nên càng khó khăn và cần thiết hơn. Bên cạnh đó, dịch Covid-19 cũng tạo ra những cơ hội và thách thức cho thị trường chứng khoán Việt Nam, đòi hỏi các chính sách và pháp luật về chứng khoán phải được hoàn thiện và linh hoạt hơn. Theo một bài báo của Bộ Tài chính, thị trường chứng khoán Việt Nam đã phục hồi nhanh và tăng trưởng vượt kỳ vọng, đạt được nhiều kỷ lục mới trong năm 2020. Một số ngành như y tế, công nghệ, tiêu dùng, v.v. đã được hưởng lợi từ dịch Covid-19 và thu hút được nhiều dòng tiền đầu tư. Theo một bài viết của Tạp chí Ngân hàng, một số chính sách và pháp luật đã được ban hành để ổn định thị trường chứng khoán trước tác động của dịch Covid-19, như việc giảm phí giao dịch, nới lỏng điều kiện cho vay chứng khoán, tăng hạn mức cho vay bằng cổ phiếu, v.v. Tuy nhiên, vẫn còn nhiều thách thức đặt ra, như việc cải thiện minh bạch thông tin, nâng cao chất lượng quản trị doanh nghiệp, phát triển các sản phẩm và dịch vụ chứng khoán, v.v.

CHƯƠNG 2: TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

2.1. Bối cảnh nghiên cứu

Trong thị trường cổ phiếu Việt Nam, việc dự đoán giá cổ phiếu là một thách thức lớn đối với các nhà đầu tư và các chuyên gia tài chính do ảnh hưởng của nhiều yếu tố phức tạp như biến động thị trường, thông tin không chắc chắn, và ảnh hưởng của các yếu tố kinh tế chính trị. Tại Việt Nam, các nghiên cứu về dự đoán giá cổ phiếu thường tập trung vào các phương pháp truyền thống như phân tích kỹ thuật và phân tích cơ bản. Tuy nhiên, sự phát triển của AI và Machine Learning đã mở ra một lĩnh vực mới cho nghiên cứu về dự đoán giá cổ phiếu, và phương pháp kết hợp giữa phân cụm và Deep Learning là một trong những hướng tiên tiến được quan tâm. Tính đến thời điểm hiện tại, số lượng nghiên cứu trong lĩnh vực này tại Việt Nam vẫn còn hạn chế và hầu hết là ở giai đoạn tiền nghiên cứu.

Việc sử dụng phương pháp phân cụm đã được chứng minh là một công cụ hữu ích để nhận biết các nhóm cổ phiếu có xu hướng di chuyển cùng nhau trong thị trường. Tuy nhiên, đối với thị trường cổ phiếu Việt Nam, việc áp dụng phân cụm đôi khi gặp khó khăn do sự biến động lớn và tính không ổn định của thị trường. Do đó, việc kết hợp phân cụm với Deep Learning có thể tạo ra những mô hình dự đoán mạnh mẽ hơn, bằng cách sử dụng các thuật toán Deep Learning để phân tích dữ liệu cổ phiếu một cách chi tiết và linh hoạt hơn.

Việc kết hợp giữa phân cụm và Deep Learning có thể giúp cải thiện đáng kể khả năng dự đoán giá cổ phiếu. Bằng cách tận dụng sức mạnh của Deep Learning trong việc tự động học và phát hiện các mẫu phức tạp từ dữ liệu, kết hợp với khả năng nhận biết các nhóm cổ phiếu có tính chất tương tự từ phân cụm, các mô hình dự đoán có thể trở nên hiệu quả hơn và chính xác hơn trong việc đưa ra các quyết định đầu tư. Điều này cũng đặt ra thách thức đối với các nhà nghiên cứu, đặc biệt là khi có ít nguồn tài liệu và dữ liệu cụ thể về thị trường cổ phiếu Việt Nam được công bố công khai.

Ngoài ra, việc áp dụng phương pháp kết hợp này cũng có thể giúp giảm thiểu rủi ro và tăng cường lợi nhuận cho các nhà đầu tư trên thị trường cổ phiếu Việt Nam. Bằng cách sử dụng các mô hình dự đoán được tạo ra từ kết hợp phân cụm và Deep Learning, nhà đầu tư có thể có cái nhìn toàn diện hơn về các cơ hội đầu tư và nguy cơ tiềm ẩn, từ đó đưa ra các quyết định đầu tư thông minh và linh hoạt hơn. Mặc dù đã có

một số nghiên cứu nhỏ và dự án thử nghiệm được thực hiện, nhưng vẫn còn rất nhiều công việc phải làm để cải thiện hiệu suất và ứng dụng thực tiễn của phương pháp này trong việc dự đoán giá cổ phiếu tại Việt Nam. Điều này bao gồm việc thu thập và xử lý dữ liệu chất lượng cao, xây dựng các mô hình dự đoán chính xác và đáng tin cậy, cũng như thử nghiệm và đánh giá kết quả trên thực tế.

Tóm lại, nghiên cứu này tại Việt Nam không chỉ là một lĩnh vực đầy tiềm năng mà còn là một lĩnh vực quan trọng đối với việc cải thiện hiệu suất đầu tư và quản lý rủi ro trên thị trường tài chính đang phát triển nhanh chóng này.

2.2. Cơ sở lý luận

2.2.1. K-means clustering

Phân cụm là cách phân loại dữ liệu thô một cách hợp lý và tìm kiếm các mẫu ẩn có thể tồn tại trong bộ dữ liệu (Huan và cộng sự, 1998). Đó là một quá trình nhóm các đối tượng dữ liệu thành các cụm rời rạc sao cho dữ liệu trong cùng một cụm giống nhau, tuy nhiên dữ liệu thuộc các cụm khác nhau lại khác nhau. Nhu cầu tổ chức dữ liệu ngày càng tăng và tìm hiểu thông tin có giá trị từ dữ liệu, điều này khiến cho các kỹ thuật phân cụm trở nên khó khăn hơn. được áp dụng rộng rãi trong nhiều lĩnh vực ứng dụng như trí tuệ nhân tạo, sinh học, quản lý quan hệ khách hàng, nén dữ liệu, khai thác dữ liệu, truy xuất thông tin, xử lý hình ảnh, học máy, tiếp thị, y học, nhận dạng mẫu, tâm lý học, thống kê (Shibao và cộng sự, 2007), v.v

K-means là một phương pháp số, không giám sát, không xác định, lặp lại. Nó đơn giản và rất nhanh nên trong nhiều ứng dụng thực tế, phương pháp này được chứng minh là một phương pháp rất hiệu quả và có thể cho kết quả phân cụm tốt. Nhưng nó rất thích hợp để tạo ra các cụm hình cầu. Một số nỗ lực đã được các nhà nghiên cứu thực hiện để nâng cao hiệu quả của thuật toán K-means (Fahim và cộng sự, 2006). Trong nghiên cứu của Shibao và cộng sự (2007), có một thuật toán K-means được cải tiến dựa trên trọng số. Đây là một thuật toán phân cụm phân vùng mới, có thể xử lý dữ liệu thuộc tính số và cũng có thể xử lý dữ liệu thuộc tính ký hiệu. Trong khi đó, phương pháp này làm giảm tác động của các điểm cô lập và “nhiều” nên nâng cao hiệu quả phân cụm. Tuy nhiên, phương pháp này không cải thiện được độ phức tạp của thời gian. Trong nghiên cứu của Yuang và cộng sự (2004), nó đã đề xuất một phương pháp có hệ thống để tìm các trung tâm cụm ban đầu. Các trung tâm thu được bằng phương

pháp này phù hợp với việc phân phối dữ liệu. Do đó phương pháp này có thể tạo ra kết quả phân cụm chính xác hơn thuật toán k-mean tiêu chuẩn, nhưng phương pháp này không có bất kỳ cải tiến nào về thời gian thực hiện và độ phức tạp về thời gian của thuật toán.

James MacQueen, người đã đề xuất thuật ngữ "k-mean" (Shalove Agarwal và cộng sự, 2012) vào năm 1967. Nhưng thuật toán tiêu chuẩn được Stuart Lloyd giới thiệu lần đầu tiên vào năm 1957 dưới dạng kỹ thuật điều chế xung mã. Thuật toán phân cụm K-Means là phương pháp phân tích cụm dựa trên phân vùng (Juntao Wang và cộng sự, 2011). Theo thuật toán, trước tiên chọn k đối tượng làm tâm cụm ban đầu, sau đó tính khoảng cách giữa mỗi tâm cụm và từng đối tượng và gán nó cho cụm gần nhất, cập nhật giá trị trung bình của tất cả các cụm, lặp lại quá trình này cho đến khi hàm tiêu chí hội tụ. Tiêu chí lỗi bình phương để phân cụm.

$$\sum_{i=1}^k \sum_{j=1}^{n_i} |x_{ij} - m_i|$$

x_{ij} là mẫu j của i-class, m_i là tâm của i-class, n_i là số lượng mẫu i-class, Bước thuật toán được thể hiện trên Hình 1.

Thuật toán phân cụm K-mean được mô tả đơn giản như sau:

Đầu vào: N đối tượng làm cụm $\{x_1, x_2, \dots, x_n\}$, số cụm k;

Đầu ra: k cụm và tổng độ khác nhau giữa mỗi đối tượng và tâm cụm gần nhất của nó là nhỏ nhất;

- Chọn tùy ý k đối tượng làm tâm cụm ban đầu (m_1, m_2, \dots, m_k);
- Tính khoảng cách giữa mỗi đối tượng x_i và tâm mỗi cụm, sau đó gán từng đối tượng vào cụm gần nhất, công thức tính khoảng cách như sau:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^d (x_{i1} - m_{j1})^2}$$

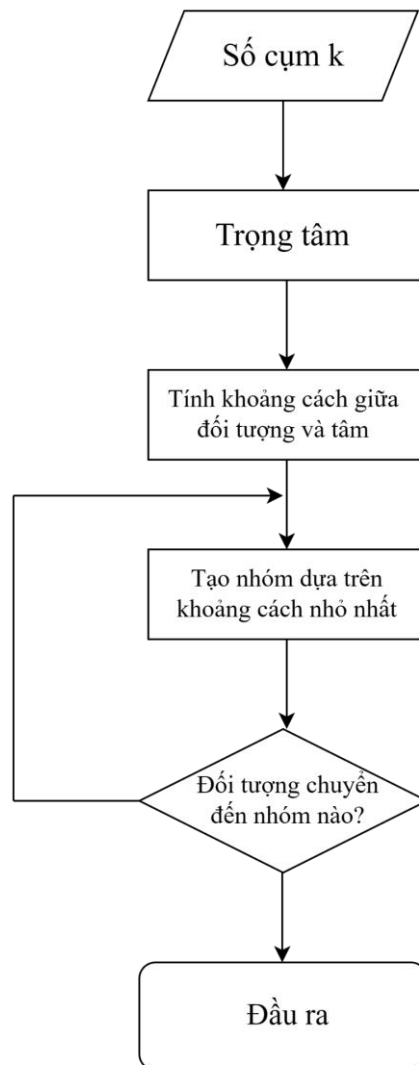
$i = 1, 2, \dots, N$

$j = 1, 2, \dots, k$

$d(x_i, m_i)$ là khoảng cách giữa dữ liệu i và cụm j;

- Tính giá trị trung bình của các đối tượng trong mỗi cụm khi là tâm của cụm mới, $m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$

$i=1, 2 \dots k$; N_i là số mẫu của cụm i hiện tại;



Hình 1. Sơ đồ của K-Means

Ưu và nhược điểm của K-Means clustering:

Ưu điểm	Nhược điểm
<ul style="list-style-type: none"> - Thuật toán đơn giản, dễ dàng sử dụng tốt cho các bài toán phân cụm. - K-Means thực hiện phân cụm tốt mà không cần biết nhãn dữ liệu đầu vào. 	<ul style="list-style-type: none"> - Trong phân cụm K-Means, người dùng cần chỉ định số cụm nâng cao. (Serban và cộng sự, 2006) - Hiệu suất của thuật toán phân cụm K-

(Học không giám sát)	Means phụ thuộc vào trọng tâm ban đầu mà tại sao thuật toán không đảm bảo cho giải pháp tối ưu (Serban và cộng sự, 2006)
- K-Means là nền tảng cho nhiều thuật toán phức tạp sau này.	- K-Means sẽ không hiệu quả nếu các cụm chênh lệch về số lượng điểm, phân bố dữ liệu không có dạng cầu, hay bài toán với 1 điểm dữ liệu có thể là con của 2 cụm.

Bảng 1. Ưu và nhược điểm của K-Means Clustering

2.2.2. Các chỉ số tài chính

2.2.2.1. Chỉ số ROE

ROE (Return on equity) là tỷ số lợi nhuận ròng trên vốn chủ sở hữu. Theo Horngren và Harrison (2007) cho rằng tỷ suất lợi nhuận trên vốn chủ sở hữu thể hiện mối quan hệ giữa lợi nhuận ròng với vốn chủ sở hữu của cổ đông phổ thông, lợi nhuận kiếm được trên mỗi 1 USD đầu tư của cổ đông phổ thông là bao nhiêu.

Công thức của ROE được tính như sau:

$$ROE = \frac{\text{Lợi nhuận ròng}}{\text{Vốn chủ sở hữu trung bình của cổ đông}}$$

- *Lợi nhuận ròng* là lợi nhuận sau khi trừ đi các chi phí và thuế.
- *Vốn chủ sở hữu trung bình của cổ đông* là trung bình của tổng vốn chủ sở hữu của công ty trong một khoảng thời gian.

Trong nghiên cứu của Asraf và Desda (2020), tỷ suất lợi nhuận ròng trên vốn chủ sở hữu càng cao thì càng tốt. Điều này bởi vì vị thế của chủ sở hữu công ty ngày càng vững chắc, và nhà đầu tư có thể đo lường được mức độ lợi tức đầu tư mà họ đã thực hiện.

Ngoài ra, khi phân cụm các công ty dựa trên ROE, ta có thể nhận ra các đặc điểm chung giữa các công ty có hiệu suất tương tự. Theo nghiên cứu của Yuli và Lia (2019), ROE có ảnh hưởng tới giá cổ phiếu. Các nhóm công ty có ROE cao có thể cho thấy khả năng tạo ra lợi nhuận cao từ vốn chủ sở hữu của họ, trong khi các nhóm có ROE thấp hơn có thể đang gặp khó khăn trong việc sinh lợi nhuận. Điều này có thể

giúp nhà đầu tư hiểu rõ hơn về tiềm năng sinh lợi nhuận và rủi ro của từng nhóm công ty.

Nhóm nghiên cứu chọn ROE làm một chỉ số quan trọng trong nghiên cứu này vì nó tập trung vào khả năng tạo lợi nhuận từ vốn sở hữu, không phụ thuộc vào cơ cấu tài chính hay quản lý nợ. Điều này giúp loại bỏ những yếu tố khác nhau giữa các công ty và tập trung vào khía cạnh tạo lợi nhuận cốt lõi của doanh nghiệp.

2.2.2.2. Chỉ số P/E

Chỉ số P/E (Price to Earnings) là một chỉ số được sử dụng để đánh giá mối quan hệ giữa giá thị trường của một cổ phiếu và thu nhập trên mỗi cổ phiếu (EPS). Tỷ lệ P/E được tính bằng cách chia giá thị trường của mỗi cổ phiếu cho EPS tương ứng. Trong một nghiên cứu được thực hiện bởi Chisholm (2009), tác giả tập trung vào việc phân tích tỷ lệ P/E một cách cụ thể. Mục tiêu của nghiên cứu là đánh giá xem cổ phiếu nào trong cùng một lĩnh vực được định giá cao hay thấp dựa trên chỉ số P/E. Ta có thể so sánh chỉ số P/E của các công ty hoạt động cùng ngành cùng lĩnh vực và hiệu quả hoạt động đều bị ảnh hưởng bởi cùng loại yếu tố. Tuy nhiên, khi so sánh chỉ số P/E với các công ty hoạt động khác lĩnh vực sẽ gặp vấn đề. Để định giá cổ phiếu, các chuẩn mực kế toán khác nhau sẽ được sử dụng và có thể ước tính tỷ lệ P/E của một công ty có lãi hoặc thậm chí là lỗ. Tuy nhiên, độ tin cậy còn hạn chế. Nhà đầu tư sẵn sàng trả giá cao hơn cho kỳ vọng tăng trưởng cao dưới dạng tỷ lệ P/E cao. Tỷ lệ P/E bị ảnh hưởng bởi mức lãi suất chung của thị trường và những thay đổi về lãi suất này có xu hướng ảnh hưởng đến thu nhập của doanh nghiệp.

Công thức xác định chỉ số P/E:

$$P/E = \frac{\text{Giá trị thị trường trên mỗi cổ phiếu}}{\text{Thu nhập trên mỗi cổ phiếu}}$$

Tỷ lệ P/E vẫn là công cụ định giá được sử dụng rộng rãi nhất nhằm giúp các nhà đầu tư xác định giá trị thị trường của một cổ phiếu so với thu nhập của công ty. Các nhà phân tích sử dụng chỉ số này để định giá cổ phiếu lần đầu phát hành ra công chúng. Chỉ số này cũng được sử dụng làm thước đo giá trị tương đối khi so sánh các công ty niêm yết. Tỷ lệ P/E triển vọng được xây dựng dựa trên kỳ vọng trung bình về triển vọng tăng trưởng trong tương lai. Tỷ lệ P/E cao thường cho thấy toàn bộ thị trường kỳ vọng thu nhập sẽ tăng trưởng đáng kể trong tương lai. Mặc khác, dễ hiểu hơn, tỷ lệ P/E cho thấy thị trường sẵn sàng trả bao nhiêu hiện nay cho một cổ phiếu

dựa trên thu nhập trong quá khứ hoặc tương lai của cổ phiếu đó. P/E cao có thể có nghĩa là giá cổ phiếu cao so với thu nhập và có thể được định giá quá cao. Ngược lại, P/E thấp có thể chỉ ra rằng giá cổ phiếu hiện tại thấp so với thu nhập.

Arslan và cộng sự (2014) đã phân tích tác động của tỷ suất cổ tức và tỷ lệ P/E đến lợi nhuận cổ phiếu và xác định mối quan hệ giữa quy mô và giá cổ phiếu dựa trên dữ liệu từ 111 công ty niêm yết phi tài chính KSE trong giai đoạn 1998-2009. Họ báo cáo rằng tỷ lệ P/E và quy mô của công ty có tác động tích cực đáng kể đến giá cổ phiếu. Họ cũng báo cáo mối quan hệ tiêu cực giữa tỷ suất cổ tức và giá cổ phiếu. Kết quả của họ cũng khuyến nghị rằng các nhà đầu tư có thể sử dụng các tiêu chí đầu tư bao gồm quy mô công ty và những bất thường về tỷ lệ P/E để kiếm được lợi nhuận bất thường.

Trong phân cụm, chỉ số P/E có thể được sử dụng để phân loại các công ty vào các nhóm có giá trị tương tác tương đồng. Cổ phiếu của công ty với P/E thấp có thể thuộc vào một nhóm, trong khi cổ phiếu với P/E cao có thể thuộc vào một nhóm khác. Điều này giúp tạo ra các phân khúc cổ phiếu có đặc điểm giá trị tương tự, làm cho mô hình dự đoán có thể được tinh chỉnh theo từng phân khúc cụ thể.

Khi phân cụm cổ phiếu theo chỉ số P/E, mô hình có thể được sử dụng để quản lý rủi ro theo từng nhóm. Nhóm cổ phiếu với P/E thấp có thể được coi là có rủi ro thấp hơn và có thể được quản lý khác biệt so với nhóm có P/E cao.

Theo dõi sự biến động của chỉ số P/E theo thời gian có thể cung cấp thông tin quan trọng về xu hướng thị trường và giúp mô hình dự đoán tốt hơn bằng cách tích hợp thông tin thời gian trong quá trình phân cụm.

2.2.2.3. Chỉ số P/B

Mối quan hệ giữa giá trị thị trường hiện tại và giá trị sổ sách của một công ty được đo bằng tỷ lệ P/B, còn được gọi là tỷ lệ thị trường trên sổ sách. Tỷ lệ này đã được một số nhà phân tích chứng khoán sử dụng để phân loại chứng khoán hoặc danh mục đầu tư theo giá trị chia kích thước tăng trưởng của chúng.

Công thức tính chỉ số P/B:

$$P/B = \frac{\text{Giá trị cổ phiếu trên thị trường}}{\text{Giá trị ghi sổ của cổ phiếu}}$$

Tỷ lệ P/B giúp nhà đầu tư đánh giá xem cổ phiếu có đang bị định giá thấp hơn so với giá trị thực hay không. Từ đó, họ có thể đưa ra quyết định mua vào hoặc bán ra để tăng lợi nhuận và giảm rủi ro. Đối với doanh nghiệp, chỉ số P/B phụ thuộc vào nhiều yếu tố như tốc độ tăng trưởng, doanh thu, lợi nhuận, lợi thế cạnh tranh, ngành nghề kinh doanh, lạm phát, GDP và nhiều yếu tố khác. Điều này là do những yếu tố này có thể gây ra sự thay đổi liên tục trong giá cổ phiếu trên thị trường.

Một số nhà nghiên cứu đã sử dụng chỉ số P/B để phân tích quan hệ giữa giá trị thị trường và giá trị sổ sách của công ty. Ví dụ, Capaul và cộng sự (1993) đã định nghĩa tỷ lệ P/B là "giá hiện tại trên mỗi cổ phiếu chia cho giá trị sổ sách được báo cáo gần đây nhất trên mỗi cổ phiếu. Giá của chứng khoán thể hiện đánh giá của nhà đầu tư về triển vọng trong tương lai, trong khi giá trị sổ sách của nó thể hiện sự thể hiện của kế toán về chi phí kèm theo của nó càng lớn". Sự tương tác giữa các biến kế toán và giá trị thị trường của công ty là một vấn đề quan trọng trong tài chính. Chỉ số P/B cho thấy sự đánh giá của nhà đầu tư về triển vọng tương lai, trong khi giá trị sổ sách thể hiện sự thể hiện của kế toán về chi phí trong quá khứ. Nếu triển vọng tương lai của một công ty lớn hơn, tỷ lệ giữa triển vọng tương lai và chi phí kèm theo của nó càng lớn.

Một số tác giả như Garza-Gómez (2001) coi tỷ lệ P/B là một biến số tốt để giải thích mối quan hệ giữa hai yếu tố giá trị thị trường và giá trị sổ sách. Các nhà đầu tư thấy P/B rất hữu ích trong phân tích đầu tư. Tỷ lệ này không chỉ cho phép so sánh giữa giá trị thị trường và giá trị sổ sách mà còn tương đối đơn giản để hiểu và vận dụng tính toán. Tỷ lệ P/B của các công ty tương tự cùng trong một ngành, một lĩnh vực có thể được so sánh để xác định xem công ty đó có được định giá quá cao hay quá thấp hay không. Các tác giả cũng nhận thấy rằng sự kết hợp giữa tỷ lệ P/B và giá trị thị trường của vốn chủ sở hữu đã thể hiện được vai trò của đòn bẩy tài chính và giải thích được lợi nhuận cổ phiếu. Barber and Lyon (1997) đã sử dụng cách tiếp cận tương tự được Fama và French áp dụng với các công ty tài chính được niêm yết trên sàn chứng khoán từ năm 1973 đến năm 1994. Phân tích cho thấy mối quan hệ giữa quy mô, giá trị sổ sách và lợi nhuận chứng khoán là tương đương nhau đối với công ty tài chính và công ty phi tài chính.

Tóm lại, chỉ số P/B được sử dụng để đánh giá mối quan hệ giữa giá trị thị trường và giá trị sổ sách của một công ty. Nó cung cấp thông tin quan trọng cho nhà đầu tư

để đưa ra quyết định giao dịch và giúp doanh nghiệp hiểu rõ hơn về giá trị của mình trên thị trường chứng khoán.

2.2.3. Deep Learning

2.2.3.1. Giới thiệu về Deep Learning

Deep Learning được phát triển từ mạng lưới thần kinh nhân tạo và hiện nay nó là một lĩnh vực học máy phổ biến. Việc nghiên cứu mạng nơ-ron nhân tạo bắt đầu từ những năm 1940. McCulloch và cộng sự (1943) đã đề xuất mô hình McCulloch-Pitts (MP) bằng cách phân tích và tóm tắt các đặc điểm của tế bào thần kinh. Hebb và cộng sự đề xuất một lý thuyết lắp ráp tế bào để giải thích sự thích nghi của tế bào thần kinh não trong quá trình học tập. Lý thuyết này có ảnh hưởng quan trọng đến sự phát triển của mạng lưới thần kinh. Sau đó Rosenblatt và cộng sự (1958) đã phát minh ra thuật toán perceptron. Thuật toán này là một loại phân loại nhị phân thuộc về học tập có giám sát. Widrow đã đề xuất phần tử tuyến tính thích ứng và đó là mạng nơ-ron nhân tạo một lớp dựa trên mô hình MP. Thật không may, Minsky và Papert đã chỉ ra rằng thuật toán perceptron có những hạn chế lớn về mặt lý thuyết và đưa ra đánh giá tiêu cực về triển vọng của mạng lưới thần kinh, khiến sự phát triển của mạng lưới thần kinh đạt đến điểm thấp nhất. Tuy nhiên, Hopfield và cộng sự (1982) đã đề xuất mạng Hopfield vào đầu những năm 1980. Điều này làm cho mạng lưới thần kinh nhân tạo hồi sinh. Sau đó Hinton và cộng sự (2012) đã đề xuất máy Boltzmann bằng cách sử dụng thuật toán ủ mô phỏng. Vào những năm 1990, nhiều phương pháp học máy nông khác nhau lần lượt được đề xuất, chẳng hạn như máy vector hỗ trợ (Cortes và cộng sự, 1995), Boosting (Freund và cộng sự, 1997). Do những ưu điểm của các phương pháp này cả về mặt lý thuyết và ứng dụng, mạng nơ-ron nhân tạo lại đạt đến mức thấp nhất. Sau Hinton và cộng sự. Đưa ra khái niệm Deep Learning trên tạp chí Science vào năm 2006, mạng nơ-ron nhân tạo một lần nữa nhận được nhiều sự quan tâm từ cộng đồng nghiên cứu.

Các mô hình Deep Learning thường áp dụng cấu trúc phân cấp để kết nối các lớp của chúng. Đầu ra của lớp thấp hơn có thể được coi là đầu vào của lớp cao hơn thông qua các phép tính tuyến tính hoặc phi tuyến đơn giản. Các mô hình này có thể chuyển đổi các tính năng cấp thấp của dữ liệu thành các tính năng trừu tượng cấp cao. Nhờ đặc điểm này, các mô hình Deep Learning có thể mạnh hơn các mô hình Machine Learning nông trong việc biểu diễn tính năng. Hiệu suất của các phương pháp Machine

Learning truyền thống thường dựa vào trải nghiệm của người dùng, trong khi các phương pháp Deep Learning lại dựa vào dữ liệu. Do đó, chúng ta có thể nhận thấy rằng các phương pháp Deep Learning đã làm giảm nhu cầu của người dùng. Với sự tiến bộ của công nghệ máy tính, hiệu suất của máy tính được cải thiện nhanh chóng. Trong khi đó, thông tin trên Internet cũng tràn lan. Những yếu tố này tạo động lực mạnh mẽ cho Deep Learning phát triển và đưa Deep Learning trở thành phương pháp phổ biến trong Machine Learning.

2.2.3.2. Quá trình nghiên cứu

Khái niệm Deep Learning lần đầu tiên được đưa ra vào năm 2006. Sau đó, Deep Learning vẫn không ngừng phát triển ở nước ngoài. Hiện tại, có rất nhiều nhân vật kiệt xuất như Geoffrey Hinton, Yoshua Bengio, Yann LeCun và Andrew Ng. Họ đang dẫn đầu hướng nghiên cứu về Deep Learning. Một số công ty như Google và Facebook đã đạt được nhiều thành tựu nghiên cứu về Deep Learning và áp dụng chúng vào nhiều lĩnh vực khác nhau. Năm nay, chương trình AlphaGo của Google đã đánh bại Lee Sedol trong cuộc thi cờ vây, điều này cho thấy Deep Learning có khả năng học tập mạnh mẽ. Hơn nữa, DeepDream của Google là một phần mềm tuyệt vời không chỉ có thể phân loại hình ảnh mà còn tạo ra những bức tranh nhân tạo và kỳ lạ dựa trên kiến thức của chính nó. Facebook công bố hệ thống trí tuệ nhân tạo mới có tên Deep Text. Deep Text là một công cụ hiểu văn bản dựa trên Deep Learning, có thể phân loại lượng dữ liệu khổng lồ, cung cấp các dịch vụ tương ứng sau khi xác định tin nhắn trò chuyện của người dùng và dọn sạch tin nhắn rác.

Deep Learning bắt đầu tương đối muộn nhưng phát triển rất nhanh. Đã đạt được tiến bộ đáng kể ở các trường cao đẳng, đại học, viện nghiên cứu và công ty. Baidu đã thành lập một học viện Deep Learning để khám phá cách hoàn thành nhiều nhiệm vụ với Deep Learning. Xe mặt đất không người lái của Baidu đã hoàn thành thử nghiệm trên đường trong điều kiện đường phức tạp. IFLYTEK bắt đầu nghiên cứu nhận dạng giọng nói dựa trên Deep Neural Network (DNN) vào năm 2010. Họ đã cho ra mắt hệ thống nhận dạng giọng nói trực tuyến đầu tiên của Trung Quốc và một công nghệ tiên tiến để nhận dạng các ngôn ngữ khác nhau. Và bây giờ, họ đã xuất bản nền tảng điện toán hiệu năng cao (HPC) hợp tác với Intel.

2.2.3.3. Các mô hình Deep Learning

2.2.3.3.1. Autoencoder

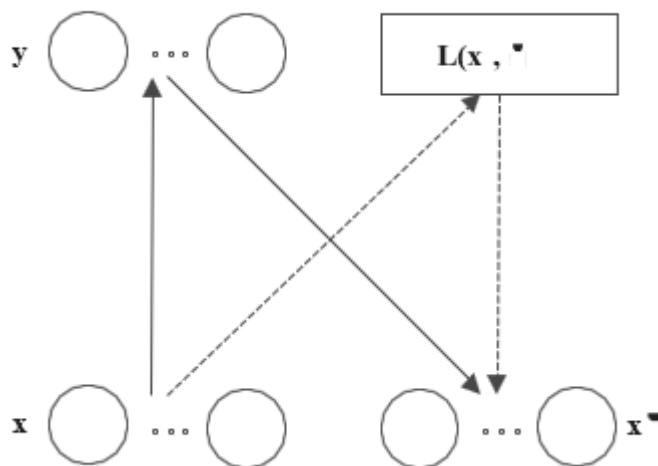
Autoencoder chủ yếu được sử dụng để xử lý dữ liệu nhiều chiều phức tạp. Mục đích của nó là tìm hiểu cách biểu diễn một tập hợp dữ liệu thông qua việc giảm kích thước. Khi xử lý đầu vào x bằng cách sử dụng một loạt các phương pháp ánh xạ và trọng số, có thể nhận được đầu ra y có chiều thấp. Sau đó, áp dụng các phương pháp ánh xạ và trọng số nghịch đảo để biến đổi y thành đầu ra x' có kích thước giống với đầu vào x . Bây giờ, tất cả những gì chúng ta phải làm là làm cho hàm lỗi $L(x, x')$ nhỏ nhất bằng cách huấn luyện lặp đi lặp lại các trọng số của mạng. Nguyên lý cơ bản của AE được thể hiện trong Hình 1.

AE cũng có nhiều cấu trúc cải tiến như Denoising Autoencoder (Vincent và cộng sự, 2008), và Sparse Autoencoder (Andrew và cộng sự, 2011). Đối với Denoising Autoencoder, nó sử dụng dữ liệu gốc có nhiều ngẫu nhiên để huấn luyện trọng số mạng, điều này làm cho các tính năng được trích xuất trở nên mạnh mẽ hơn. Đối với Sparse Autoencoder, bên cạnh việc tăng số lượng lớp và nơ-ron ẩn, Sparse Autoencoder giới hạn trạng thái kích hoạt của các nút ẩn, trong đó chỉ một số lượng nhỏ các nút ẩn ở trạng thái được kích hoạt và hầu hết các nút ẩn ở trạng thái không được kích hoạt.

Xiong và cộng sự (2016) đã đề xuất một mạng mã hóa tự động được sửa đổi để nhận biết và tách các chất dị thường khỏi một tập hợp các mẫu địa hóa. Continuous Restricted Boltzmann Machine (CRBM) được sử dụng như một phần của mạng bộ mã hóa tự động (Vincent và cộng sự, 2008). Các tác giả áp dụng ba bước để huấn luyện mô hình, đó là huấn luyện trước CRBM, hủy kiểm soát CRBN để xây dựng mạng và tinh chỉnh các tham số thông qua lan truyền ngược. Cuối cùng, phương pháp này đạt được kết quả tốt trong việc nhận biết các dị thường địa hóa đa biến.

Louizos và cộng sự (2011) đã đề xuất một mô hình bộ mã hóa tự động công bằng đa dạng, có thể làm cho các biểu diễn tiềm ẩn có nhiều thông tin nhất về các biến ngẫu nhiên được quan sát nhưng có ít thông tin về các biến nhảy cảm hoặc gây phiền toái. Nói cách khác, mô hình có thể tách các yếu tố không mong muốn khỏi các biến thể trong khi vẫn giữ lại càng nhiều thông tin càng tốt từ những gì còn lại. Để loại bỏ các biến nhảy cảm hoặc gây phiền toái khỏi các biểu diễn tiềm ẩn, Louizos và cộng sự

(2011) đã thêm một điều khoản phạt dựa trên thước đo Maximum Mean Discrepancy vào mô hình. Cuối cùng, họ đã áp dụng mô hình này vào một số nhiệm vụ và thu được kết quả rất tốt.



Hình 2. Nguyên lý cơ bản của Autoencoder

2.2.3.3.2. Deep Belief Network

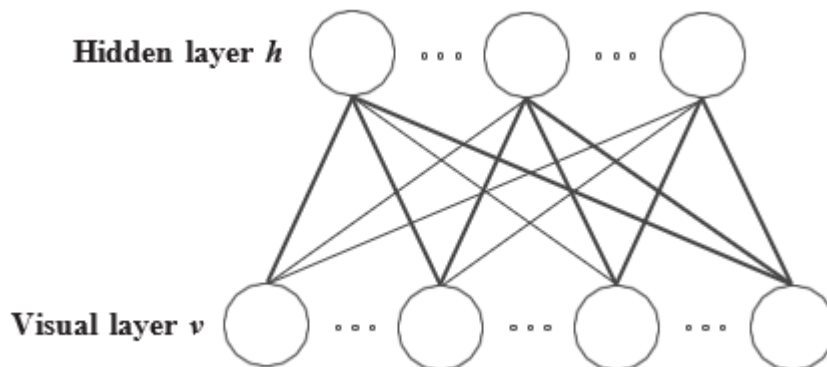
Deep Belief Network là một loại mạng thần kinh được xếp chồng lên nhau bởi một số Restricted Boltzmann Machines (RBMs). RBM là một loại mô hình mạng lưới thần kinh ngẫu nhiên tổng quát, xuất phát từ máy Boltzmann. Mặc dù RBM kế thừa cấu trúc nơ-ron hai lớp của máy Boltzmann nhưng không có sự kết nối nào giữa các nơ-ron trong cùng một lớp mà chỉ có toàn bộ kết nối giữa lớp thị giác và lớp ẩn. Cấu trúc cơ bản của RBM được thể hiện trong Hình 3.

Sau khi tăng số lượng lớp RBMs ẩn, chúng ta có thể có được máy Boltzmann sâu. Sau đó, áp dụng kết nối được định hướng từ trên xuống gần lớp trực quan để có thể có được mô hình DBN. Khi huấn luyện mạng, phương pháp huấn luyện trước theo từng lớp không được giám sát tham lam có thể được sử dụng để lấy trọng số của mạng. Nó chỉ đào tạo một lớp tại một thời điểm với đầu ra của lớp thấp hơn được sử dụng làm đầu vào của lớp cao hơn. Sau đó, thuật toán lan truyền ngược được sử dụng để tinh chỉnh toàn bộ mạng.

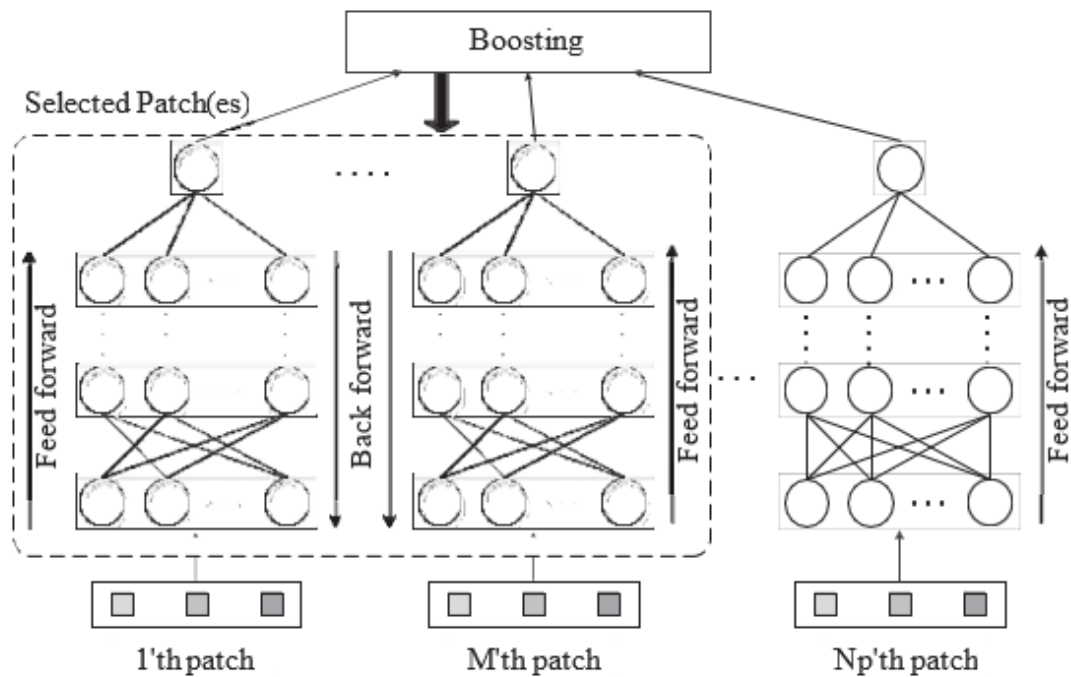
Liu và cộng sự (2014) đã đề xuất một Boosted Deep Belief Network (BDBN), bao gồm một số DBN. Mỗi DBN được sử dụng để học các biểu diễn tính năng phân cấp và tất cả các DBN được coi là người học yếu sẽ được kết nối với nhau thông qua một bộ phân loại nâng cao. BDBN áp dụng quy trình học tính năng không giám sát

(BU-UFL) từ dưới lên và quy trình tăng cường tính năng được giám sát từ trên xuống (BTD-SFS). Nó được sử dụng để nhận dạng nét mặt. Mạng chia hình ảnh khuôn mặt thành nhiều mảng chồng chéo một phần. Sau đó, nó sử dụng quy trình BU-UFL để tìm hiểu cách trình bày tính năng từ mỗi bản vá bằng một quy trình DBN và BTD-SFS nhằm tinh chỉnh các tính năng bằng cách xử lý các lỗi phân loại do trình phân loại nâng cao và trình học yếu tạo ra. Cuối cùng, mô hình đạt được kết quả tốt hơn so với các công việc liên quan khác. Cấu trúc của mô hình này được thể hiện trong Hình 4.

Kim và cộng sự (2016) đã đề xuất một phương pháp phát hiện độ thật của dấu vân tay có thể phân biệt dấu vân tay được quét là thật hay giả trước khi nhận dạng. Mô hình trong bài nghiên cứu của Kim và cộng sự (2016) sử dụng mạng lưới niềm tin sâu sắc. Cấu trúc của DBN giống như DBN bình thường ngoại trừ lớp cuối cùng có hai nút đầu ra để đưa ra quyết định về độ sống. Trước khi nhập dữ liệu vào DBN, hình ảnh dấu vân tay phải được xử lý. Các tác giả sử dụng máy dò góc Harris hai chiều để suy ra vị trí trung bình chứa vùng quan tâm. Mô hình này được huấn luyện trước bằng phương pháp học không giám sát và được tinh chỉnh bằng tập huấn luyện gồm các đầu vào được gắn nhãn. Cuối cùng, các thử nghiệm cho thấy phương pháp đề xuất có thể cung cấp khả năng phát hiện độ thật của dấu vân tay hiệu quả và hiệu quả để ngăn chặn việc giả mạo dấu vân tay giả.



Hình 3. Cấu trúc cơ bản của Restricted Boltzmann Machine



Hình 4. Cấu trúc của Boosted Deep Belief Network được tăng cường bao gồm nhiều Deep belief networks. Chỉ những cái trong hộp chấm sẽ được tinh chỉnh chung.

2.2.3.3.3. Convolutional Neural Network

Vào những năm 1960, khái niệm trường tiếp nhận đã được đề xuất. Neocognitron (Fukushima và cộng sự, 1983) dựa trên lĩnh vực tiếp nhận được đề xuất vào những năm 1980 được coi là tiền thân của CNN. Đặc điểm nổi bật của CNN là mạng sử dụng trường tiếp nhận cục bộ và chia sẻ trọng số. Bằng cách sử dụng hai chiến lược này, số lượng tham số huấn luyện sẽ giảm đáng kể, điều này có thể làm cho mạng trở nên ít phức tạp hơn. Cấu trúc CNN điển hình bao gồm một số lớp tích chập, lớp gộp và lớp được kết nối đầy đủ. Lớp chập được sử dụng để trích xuất đặc trưng. Mỗi đầu vào của nơron trong lớp này được kết nối với trường tiếp nhận cục bộ của lớp trước. Lớp tổng hợp được sử dụng để ánh xạ tính năng. Nó có thể giảm kích thước của dữ liệu và có thể duy trì tính bất biến của cấu trúc mạng.

Những năm gần đây, CNN nhận được rất nhiều sự quan tâm của các nhà nghiên cứu. Đó là một mô hình tuyệt vời có thể hoàn thành nhiệm vụ một cách hiệu quả. Có nhiều loại cấu trúc CNN như LeNet (Yecum và cộng sự, 1998), AlexNet (Krizhevsky và cộng sự, 2012), ZFNet (Zeiler và cộng sự, 2014), VGGNet (Simonyan và cộng sự, 2014) và GoogleNet (Szegedy và cộng sự, 2015). LeCun và cộng sự (1998) đề xuất một mạng lưới thần kinh tích chập có tên là LeNet và áp dụng vào nhận dạng chữ viết

tay. AlexNet chủ yếu được sử dụng để phát hiện đối tượng. Sau đó, ZFNet, VGGNet và GoogleNet được đưa ra dựa trên AlexNet. Hiện nay CNN vẫn đang là một chủ đề sôi động với nhiều hướng khai thác. Một số nhà nghiên cứu muốn tăng độ phức tạp của cấu trúc CNN. Những người khác muốn kết hợp CNN với các phương pháp học máy truyền thống khác.

Liu và cộng sự (2015) đã đề xuất một mô hình SSD có thể phát hiện vật thể một cách hiệu quả với độ chính xác cao. Mô hình bao gồm cấu trúc mạng cơ sở được cắt ngắn và cấu trúc phụ trợ. Mạng cơ sở rút gọn trong (Liu và cộng sự, 2015) sử dụng VGG-16 và cấu trúc phụ trợ sử dụng một số lớp tính năng ở cuối VGG-16. Mạng có thể tạo ra một tập hợp các hộp giới hạn có kích thước cố định từ nhiều bản đồ đặc trưng. Nó cũng có thể cho điểm danh mục nếu có một đối tượng trong các hộp giới hạn và độ lệch tương ứng. Khi huấn luyện SSD, hàm mất mát là tổng trọng số của mất mát nội địa hóa và mất tin cậy sẽ được tạo ra khi truyền lan về phía trước. Cuối cùng, hàm mất mát có thể được sử dụng để tinh chỉnh mô hình lan truyền ngược. Cấu trúc của mô hình này được thể hiện trong Hình 5.

Kontschieder và cộng sự (2015) đã đề xuất Deep Neural Decision Forests, một cấu trúc mới hợp nhất các cây phân loại với CNN. Trong bài báo của họ, cấu trúc mạng rất rõ ràng rằng họ thay thế lớp softmax bằng mô hình cây quyết định ngẫu nhiên và khả vi. Cây quyết định là một loại phân loại có cấu trúc cây bao gồm các nút quyết định và nút dự đoán. Các nút quyết định quyết định các tuyến đường mà các mẫu đi dọc theo cây. Các nút dự đoán sẽ tính toán dự đoán của chúng. Cuối cùng, tất cả các dự đoán sẽ được tính trung bình riêng biệt và các mẫu có thể được đánh giá xem chúng thuộc loại nào. Khi truyền ngược, họ áp dụng nguyên tắc rủi ro thực nghiệm tối thiểu để tinh chỉnh mạng. Cấu trúc của mô hình cây được thể hiện trong Hình 6.

Levine và cộng sự (2016) đề xuất một phương pháp phối hợp tay và mắt của robot để nắm bắt mọi thứ từ hình ảnh một mắt. Cách tiếp cận này thống nhất việc học tăng cường với Deep Learning. Trong bài báo của mình, các tác giả sử dụng mạng lưới thần kinh tích chập để đưa ra dự đoán rằng liệu chuyển động của dụng cụ kẹp có thể dẫn đến việc nắm bắt thành công hay không. Ảnh hiện tại It và ảnh gốc I0 được coi là đầu vào của mạng. Nó cũng cung cấp vector lệnh vt làm đầu vào cho mạng sau khi hai hình ảnh được xử lý bởi năm lớp đầu tiên. Sau đó, chúng được nhập vào nhóm lớp tiếp theo để có xác suất nắm bắt thành công.

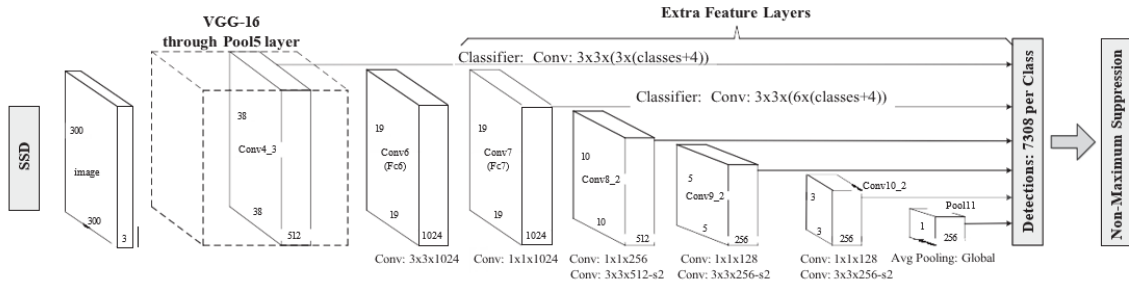
2.2.3.3.4. Recurrent Neural Network

Recurrent Neural Network là một loại mạng thần kinh nhân tạo. Ngoài việc có cấu trúc của mạng nơ-ron tiếp liệu, còn tồn tại các chu trình có hướng trong RNN. Cấu trúc này cho phép thông tin được lưu chuyển trong mạng, do đó đầu ra của mỗi thời điểm không chỉ liên quan đến đầu vào hiện tại mà còn liên quan đến đầu vào ở các dấu thời gian trước đó.

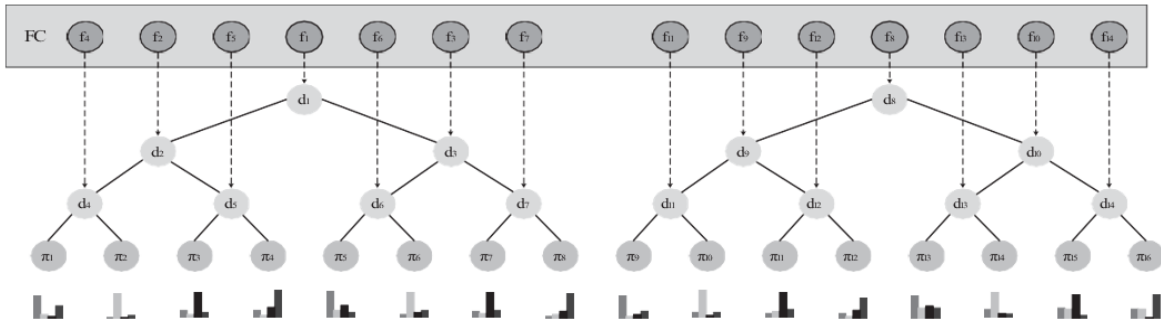
Mặc dù RNN truyền thống có thể xử lý dữ liệu chuỗi thời gian nhưng vẫn tồn tại một vấn đề nghiêm trọng về sự biến mất độ dốc trong quá trình truyền ngược. Do đó RNN chỉ có thể được sử dụng cho bộ nhớ ngắn hạn trong hầu hết các trường hợp. Để giải quyết vấn đề này, nhiều nhà nghiên cứu bắt đầu đưa ra một số loại cấu trúc cải tiến, chẳng hạn như Long Short-Term Memory (LSTM). Khác với RNN truyền thống, LSTM có ô nhớ và cấu trúc cổng đầu vào-đầu ra. Ô nhớ được sử dụng để ghi thông tin và cổng vào-ra xác định xem thông tin có khả năng chảy vào hay ra khỏi ô nhớ hay không. Do những đặc điểm này, LSTM có hiệu suất tốt hơn RNN trong các tác vụ bộ nhớ dài hạn.

Byeon và cộng sự (2015) đã đề xuất một cách tiếp cận hoàn toàn dựa trên học tập để ghi nhận cảnh bằng cách sử dụng một loại mạng thần kinh tái phát 2D LSTM. Mạng được chia thành ba lớp chính: lớp đầu vào, lớp ẩn và lớp đầu ra. Hình ảnh đầu vào được chia thành nhiều cửa sổ không chồng lên mạng đầu vào. Lớp ẩn bao gồm lớp LSTM 2D và lớp tiếp liệu. Lớp LSTM 2D được sử dụng để ghi nhớ thông tin ngữ cảnh theo mọi hướng và lớp phản hồi kết hợp thông tin lại với nhau. Lớp đầu ra chuẩn hóa các đầu ra từ lớp ẩn cuối cùng bằng hàm softmax và tạo ra xác suất về việc mục tiêu thuộc về lớp nào. Kết quả thực nghiệm cho thấy tính hiệu quả của mô hình đề xuất. Mạng được hiển thị trong Hình 7.

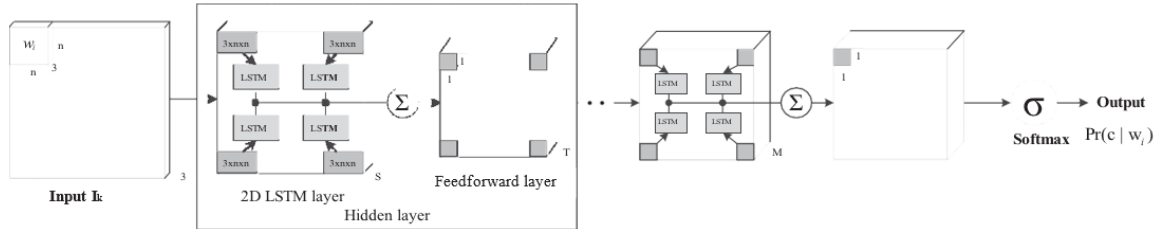
Luu và cộng sự (2016) đã giới thiệu một cách tiếp cận mới để cùng học cách biểu diễn tính năng qua nhiều nhiệm vụ liên quan. Điểm mới lạ của phương pháp này là chúng tích hợp LSTM vào khung học tập đa tác vụ. Trong bài báo của mình, họ đã đề xuất ba kiến trúc chia sẻ thông tin để mô hình hóa văn bản. Cái đầu tiên chỉ chia sẻ một lớp LSTM cho tất cả các tác vụ. Lớp thứ hai chỉ định một lớp LSTM cho mỗi tác vụ và mọi lớp LSTM có thể sử dụng thông tin từ các lớp khác. Cái cuối cùng không chỉ có đặc điểm của cái thứ hai mà còn xây dựng lớp LSTM hai chiều cho tất cả các tác vụ.



Hình 5. Mô hình sử dụng VGG-16 làm mạng cơ sở. Các lớp đối tượng bổ sung được thêm vào mạng cơ sở để có thể sử dụng các lớp đối tượng để đưa ra dự đoán phát hiện của mọi bản đồ đối tượng.



Hình 6. Mô hình cây của Deep Neural Decision Forests bao gồm một số cây quyết định. Các nút quyết định quyết định cách các mẫu đi qua cây. Các nút dự đoán tạo ra xác suất quyết định sau khi các mẫu đến các nút cuối cùng tương ứng.



Hình 7. Kiến trúc mạng bộ nhớ ngắn hạn dài 2D. Ảnh đầu vào I_k được chia thành một số cửa sổ. Mỗi cửa sổ được đưa vào bốn khối bộ nhớ LSTM riêng biệt. Đầu ra của mỗi khối LSTM được chuyển đến lớp tiếp liệu. Ở lớp cuối cùng, đầu ra của các khối LSTM cuối cùng được tổng hợp và gửi đến lớp softmax. Cuối cùng, các mạng đưa ra xác suất của lớp cho từng cửa sổ đầu vào

2.2.3.4. Ứng dụng của Deep Learning

Sau khi thảo luận về các mô hình và khung này, chúng ta có thể thấy rằng các phương pháp Deep Learning có thể giúp chúng ta đạt được hiệu suất trong các ứng

dụng khác nhau. Trong phần này, chúng tôi giới thiệu một số ứng dụng của Deep Learning trong thị giác máy tính và xử lý ngôn ngữ tự nhiên.

Deep Learning đã có sự phát triển vượt bậc trong lĩnh vực thị giác máy tính, chẳng hạn như phát hiện đối tượng, theo dõi đối tượng và phân đoạn hình ảnh. Phát hiện đối tượng nhằm mục đích nhận ra một lớp đối tượng từ một số lượng lớn hình ảnh. Các phương pháp phát hiện đối tượng truyền thống chủ yếu bao gồm lựa chọn vùng ứng cử viên, trích xuất và phân loại đặc trưng. Phương pháp trích xuất tính năng thủ công này cần người dùng thiết kế những tính năng họ nên trích xuất. Và những quá trình này thường tốn kém chi phí và thời gian. Deep Learning có khả năng học đặc điểm không giám sát và nó có thể trích xuất các đặc điểm của hình ảnh mà không cần bất kỳ sự can thiệp nào của con người. Vì vậy, nó dần dần thu hút ngày càng nhiều sự quan tâm của các nhà nghiên cứu. Sau Krizhevsky và cộng sự (2012) có bước đột phá nhờ sử dụng CNN trong ImageNet LSVRC 2012, Deep Learning ngày càng trở nên phổ biến trong lĩnh vực thị giác máy tính và có bước đột phá xuất sắc cho đến nay. Cho đến nay, phương pháp kết hợp ba mạng Inception còn lại với một Inception-v4 (Szegedy và cộng sự, 2016) khiến nhiệm vụ nhận dạng hình ảnh đạt được sai số 3,08% trong top 5. Learned-Miller và cộng sự (2012) đã đề xuất một phương pháp Deep Learning giúp độ chính xác của nhận dạng khuôn mặt tăng lên khoảng 87%. Hiện tại, các nhà nghiên cứu tại Đại học Trung Quốc Hồng Kông đã tăng độ chính xác nhận dạng khuôn mặt lên trên 99% (Sun và cộng sự, 2015).

Deep Learning đã và đang được phát triển không ngừng trong lĩnh vực xử lý ngôn ngữ tự nhiên và đạt được nhiều thành tựu trong nhiều ứng dụng, bao gồm nhận dạng giọng nói, tổng hợp giọng nói và trả lời câu hỏi. Các hệ thống nhận dạng giọng nói truyền thống chủ yếu dựa trên mô hình hỗn hợp Gaussian và mô hình Markov ẩn trong một thời gian dài. Tuy nhiên, các phương pháp này không thể xử lý tốt các đặc tính sâu và nhạy cảm với các nhiễu loạn từ môi trường bên ngoài. Sau khi áp dụng Deep Learning trong nhận dạng giọng nói, hiệu suất của hệ thống đã được cải thiện đáng kể. Giờ đây, hệ thống nhận dạng giọng nói Deep Speech 2 do Baidu thiết kế đã giảm tỷ lệ lỗi xuống còn 3,7% trong bài kiểm tra giọng nói tiếng Trung. Hiện tại, Google DeepMind đã xuất bản một hệ thống tổng hợp giọng nói mới được đặt tên là WaveNet (Simonyan và cộng sự, 2016). WaveNet là một loại Deep Neural Network và có thể tạo ra dạng sóng âm thanh thô. So với các hệ thống chuyển văn bản thành

giọng nói khác, WaveNet có thể tạo ra âm thanh cũng như âm nhạc chân thực hơn. Từ DeepMind, nó cho thấy WaveNet đã giảm hơn 50% khoảng cách giữa giọng nói của con người và giọng nói tổng hợp bằng tiếng Anh và tiếng Trung. Ask-Answering (QA) là một hướng nghiên cứu hot về xử lý ngôn ngữ tự nhiên, có thể đưa ra câu trả lời đúng và ngắn gọn dưới dạng ngôn ngữ tự nhiên cho các bài toán ngôn ngữ tự nhiên. Chiến thắng của Watson (Ferrucci và cộng sự, 2013) trước nguy cơ đã cho thấy rằng QA dựa trên Deep Learning có ưu thế vượt trội riêng.

2.3. Thực trạng vấn đề nghiên cứu: Khái quát các kết quả nghiên cứu đã đạt được

2.3.1. Trên thế giới

Thị trường chứng khoán là một hệ thống năng động phức tạp. Việc dự đoán giá cổ phiếu trong tương lai bị ảnh hưởng rất nhiều bởi các yếu tố bên trong và bên ngoài. Một số nhà nghiên cứu đã xuất bản các bài báo trong lĩnh vực này để dự báo dữ liệu chứng khoán. Tuy nhiên, kết quả của họ hơi khác so với dữ liệu chứng khoán ban đầu. Dự đoán dữ liệu chứng khoán hiện tại được coi là mục tiêu nghiên cứu.

Kim và cộng sự (2003) đề xuất máy vector hỗ trợ tài chính dự đoán chuỗi thời gian. Một số tính năng ban đầu được chọn như đầu vào của mạng. Ngoài ra, họ còn so sánh các mô hình SVM với mô hình ANN và mô hình suy luận dựa trên trường hợp. Armano (2005) đã nghiên cứu một phương pháp mới để thực hiện dự báo thị trường chứng khoán bằng cách sử dụng kiến trúc thần kinh di truyền lai. Kim và cộng sự (2005) điều tra tính hiệu quả của phương pháp kết hợp với mạng thần kinh trễ thời gian (TDNN) và thuật toán di truyền (GA) trong việc phát hiện các mô hình thời gian cho nhiệm vụ dự đoán thị trường chứng khoán và cho thấy phương pháp GA-TDNN tích hợp được đề xuất cho nghiên cứu này hoạt động tốt hơn so với TDNN tiêu chuẩn và Mạng thần kinh tái phát để phản ánh mô hình thời gian

Cao (2005) đã sử dụng mạng nơ-ron nhân tạo để dự đoán biến động giá cổ phiếu (tức là lợi nhuận của giá) cho các công ty giao dịch trên Sở giao dịch chứng khoán Thượng Hải và so sánh khả năng dự đoán của các mô hình mạng nơ-ron đơn biến và đa biến và kết quả cho thấy mạng nơ-ron vượt trội hơn các mô hình tuyến tính. Hassan (2007) đã sử dụng phương pháp mô hình markov ẩn (HMM) để dự báo giá cổ phiếu cho các thị trường có liên quan với nhau. HMM được sử dụng cho các vấn đề nhận dạng và phân loại mẫu vì tính phù hợp đã được chứng minh của nó đối với việc mô

hình hóa hệ thống động. Tác giả tóm tắt ưu điểm của HMM là nền tảng thống kê vững chắc. Nó có thể xử lý dữ liệu mới một cách mạnh mẽ và hiệu quả về mặt tính toán để phát triển và đánh giá các mẫu tương tự. Takashi Yamashita và cộng sự đã sử dụng mạng thần kinh đa nhánh (MBNN) để dự đoán giá cổ phiếu và mô phỏng đã được thực hiện để điều tra tính chính xác của dự đoán. Kết quả cho thấy MBNN có ít tham số hơn có thể có độ chính xác cao hơn NN thông thường khi dự đoán Nikkei-225 tại thời điểm $(t+1)$.

O'Connor và cộng sự (2006) đánh giá hiệu quả của việc sử dụng các chỉ số bên ngoài như giá hàng hóa và tỷ giá hối đoái trong việc dự đoán biến động của chỉ số Trung bình Công nghiệp Dow Jones và hiệu suất của từng kỹ thuật cũng được đánh giá bằng cách sử dụng các ma trận miền cụ thể khác nhau. Hassan và cộng sự (2007) đã sử dụng mô hình tổng hợp HMM, ANN và GA để dự báo hành vi thị trường tài chính. Công cụ này được sử dụng để phân tích chuyên sâu về thị trường chứng khoán. Giá cổ phiếu hàng ngày được chuyển đổi thành các tập hợp giá trị độc lập trở thành đầu vào cho HMM và GA để tối ưu hóa các tham số ban đầu của HMM. Sau đó HMM được sử dụng để xác định và định vị các mẫu tương tự. Trung bình có trọng số của chênh lệch giá của các mẫu tương tự được lấy để chuẩn bị dự báo cho ngày tiếp theo cần thiết.

Yildiz và cộng sự (2008) đề xuất mô hình ANN để dự báo hướng đi của Sở giao dịch chứng khoán Istanbul. Kết quả cho thấy mô hình ANN cho độ chính xác đạt 74,51%. TilaKaratne và cộng sự (2009) đã nghiên cứu các thuật toán mạng lưới thần kinh đã được sửa đổi để dự đoán liệu mua, nắm giữ hoặc bán cổ phiếu của các chỉ số thị trường chứng khoán là tốt nhất.

Akinwale và cộng sự (2005) đã kiểm tra việc sử dụng phân tích hồi quy và lan truyền ngược lỗi để dự đoán Giá thị trường chứng khoán Nigeria (NSMP) chưa được dịch và dịch. Tác giả đã sử dụng cấu trúc liên kết mạng $5 - j - 1$ để áp dụng năm biến đầu vào. Số lượng nơ-ron ẩn xác định các biến j trong quá trình chọn mạng. Cả những câu chưa dịch và đã dịch đều được phân tích và so sánh. Hiệu suất của NSMP được dịch bằng cách sử dụng phân tích hồi quy hoặc lan truyền lỗi vượt trội hơn so với NSMP chưa được dịch. Kết quả cho thấy trên NSMP chưa dịch dao động ở mức 11,3% trong khi ở NSMP là 2,7%.

George (2009) đã sử dụng bộ điều khiển hệ thống mờ thần kinh thích ứng để điều khiển mô hình quy trình thị trường chứng khoán và cũng đánh giá nhiều loại cổ phiếu. Giả thuyết thị trường hiệu quả được sử dụng để cải thiện dự đoán về xu hướng thị trường chứng khoán ngắn hạn. Kết quả thể hiện rõ ràng việc sử dụng tỷ suất lợi nhuận đề xuất (ROR). Lợi nhuận tốt hơn thu được khi nhà đầu tư phân bổ tài sản vào trái phiếu chính phủ không rủi ro khi lợi nhuận cổ phiếu dự đoán chuyển sang âm. Điều này được gọi là kết quả bất cân xứng của thị trường chứng khoán. Lần thứ hai nhà đầu tư phân bổ tài sản vào trái phiếu chính phủ không rủi ro đã có một số lợi nhuận tích cực. Điều này có nghĩa là lợi nhuận từ dự đoán chính xác và tổn thất từ dự đoán sai. Hệ thống thần kinh mờ thể hiện rõ ràng tiềm năng của dự đoán thị trường tài chính

Huang (2009) đã sử dụng mô hình dự đoán ngoại sinh tự hồi quy (ARX) được kết hợp với lý thuyết hệ thống xám và lý thuyết tập thô để tạo ra cơ chế dự báo thị trường chứng khoán và lựa chọn danh mục đầu tư tự động. Dữ liệu tài chính được thu thập tự động hàng quý và được đưa vào mô hình dự đoán ARX để dự báo xu hướng trong tương lai. Được phân cụm bằng thuật toán phân cụm K có nghĩa là và sau đó được cung cấp cho mô-đun phân loại RS để chọn cổ phiếu đầu tư phù hợp theo quy tắc ra quyết định. Ưu điểm là kết hợp các kỹ thuật dự báo khác nhau để nâng cao hiệu quả và độ chính xác của dự đoán tự động. Hiệu quả của các mô hình tổng hợp được đánh giá bằng cách so sánh độ chính xác dự báo của mô hình ARX với mô hình GM (1, 1). Mô hình lai cung cấp hiệu suất dự báo có độ chính xác cao.

Abdulsalam và cộng sự (2010) đã sử dụng phương pháp trung bình động [MA] để khám phá các mẫu, mối quan hệ và trích xuất giá trị của các biến từ cơ sở dữ liệu để dự đoán giá trị tương lai của các biến khác thông qua việc sử dụng dữ liệu chuỗi thời gian. Ưu điểm của phương pháp MA là một công cụ giúp giảm biến động và thu được xu hướng với độ chính xác khá cao. Kỹ thuật này đã chứng minh phương pháp dự báo số bằng cách sử dụng phân tích hồi quy với đầu vào là thông tin tài chính thu được từ hoạt động chứng khoán hàng ngày do sàn giao dịch chứng khoán Nigeria công bố.

Hsien-Lun và cộng sự (2010) đã sử dụng mô hình ARIMA và mô hình vector ARMA với phương pháp chuỗi thời gian mờ để dự báo. Phương pháp chuỗi thời gian mờ đặc biệt là mô hình heuristic thực hiện khả năng dự báo tốt hơn trong dự báo chu kỳ ngắn hạn. Mô hình ARIMA tạo ra các lỗi dự báo nhỏ trong khoảng thời gian thử

nghiệm dài hơn. Trong bài báo này, tác giả nghiên cứu xem độ dài của khoảng thời gian có ảnh hưởng đến khả năng dự báo của các mô hình hay không.

Agrawal và cộng sự (2010) đã trình bày một cách tiếp cận sáng tạo để đưa ra các quyết định trên thị trường chứng khoán bằng cách giảm thiểu rủi ro liên quan đến việc đầu tư. Hệ thống này đã sử dụng hệ thống suy luận mờ thần kinh thích ứng (ANFIS) để đưa ra quyết định dựa trên các chỉ báo kỹ thuật. Trong số các chỉ báo kỹ thuật khác nhau có sẵn, hệ thống đã sử dụng đường trung bình động có trọng số, độ phân kỳ và Chỉ số sức mạnh tương đối (RSI). Kara và cộng sự (2008) đã sử dụng mạng lưới thần kinh nhân tạo và Máy Vector hỗ trợ để dự đoán Sở giao dịch chứng khoán Istanbul. Kara và cộng sự (2008) đã sử dụng các mô hình có 10 chỉ báo kỹ thuật làm đầu vào cho mạng của họ. Kết quả cho thấy mô hình ANN được cải thiện rõ rệt so với mô hình SVM.

George và cộng sự (2011) đề xuất phân tích sóng dự đoán cổ phiếu dựa trên mô hình thần kinh mờ. Các kỹ thuật này được sử dụng để dự báo xu hướng giá cổ phiếu và kết quả thu được. Nguyên lý sóng Elliott được kết nối với dãy Fibonacci, dãy số Fibonacci bắt nguồn từ việc cộng hai số trước đó. Lý thuyết sóng của Elliott không thể liên tục giải thích thị trường một cách hoàn hảo nhưng những ước tính mờ nhạt về hành vi thị trường một cách chính xác sẽ cải thiện việc dự báo thị trường chứng khoán

Sureshkumar và cộng sự (2011) đã sử dụng các thuật toán và chức năng dự đoán để dự đoán giá cổ phiếu trong tương lai và so sánh hiệu suất của chúng. Kết quả phân tích cho thấy hàm hồi quy đẳng trương được sử dụng mang lại khả năng dự đoán giá cổ phiếu chính xác hơn các kỹ thuật hiện có khác. Ayodele và cộng sự (2012) đã sử dụng phương pháp lai để dự đoán giá cổ phiếu. Họ đã chứng minh rằng phương pháp kết hợp tốt hơn đáng kể so với các kỹ thuật khác. Budhani và cộng sự (2012) đã sử dụng mạng truyền lan truyền ngược để dự đoán thị trường chứng khoán. Quy trình đào tạo của họ giúp cải thiện tốt hơn khả năng dự đoán thị trường chứng khoán.

Devadoss và cộng sự (2013) đã đề xuất các mô hình ANN để dự đoán giá trị đóng cửa của sàn giao dịch chứng khoán Bombay. Đầu vào của mạng là cao, thấp, giá mở cửa, giá đóng cửa và khối lượng. Sai số bình phương trung bình gốc, độ lệch tuyệt đối trung bình và sai số phần trăm tuyệt đối trung bình được sử dụng làm chỉ báo hiệu suất cho mạng.

Wang và Choi (2014) đã đề xuất các kỹ thuật học máy như phân tích thành phần chính (PCA) để xác định các thành phần chính và máy vector hỗ trợ được sử dụng để phân loại cho diễn biến thị trường chứng khoán trong tương lai. Họ coi chỉ số giá cổ phiếu tổng hợp của Hàn Quốc (KOSPI) và chỉ số Hang seng (HSI) là dữ liệu chứng khoán cho nghiên cứu của họ. Masoud (2014) đã sử dụng mô hình ANN để dự đoán hướng đi của thị trường tài chính Libya. Sai số bình phương trung bình và sai số bình phương trung bình gốc được sử dụng làm chỉ báo hiệu suất. Lai và Lu (2014) đã sử dụng máy vector hỗ trợ và máy vector hỗ trợ bình phương nhỏ nhất để dự báo giá cổ phiếu. Yan và cộng sự (2014) đề xuất máy học cực đoan với chế độ phân rã theo kinh nghiệm để dự báo giá tài nguyên uranium.

2.3.2. Tại Việt Nam

Cho tới hiện tại, tại Việt Nam số nghiên cứu về việc dự đoán giá cổ phiếu còn nhiều hạn chế.

Luu và các cộng sự (2020) đã nghiên cứu và phát triển mô hình kết hợp Sequence-to-Sequence với LSTM và các thành phần thời gian cấu trúc để dự đoán giá cổ phiếu. Kết quả cho thấy mô hình này hiệu quả hơn so với các mô hình hiện có và mang lại lợi nhuận tích cực trong giao dịch tương lai.

Ta và Liu (2016) đã nghiên cứu sử dụng các kỹ thuật khai thác dữ liệu để khám phá nhiều khía cạnh của thị trường tài chính, bao gồm dự đoán chỉ số và giá cổ phiếu, quản lý rủi ro danh mục và phát hiện xu hướng. Nghiên cứu này điều tra tác động của sở hữu nước ngoài đối với biến động thị trường chứng khoán tại Việt Nam bằng cách sử dụng dữ liệu giao dịch hàng ngày của 100 cổ phiếu lớn trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) trong giai đoạn từ ngày 1 tháng 9 năm 2015 đến ngày 1 tháng 9 năm 2016. Các công ty thường được phân loại dựa trên hoạt động kinh doanh của họ, nhưng điều này không nhất thiết phản ánh cách giá trị thị trường của họ biến động. Thuật toán phân cụm K-mean và phương pháp phân cụm phân cấp được sử dụng để trực quan hóa phân tích về khối lượng giao dịch net, biến động giá và tỷ lệ biến động lợi nhuận. Dựa trên các mô hình hóa thị trường chứng khoán, chúng ta có thể quan sát tác động của vốn nước ngoài đối với biến động thị trường. Hơn nữa, những mô hình này cho thấy các hành vi đầu tư có thể giúp tối ưu hóa quản lý đầu tư danh mục.

Cần và cộng sự (2023) đã nghiên cứu này sử dụng các kỹ thuật khai thác dữ liệu để phân cụm các nhóm có liên quan chính trị bằng máy học. Mẫu dữ liệu trong nghiên cứu bao gồm các doanh nghiệp được niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh và Sở Giao dịch Chứng khoán Hà Nội, dữ liệu nghiên cứu trong giai đoạn từ 2015 đến 2020. Dữ liệu sử dụng trong nghiên cứu bao gồm tỷ lệ sở hữu của nhà nước, mức độ kết nối chính trị của các nhà lãnh đạo doanh nghiệp và các chỉ số tài chính trong báo cáo tài chính niêm yết của các doanh nghiệp. Nghiên cứu của tác giả đo lường mối quan hệ chính trị bằng thuật toán K-Means và sau đó so sánh kết quả của việc phân cụm K-Means với phương pháp truyền thống đo lường kết nối chính trị bằng hai giá trị 0 và 1, trong đó 0 là không có liên hệ chính trị và 1 là có liên hệ chính trị. Đồng thời, tác giả chạy ba nhóm để có cái nhìn sâu hơn. Các tác giả kết luận rằng phân cụm máy học sử dụng mô hình K-Means có thể thay thế phương pháp truyền thống. Các doanh nghiệp có liên hệ chính trị niêm yết trên HOSE và HNX mang lại nhiều lợi ích cho doanh nghiệp trong hoạt động đầu tư, trong việc tiếp cận nguồn lực cũng như vốn; tuy nhiên, các doanh nghiệp này có tác động tiêu cực đến hiệu suất kinh doanh. Các tác giả khuyến nghị rằng mức độ liên hệ chính trị vừa phải sẽ giúp doanh nghiệp đạt được hiệu suất tốt hơn

Nhìn chung, các nghiên cứu đều có số liệu khá cũ từ 2020 đổ về trước, thời kỳ chưa xảy ra Covid-19 đầy biến động. Và chưa có bài nào áp dụng kết hợp phân cụm và Deep Learning để dự đoán giá cổ phiếu dựa vào phân cụm các chỉ số tài chính.

CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Mẫu nghiên cứu

Mẫu nghiên cứu là 96 doanh nghiệp Việt Nam, được chia làm khoảng 20 ngành, mỗi ngành gồm 5 công ty, được niêm yết trên sàn chứng khoán HOSE trong giai đoạn từ năm 2019 đến 2023 và các chỉ số tài chính được lấy từ thư viện vnstock, phải đảm bảo các chỉ số tài chính trong thư viện tương ứng với các chỉ số trong báo cáo tài chính hợp nhất đã kiểm toán được công bố công khai của các doanh nghiệp. Các doanh nghiệp chỉ được chọn khi có đủ tất cả các chỉ số cần thiết để phục vụ cho việc tính toán, đồng thời phải có đầy đủ báo cáo tài chính đã được kiểm toán được công bố trong giai đoạn nghiên cứu. Tuy nhiên, trong quá trình chọn mẫu, còn tồn tại một vài nguyên nhân làm hạn chế số lượng mẫu của nhóm tác giả, như một số ngành ít các doanh nghiệp niêm yết, một số doanh nghiệp mới niêm yết nên còn thiếu các chỉ số tài chính trong một số năm.

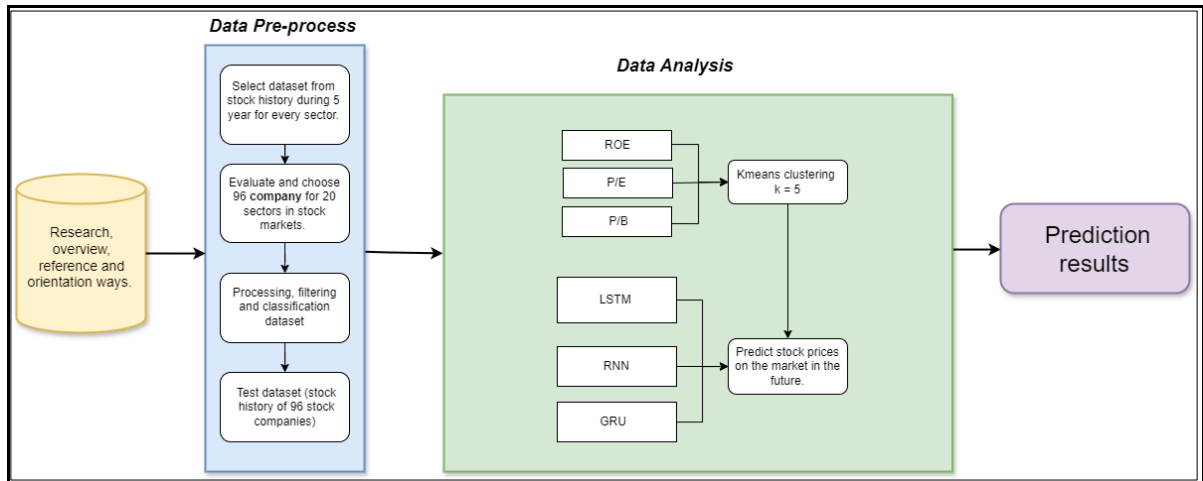
STT	Nhóm ngành	STT	Nhóm ngành
1	Bán lẻ	11	Xây dựng
2	Đầu tư công	12	Thực phẩm
3	Thép	13	Bưu chính - Viễn thông
4	Hàng Không	14	Truyền thông - Giải trí
5	Y tế	15	Cao su
6	Ngân hàng	16	Thủy hải sản
7	Chứng khoán	17	Dệt may - may mặc
8	Điện	18	Giao thông vận tải
9	Dầu khí	19	Bảo hiểm
10	Du lịch	20	Công nghệ

Bảng 2. Danh sách 20 nhóm ngành được lựa chọn nghiên cứu

3.2. Quy trình nghiên cứu

Như tên của đề tài nghiên cứu “**A hybrid approach combining clustering and deep learning for stock price prediction**”, nhóm nghiên cứu đã tiến hành phân cụm dựa vào việc tính trung bình các chỉ số ROE, P/E, P/B của 96 công ty trong khoảng thời gian 5 năm, sau đó sử dụng tiếp thuật toán K-means để tiến hành chia các nhóm công ty có chỉ số trung bình gần nhau thành 1 cụm. Sau khi đã phân thành 5 cụm, nhóm nghiên cứu tiến hành dự đoán giá cổ phiếu bằng cách lấy giá cổ phiếu mỗi ngày của mỗi công ty trên thư viện vnstock và sử dụng mô hình Deep Learning kết hợp CNN, LSTM, GRU để dự đoán giá cổ phiếu trong tương lai. Cuối cùng, nhằm đánh giá hiệu năng, ưu nhược điểm của mô hình Deep Learning cũng như tính khả quan của

nguyên cứu, nhóm nghiên cứu đã tiến hành so sánh kết quả dự đoán của mô hình Deep Learning với một mô hình khác là mô hình hồi quy tuyến tính đa biến. Tuy nhiên, điều quan trọng để nghiên cứu đề tài này, nhóm nghiên cứu đã phải tham khảo qua nhiều nguồn tài liệu, các bài nghiên cứu trước đó để có cái nhìn tổng quan nhất sau đó mới đưa ra phương pháp làm như trên.



Hình 8. Bốn giai đoạn của phương pháp nghiên cứu

- (1) Giai đoạn 1: Nhóm tìm hiểu các cách thức nghiên cứu, tổng quan về nghiên cứu, tham khảo và đưa ra định hướng nghiên cứu phù hợp với mục tiêu của nhóm.
- (2) Giai đoạn 2: Tiến hành thu thập dữ liệu. Dữ liệu được thu thập là lịch sử giá theo ngày giai đoạn 5 năm từ năm 2019 đến năm 2022 của mỗi mã chứng khoán ứng với mỗi công ty. Nhóm chọn ra 96 công ty đến từ khoảng 20 ngành khác nhau để thiết lập thành một tập dữ liệu. Sau đó, tập dữ liệu này sẽ được đưa vào bước tiền xử lý, lọc và phân loại dữ liệu để tìm ra các số liệu bị lỗi, các dữ liệu bị thiếu cụ thể trong nghiên cứu này may mắn không có biến nào bị lỗi hay thiếu. Cuối cùng, nhóm bắt đầu thực hiện thử nghiệm tập dữ liệu.
- (3) Giai đoạn 3: Nhóm chọn ra 3 chỉ số tài chính có ảnh hưởng lớn đến lịch sử giá của mã chứng khoán bao gồm tỷ suất sinh lời trên vốn chủ sở hữu (ROE), chỉ số đánh giá mối quan hệ giữa giá thị trường của cổ phiếu và thu nhập trên mỗi cổ phiếu (P/E), chỉ số so sánh giá của cổ phiếu với giá trị sổ sách của cổ phiếu (P/B). Sau đó, tính giá trị trung bình của 3 chỉ số này trong giai đoạn 5 năm và sử dụng thuật toán phân cụm K-Means để phân 96 công ty thành 5 cụm khác nhau. Sau khi đã phân thành 5 cụm, nhóm

sử dụng mô hình Deep Learning kết hợp CNN, LSTM, GRU để dự đoán giá cổ phiếu trong tương lai.

(4) Giai đoạn 4: Đưa ra kết quả so sánh hai mô hình và tiến hành dự đoán kết quả.

3.3. Phương pháp xử lý thông tin

3.3.1. Phương pháp thu thập dữ liệu

Nhóm sử dụng bộ dữ liệu của 96 công ty thuộc 20 ngành khác nhau được niêm yết trên sàn chứng khoán HOSE lấy từ thư viện vnstock trong khoảng thời gian 5 năm (2019 - 20). Dữ liệu này cung cấp cho nhóm những thông tin cần thiết và quan trọng của công ty như các chỉ số tài chính, phù hợp với mục tiêu nghiên cứu của nhóm là dự đoán giá chứng khoán trong tương lai. Tập dữ liệu này giúp dự đoán liệu nhà đầu tư có đưa ra quyết định chọn mua - bán mã chứng khoán này hay không dựa trên các tính năng dữ liệu về công ty, về dữ liệu liên quan đến kinh tế - xã hội.

3.3.2. Phương pháp phân tích dữ liệu

Nhóm sử dụng các phương pháp tính phù hợp để tính giá trị trung bình của các chỉ số ROE, P/E, P/B. Tiếp theo đó, tính toán số cụm cần gộp bằng phương pháp Elbow, và tiến hành phân cụm bằng thuật toán K-Means.

Sau khi thực hiện các phương pháp và thu được kết quả về các nhóm công ty, nhóm sẽ thực hiện thống kê mô tả các cụm, diễn giải và phân tích kết quả. Sau đó tiến hành huấn luyện mô hình Deep Learning: RNN, LSTM, CNN (Huấn luyện mô hình riêng cho từng nhóm dữ liệu được phân cụm, Dự đoán giá cổ phiếu cho từng nhóm).

CHƯƠNG 4. KẾT QUẢ VÀ ĐÁNH GIÁ

4.1. Thu thập và tính toán dữ liệu

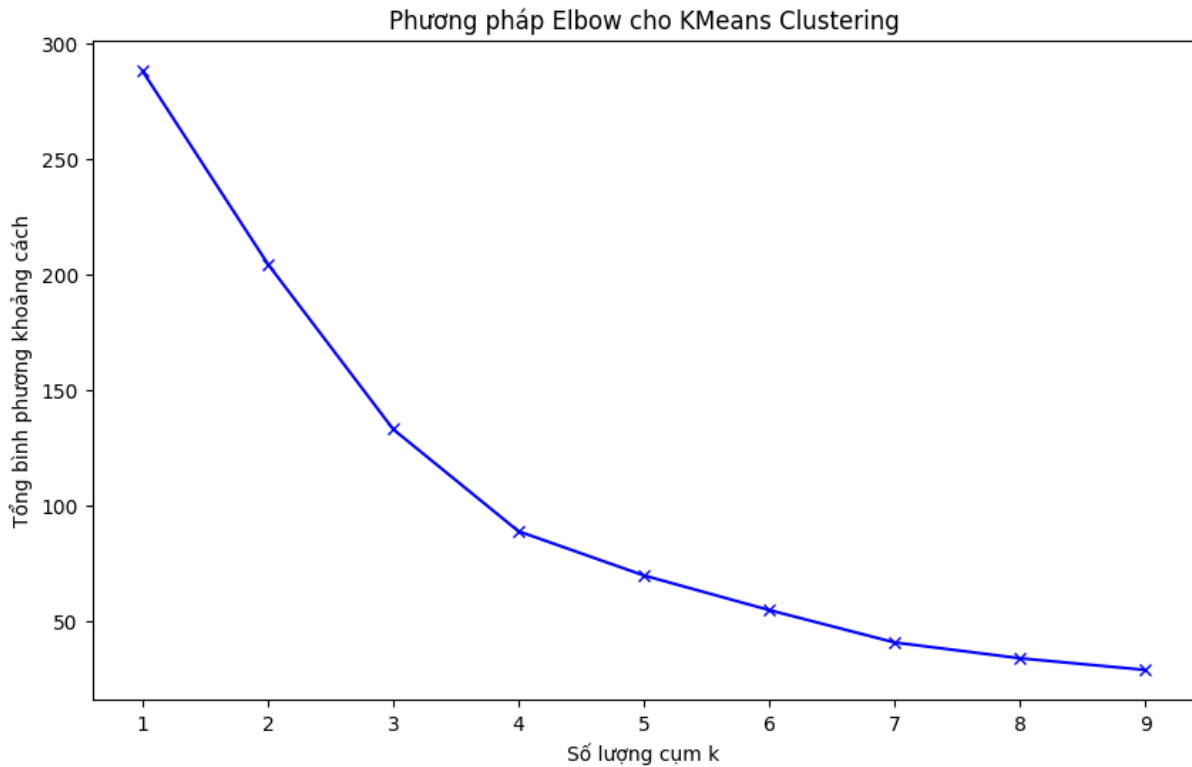
Các chỉ số được tính trung bình trong vòng 5 năm, bằng cách tính trung bình các chỉ số này trong một khoảng thời gian dài như vậy, sẽ phản ánh được xu hướng kinh doanh dài hạn của doanh nghiệp hoặc thị trường, thay vì tập trung vào biến động ngắn hạn. Điều này giúp đánh giá được giá trị cổ phiếu từ một góc nhìn dài hạn và giúp đưa ra quyết định đầu tư có sự tự tin hơn.

	Ticker	P/E	P/B	ROE
0	FRT	20.325	5.90	0.0875
1	DGW	15.360	3.48	0.2736
2	MWG	92.120	3.66	0.2224
3	SID	43.700	0.56	0.0222
4	TMC	16.920	0.74	0.0568
	...			
91	FPT	16.240	3.96	0.2572
92	ELC	19.380	1.02	0.0502
93	SGT	46.420	1.70	0.0414
94	CMG	23.380	2.32	0.1120
95	SAM	217.320	1.14	0.0232

Bảng 3. Kết quả tính toán

4.2. Phân cụm

- Phương pháp Elbow để phân cụm K-means



Hình 9. Kết quả phân cụm

Nhóm đã tiến hành kiểm nghiệm số cụm từ 1 đến 9 để đánh giá sự độ lệch của các giá trị so với tâm cụm. Độ lệch được tính bằng cách tính bình phương trung bình của khoảng cách từ các giá trị trong cụm đến giá trị trung tâm của cụm đó. Khi độ lệch càng lớn, điều này cho thấy độ tập trung của cụm thấp hơn, có xu hướng lệch tâm cao hơn, và khó xác định được đặc điểm đồng nhất và tính đại diện của cụm đó.

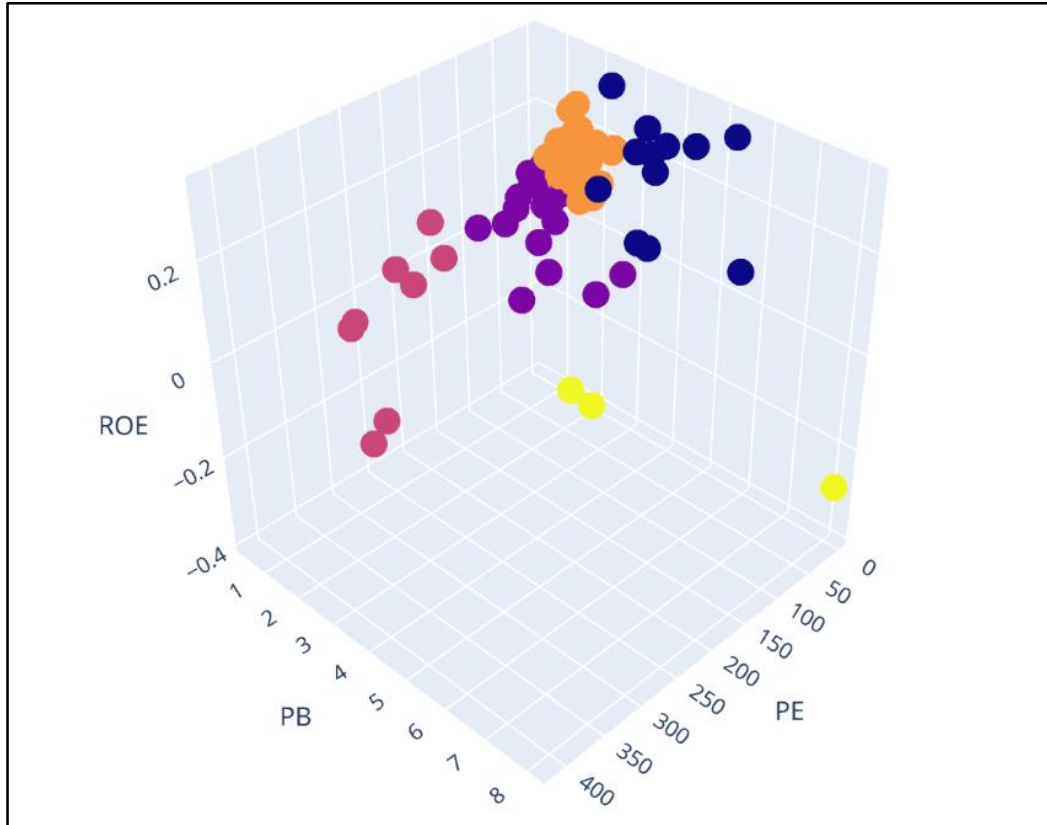
Kết quả trong Hình 9 cho thấy đường diễn tả độ lệch của các cụm giảm dần từ số cụm 1 đến 9. Tại điểm $K = 5$ (số cụm là 5), sự khác biệt về độ tập trung rõ rệt, và từ giá trị này, độ lệch thay đổi xu hướng và giảm dần đến mức rất nhỏ.

Lý do nhóm không chọn các giá trị cụm lớn hơn là vì khi số cụm tăng lên, độ lệch tâm giảm đi. Tuy nhiên, điều này sẽ tạo ra nhiều cụm hơn và dẫn đến sự phân khúc hóa cao hơn giữa các công ty. Điều này làm cho quá trình xử lý trở nên khó khăn hơn rất nhiều. Do đó, số cụm 5 đã được chọn làm số cụm phù hợp để phân cụm trong trường hợp này.

	Ticker	PE	PB	ROE	Cluster
0	FRT	20.325	5.90	0.0845	0
1	DGW	15.360	3.48	0.2736	0
2	MWG	92.120	3.66	0.2224	0
3	SID	43.700	0.56	0.0222	1

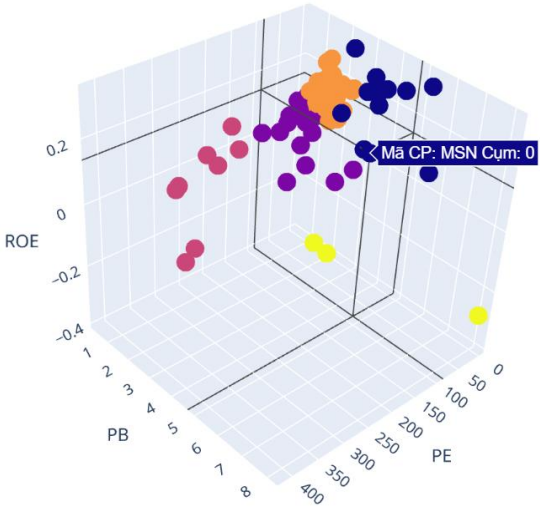
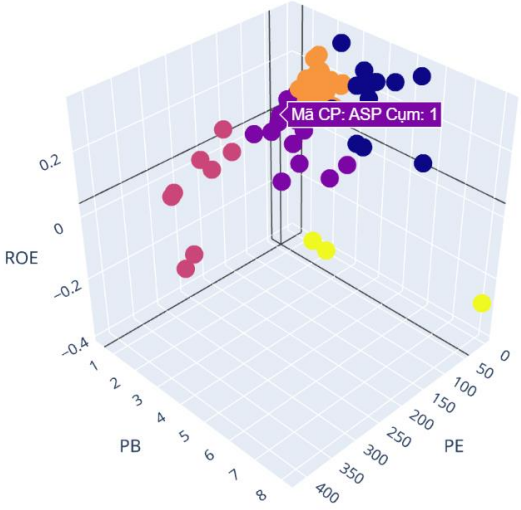
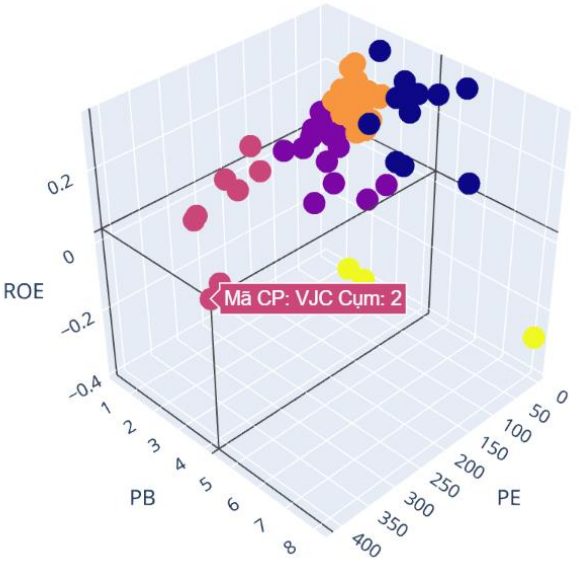
4	TMC	16.920	0.74	0.0568	1
...
91	FPT	16.240	3.96	0.2572	0
92	ELC	19.380	1.02	0.0502	1
93	SGT	46.420	1.70	0.0414	1
94	CMG	23.380	2.32	0.1120	3
95	SAM	217.320	1.14	0.0232	2

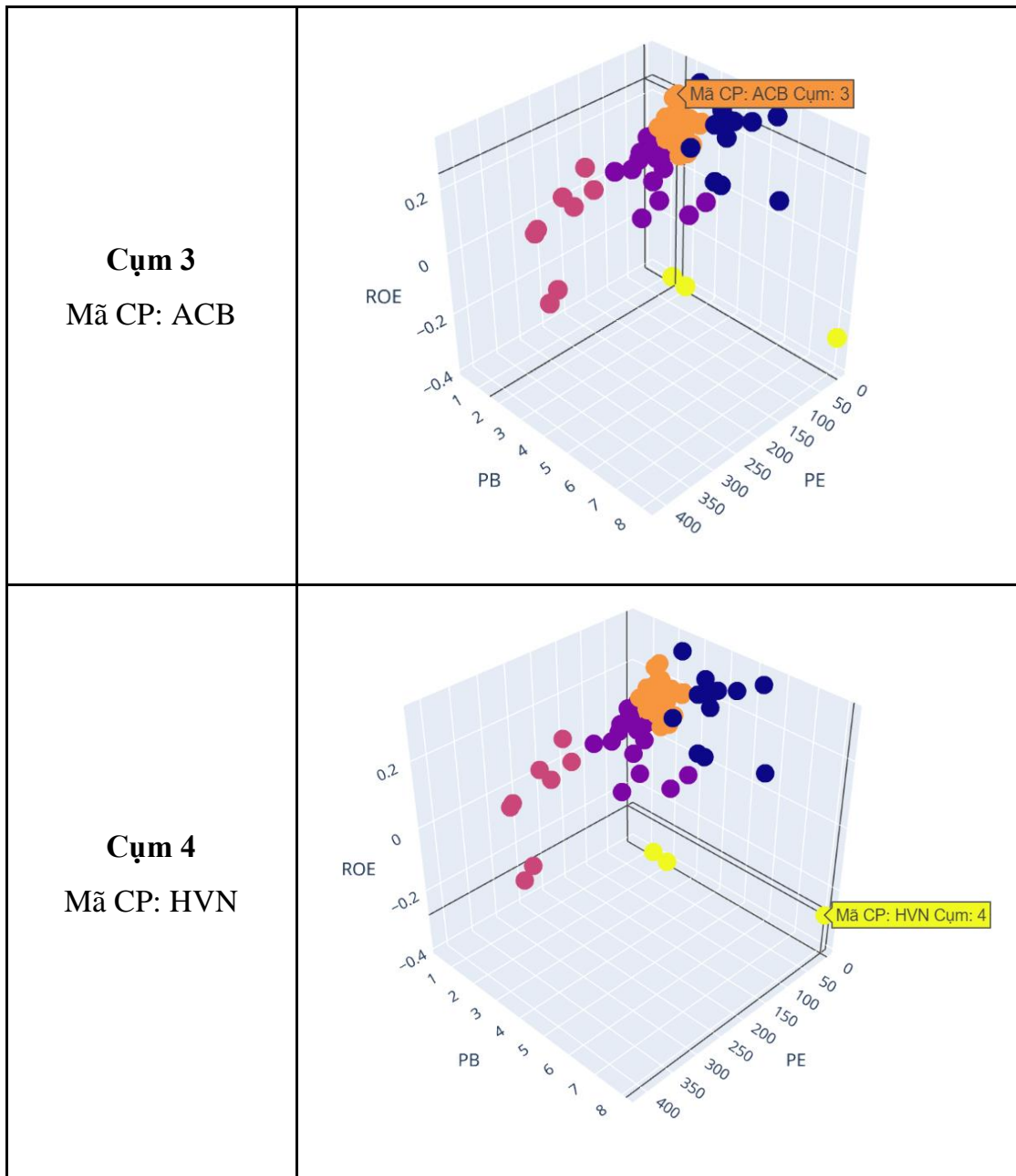
Bảng 4. Phân tích và dự đoán cụm cho mỗi công ty



Hình 10. Sơ đồ 3D biểu thị 5 cụm K-Means

Từ biểu đồ phân cụm, có thể thấy rằng các cụm được xác định là các đại diện cho các nhóm cổ phiếu với các đặc điểm về chỉ số ROE, P/E, P/B tương tự. Trong đó, cụm 4 ít nhất với 3 công ty và cụm 2 có nhiều công ty có đặc điểm tài chính tương tự nhất.

<p>Cụm 0 Mã CP: MSN</p>	
<p>Cụm 1 Mã CP: ASP</p>	
<p>Cụm 2 Mã CP: VJC</p>	



Bảng 5. Chi tiết 5 cụm K- Means

4.3. Đánh giá mô hình

Từ thư viện vnstock, nhóm nghiên cứu thu thập được số liệu về giá cổ phiếu của các công ty trong danh sách nghiên cứu theo từng ngày từ năm 2019 đến hết năm 2023. Bảng số liệu được lấy như bảng ... dưới đây là bảng dữ liệu lấy cho cụm 0.

	Time	Open	High	Low	Close	Volume	Ticker
0	02-01-2019	18130	18540	18130	18430	54890	GMC
1	03-01-2019	17860	18130	17500	17610	158460	GMC
2	04-01-2019	16930	17860	16930	17860	24090	GMC
3	07-01-2019	18100	18130	17710	18020	29620	GMC
4	08-01-2019	17990	18130	17990	18020	8440	GMC

				...			
1208	25-12-2023	7750	8000	7750	7880	800	GMC
1209	26-12-2023	7500	7920	7500	7920	1000	GMC
1210	27-12-2023	7650	7910	7650	7910	700	GMC
1211	28-12-2023	7970	8000	7970	7990	22600	GMC
1212	29-12-2023	7990	7990	7990	7990	1500	GMC

Bảng 6. Dữ liệu được lấy từ thư viện Vnstock của công ty GMC

Việc đánh giá hiệu suất của mô hình dự đoán giá cổ phiếu được thực hiện thông qua việc phân tích các chỉ số thống kê.

- MAE (Mean Absolute Error) là một phép đo được sử dụng để đánh giá hiệu suất của một mô hình hồi quy, nó đo lường trung bình của sự chênh lệch tuyệt đối giữa các giá trị dự đoán và các giá trị thực tế, với giá trị càng thấp cho thấy hiệu suất mô hình càng tốt.
- MAPE (Mean Absolute Percentage Error) là phép đo được sử dụng để đánh giá hiệu suất của một mô hình dự báo, nó đo lường trung bình của tỷ lệ phần trăm tuyệt đối giữa các giá trị dự đoán và các giá trị thực tế. Đối với MAPE, càng giảm giá trị của nó thì mô hình dự báo càng tốt. Điều này có nghĩa là mô hình dự báo có MAPE thấp hơn có xu hướng dự đoán các giá trị gần hơn với giá trị thực tế hơn so với các mô hình có MAPE cao hơn.
- MSE (Mean Squared Error) là một phép đo thường được sử dụng để đánh giá hiệu suất của một mô hình dự báo, đặc biệt là trong các bài toán hồi quy. MSE tính toán trung bình của bình phương của sự chênh lệch giữa các giá trị dự đoán và các giá trị thực tế. Càng giảm giá trị của MSE, mô hình dự báo càng chính xác, MSE tăng khi sự chênh lệch giữa các giá trị dự đoán và giá trị thực tế tăng lên. Do đó, một MSE thấp hơn cho thấy rằng mô hình dự báo đang có sự phù hợp tốt hơn với dữ liệu huấn luyện.
- RMSE (Root Mean Squared Error) có ý nghĩa tương tự như MSE, tuy nhiên, việc lấy căn bậc hai của MSE giúp chuyển đổi đơn vị trở lại gần với đơn vị gốc của dữ liệu, giúp dễ dàng so sánh và hiểu rõ hơn về mức độ chênh lệch giữa dự đoán và giá trị thực tế. Điều này làm cho RMSE thường được ưa chuộng hơn MSE trong việc diễn giải kết quả của mô hình.
- R2 được sử dụng để đánh giá hiệu suất của một mô hình hồi quy. Nó có kết quả trong khoảng từ 0 đến 1 đo lường mức độ biến thiên của biến phụ thuộc mà mô

hình có thể giải thích so với tổng biến thiên của biến phụ thuộc. R2 càng cao, mô hình càng tốt ở việc giải thích biến thiên của dữ liệu.

	Cụm 0			Cụm 1			Cụm 2			Cụm 3			Cụm 4		
	LSTM	GRU	RNN	LSTM	GRU	RNN	LSTM	GRU	RNN	LSTM	GRU	RNN	LSTM	GRU	RNN
TẬP TRAIN															
MAE	0.02	0.02	0.02	0.02	0.03	0.03	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.01	0.02
MAPE	82.39	79.92	83.44	84.61	81.57	84.49	157.75	157.28	154.03						
MSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMSE	0.03	0.03	0.03	0.03	0.04	0.04	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02
R2	0.99	0.98	0.99	0.98	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
TẬP TEST															
MAE	0.01	0.01	0.01	0.02	0.03	0.03	0.01	0.00	0.01	0.02	0.01	0.01	0.01	0.01	0.03
MAPE										33.93	32.01	32.21	33.23	33.63	38.71
MSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMSE	0.01	0.01	0.02	0.03	0.04	0.04	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.03
R2	0.92	0.84	0.81	0.89	0.85	0.81	0.98	0.98	0.95	0.94	0.97	0.97	0.92	0.92	0.52

Bảng 7. Bảng chỉ số hiệu suất đánh giá mô hình

Các mô hình học sâu như LSTM, GRU và RNN đã được chứng minh là có khả năng mạnh mẽ trong việc mô hình hóa dữ liệu chuỗi thời gian phức tạp, điều này là cần thiết cho việc dự đoán giá cổ phiếu. Tuy nhiên, một thách thức lớn trong việc dự đoán giá cổ phiếu là sự không ổn định và tính không tuyến tính của thị trường. Điều này đòi hỏi các mô hình phải có khả năng tổng quát hóa tốt và thích ứng với sự thay đổi liên tục của thị trường.

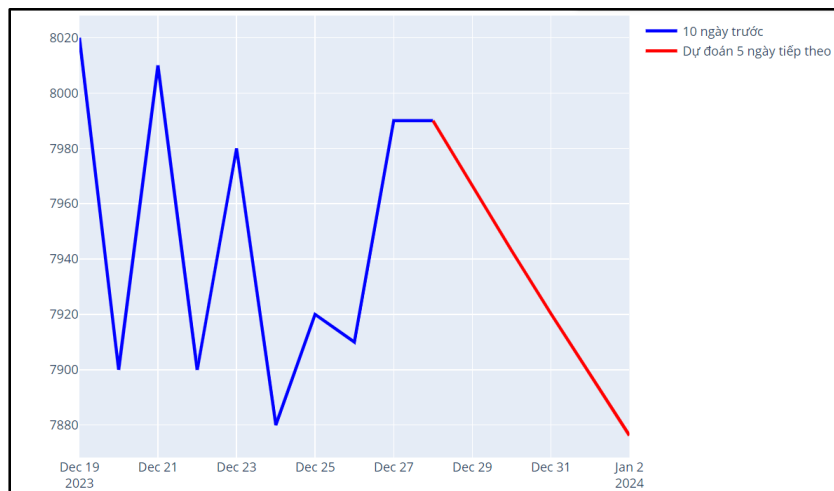
Kết quả từ bảng so sánh cho thấy rằng mỗi mô hình có những ưu và nhược điểm riêng. Mô hình LSTM cho thấy khả năng tổng quát hóa tốt trong cả giai đoạn huấn luyện và kiểm tra, với giá trị R2 cao luôn là 0.99 ở tập train của tất cả các cụm và các

chỉ số hiệu suất thấp luôn trong khoảng 0.01 hoặc 0.02. Điều này làm nổi bật khả năng của LSTM trong việc nắm bắt các mối quan hệ dài hạn trong dữ liệu, một yếu tố quan trọng cho việc dự đoán giá cổ phiếu. Theo quan sát từ bảng chỉ số hiệu suất cũng thấy ở cụm 1 và 4 chọn mô hình LSTM sẽ đạt hiệu suất lớn nhất.

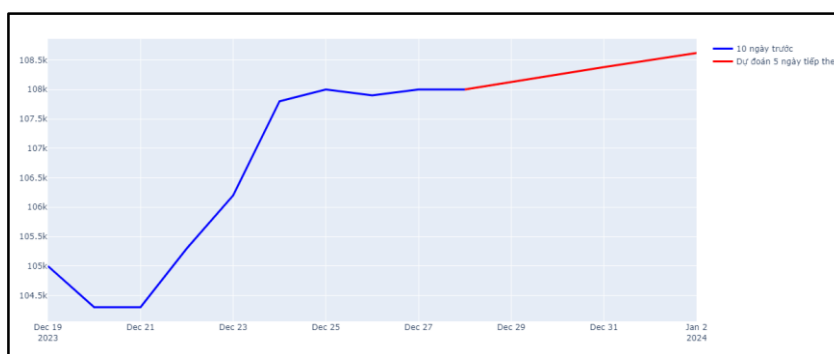
Mặt khác, mô hình GRU, mặc dù có hiệu suất tương tự trong giai đoạn huấn luyện, lại cho thấy sự biến động lớn hơn trong giai đoạn kiểm tra. Điều này có thể chỉ ra rằng GRU có thể phản ứng mạnh mẽ hơn với sự biến động ngắn hạn của dữ liệu, điều này có thể hữu ích trong việc dự đoán giá cổ phiếu trong ngắn hạn. Dựa trên các chỉ số hiệu suất nên các công ty ở cụm 0, 2, 3 sẽ phù hợp nhất với mô hình GRU.

Cuối cùng, mô hình RNN, mặc dù có khả năng học nhanh và yêu cầu ít tài nguyên tính toán hơn, lại cho thấy sự giảm hiệu suất trong giai đoạn kiểm tra, đặc biệt là với giá trị R^2 là 0.52 khá thấp cho Cụm 4. Điều này có thể chỉ ra rằng RNN có thể không phù hợp cho việc dự đoán giá cổ phiếu trong dài hạn hoặc trong các tình huống có sự biến động lớn.

4.4. Dự đoán



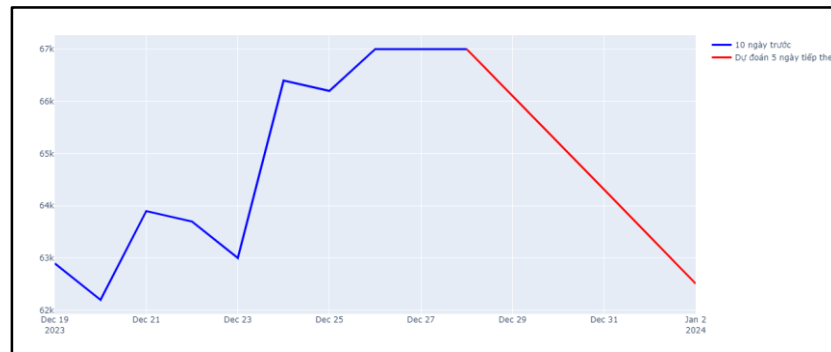
Hình 11. Cụm 0



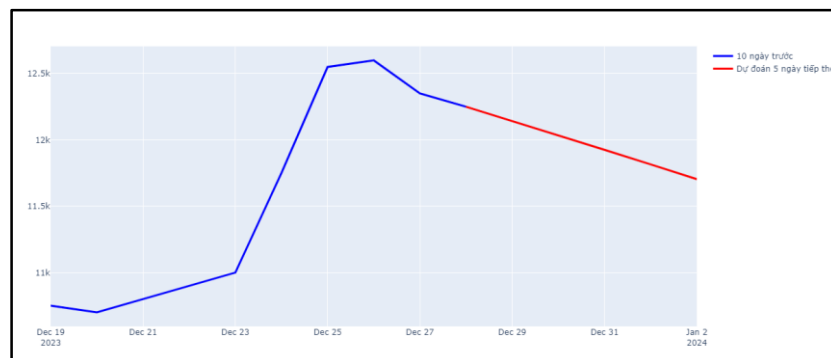
Hình 12. Cụm 1



Hình 13. Cụm 2



Hình 14. Cụm 3



Hình 15. Cụm 4

Kết quả cho thấy trong 5 ngày tiếp theo, giá cổ phiếu ở cụm 1 và 2 có xu hướng tăng; cụm 0, 3 và 4 có xu hướng giảm. Đây sẽ là những thông tin vô cùng hữu ích cho các nhà đầu tư.

CHƯƠNG 5. KẾT LUẬN VÀ ĐỀ XUẤT

5.1. Kết luận

Như vậy, bài nghiên cứu này đã giải quyết được vấn đề đặt ra đó chính là dự đoán giá cổ phiếu theo ngày trong tương lai trên thị trường chứng khoán, dựa vào việc tính toán và phân tích các chỉ số như ROE, P/B, P/E,... Dựa trên các mô hình hợp nhất hai giai đoạn hiện có trong tài liệu là phân cụm và dự đoán, trong đó phương pháp phân cụm K-Means được áp dụng để phân cụm phổ biến. Ngoài ra, nhóm cũng đã được áp dụng để cải thiện độ chính xác của dự đoán thông qua các mô hình Deep Learning trong đó GRU và LSTM cho hiệu suất khá tốt còn RNN có phần kém hơn. Cuối cùng, một mô hình dự đoán kết hợp, kết hợp cả phương pháp phân cụm k-mean đã được đề xuất. Kết quả thử nghiệm trên 96 công ty chứng khoán chứng minh rằng mô hình dự đoán đạt được độ chính xác dự đoán về giá cổ phiếu. Việc phân cụm k-mean trên các chỉ báo kỹ thuật chứng khoán có thể nâng cao hơn nữa độ chính xác dự đoán của quá trình học tập tổng hợp.

5.2. Hạn chế và đề xuất hướng phát triển

5.2.1. Hạn chế

Hạn chế thứ nhất, một trong những hạn chế chính của việc dự đoán giá cổ phiếu là phụ thuộc vào chất lượng và tính đáng tin cậy của dữ liệu. Dữ liệu được thu thập từ một thị trường cận biên như Việt Nam và dữ liệu lịch sử về giá cổ phiếu còn khá ít so với các thị trường phát triển như Mỹ hay Trung Quốc, nên do đó, việc dự đoán có thể còn mang tính chủ quan, chưa tối ưu hoàn toàn. Hơn nữa, dữ liệu có thể bị nhiễu bởi các yếu tố ngoại vi không liên quan, làm mất đi tính đáng tin cậy của kết quả dự đoán.

Hạn chế thứ hai, một vấn đề khác là tính chất không chắc chắn và phức tạp của thị trường tài chính. Thị trường có thể phản ứng mạnh mẽ và không lý giải được đối với các sự kiện không lường trước, như biến động thị trường toàn cầu, thay đổi chính sách kinh tế, hoặc các sự kiện tự nhiên. Những yếu tố này có thể làm biến đổi giá cổ phiếu một cách không đoán trước được và gây ảnh hưởng tiêu cực đến hiệu suất của mô hình dự đoán.

Hạn chế thứ ba, việc không có khả năng giải thích rõ ràng cho các dự đoán cũng là một vấn đề. Mặc dù mô hình có thể đạt được hiệu suất dự đoán cao, nhưng khả năng

giải thích của chúng thường bị hạn chế. Điều này có thể gây khó khăn cho các nhà đầu tư trong việc hiểu và tin tưởng vào các dự đoán được đưa ra.

Cuối cùng, việc dự đoán giá cổ phiếu luôn mang theo rủi ro đầu tư. Ngay cả khi sử dụng các phương pháp dự đoán tiên tiến nhất, không có phương pháp nào đảm bảo rằng các dự đoán sẽ chính xác hoặc mang lại lợi nhuận. Đối mặt với những thách thức này, việc nghiên cứu và phát triển các phương pháp dự đoán ngày càng tiên tiến và cẩn thận là điều cần thiết.

5.2.2. Hướng phát triển

Thứ nhất, việc cải thiện khả năng phân loại dữ liệu không cấu trúc cũng là một ưu tiên. Phương pháp clustering có thể được tinh chỉnh để phù hợp hơn với tính đa dạng của dữ liệu thị trường tài chính. Các phương pháp mới có thể tập trung vào việc xác định các cụm dữ liệu một cách hiệu quả và linh hoạt hơn, bao gồm việc sử dụng các kỹ thuật clustering không phụ thuộc vào dữ liệu, như Gaussian Mixture Models (GMMs) hoặc Hierarchical Clustering.

Thứ hai, việc tăng cường khả năng giải thích của mô hình là một yếu tố quan trọng cần được xem xét. Mặc dù các mô hình Deep Learning có thể đạt được hiệu suất dự đoán cao, nhưng khả năng giải thích kết quả của chúng thường bị hạn chế. Điều này có thể gây khó khăn cho các nhà đầu tư và quản lý cổ phiếu trong việc hiểu và tin tưởng vào các dự đoán. Do đó, việc phát triển các phương pháp để giải thích kết quả của mô hình, như sử dụng các phương pháp như LIME (Local Interpretable Model-agnostic Explanations) hoặc SHAP (SHapley Additive exPlanations), có thể là một hướng đi quan trọng.

Một khía cạnh quan trọng khác là quản lý rủi ro. Việc phát triển các phương pháp quản lý rủi ro đúng mức và hiệu quả có thể giúp giảm thiểu các rủi ro đầu tư và tăng cường lợi nhuận. Điều này có thể bao gồm việc sử dụng kỹ thuật diversification, optimization, và dynamic hedging để giảm thiểu rủi ro đầu tư và tối ưu hóa lợi nhuận.

Cuối cùng, các phương pháp mới cần được kiểm định và đánh giá một cách kỹ lưỡng trên dữ liệu lịch sử và thử nghiệm trên dữ liệu thực tế để đảm bảo tính hiệu quả và tin cậy. Điều này đòi hỏi việc sử dụng các phương pháp đánh giá hiệu suất như backtesting và stress testing để đánh giá khả năng dự đoán của mô hình trong các tình huống khác nhau và dưới áp lực khác nhau.

Tóm lại, các nghiên cứu tiếp theo có thể tập trung vào việc tăng cường tính linh hoạt, hiệu quả và giải thích của phương pháp kết hợp giữa clustering và deep learning để dự đoán giá cổ phiếu. Điều này sẽ đóng góp vào việc cải thiện quy trình ra quyết định đầu tư và quản lý rủi ro trong thị trường tài chính.

TÀI LIỆU THAM KHẢO

- Abdulsalam, S. O., Adewole, K. S., & Jimoh, R. G. (2011). Stock trend prediction using regression analysis—a data mining approach.
- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1), 147-169.
- Adebisi, A. A., Ayo, C. K., Adebisi, M., & Otokiti, S. O. (2012). Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, 3(1).
- Agrawal, S., Jindal, M., & Pillai, G. N. (2010, March). Momentum analysis based stock market prediction using adaptive neuro-fuzzy inference system (anfis). In *Proceedings of the International Multi Conference of Engineers and Computer Scientists* (Vol. 1, pp. 2078-0958).
- Akinwale Adio, T., Arogundade, O. T., & Adekoya Adebayo, F. (2009). Translated Nigeria stock market prices using artificial neural network for effective prediction. *Journal of theoretical and Applied Information technology*, 1(1), 36-43.
- Anh, D.L.T. and Gan, C. (2021), "The impact of the COVID-19 lockdown on stock market performance: evidence from Vietnam", *Journal of Economic Studies*, Vol. 48 No. 4, pp. 836-851.
- Anwar, Y., Rahmalia, L. (2019). The effect of return on equity, earning per share and price earning ratio on stock prices. *The Accounting Journal of BINANIAGA*. 4 (1): 57-66.
- Arslan, M., & Zaman, R.U. (2014). Impact of Dividend Yield and Price Earnings Ratio on Stock Returns: A Study Non-Financial listed Firms of Pakistan. *Research Journal of Finance and Accounting*, 5, 68-74.
- Arslan, M., Zaman, R., & Phil, M. (2014). Impact of dividend yield and price earnings ratio on stock returns: A study non-financial listed firms of Pakistan. *Research Journal of Finance and Accounting*, 5(19), 68-74.
- Asraf, A., & Desda, M. M. (2020). Analysis of the effect of operating leverage and financial leverage on companies profitability listed on Indonesia Stock Exchange. *Ilomata International Journal of Management*, 1(2), 45-50.

Atsalakis, G. S., & Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert systems with Applications*, 36(7), 10696-10707.

Atsalakis, G. S., Dimitrakakis, E. M., & Zopounidis, C. D. (2011). Elliott Wave Theory and neuro-fuzzy systems, in stock market prediction: The WASP system. *Expert Systems with Applications*, 38(8), 9196-9206.

Azzopardi, A. M. (2006). An analysis of the price/book ratio of two maltese listed companies. *Bank of Valletta Review*, 34.

Budhani, N., Jha, C. K., & Budhani, S. K. (2012). Application of neural network in analysis of stock market prediction. *International Journal of Computer Science & Engineering Technology (IJCSET)*, 3(4), 61-68.

Byeon, W., Breuel, T. M., Raue, F., & Liwicki, M. (2015). Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3547-3555).

Capaul, C., Rowley, I., & Sharpe, W. F. (1993). International value and growth stock returns. *Financial Analysts Journal*, 49(1), 27-36.

Chandar, S. K. (2019). Stock market prediction using subtractive clustering for a neuro fuzzy hybrid approach. *Cluster Computing*, 22(Suppl 6), 13159-13166.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273-297.

Devadoss, A. V., & Ligor, T. A. A. (2013). Stock prediction using artificial neural networks. *International Journal of Data Mining Techniques and Applications*, 2(1), 283-291.

Du, X., Cai, Y., Wang, S., & Zhang, L. (2016, November). Overview of deep learning. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (pp. 159-164). IEEE.

Fahim, A. M., Salem, A. M., Torkey, F. A., & Ramadan, M. (2006). An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, 7, 1626-1633.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2), 427-465.

- Fama, E. F., & French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The journal of finance*, 50(1), 131-155.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.
- Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5), 826-834.
- Garza-Gómez, X. (2001). The information content of the book-to-market ratio. *Financial Analysts Journal*, 57(6), 78-95.
- Ghaeli, M. R. (2016). Price-to-earnings ratio: A state-of-art review. *Accounting*, 3(2), 131-136.
- Gottwald, R. (2012). The use of the P/E ratio to stock valuation. *European Grants Projects Journals*, 2012, 21-24.
- Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert systems with Applications*, 33(1), 171-180.
- Hebb, D. O. (1949). The organization of behavior. *rJ Appl Behav Anal*, 25, 575.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8), 2554-2558.
- Huang, G. B., Lee, H., & Learned-Miller, E. (2012, June). Learning hierarchical representations for face verification with convolutional deep belief networks. In 2012 IEEE conference on computer vision and pattern recognition (pp. 2518-2525). IEEE.
- Huang, K. Y., & Jane, C. J. (2009). A hybrid model for stock market forecasting and portfolio selection based on ARX, grey system and RS theories. *Expert systems with applications*, 36(3), 5387-5392.
- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304.
- Kamalov, F., Smail, L., & Gurrib, I. (2020). Stock price forecast with Deep Learning. 2020 International Conference on Decision Aid Sciences and Application (DASA).

Kamar, K. (2017). Analysis of the effect of return on equity (ROE) and debt to equity ratio (DER) on stock price on cement industry listed in Indonesia stock exchange (IDX) in the year of 2011-2015. *IOSR Journal of Business and Management*, 19(05), 66-76.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Applications*, 38(5), 5311-5319.

Kim, S., Park, B., Song, B. S., & Yang, S. (2016). Deep belief network based statistical feature learning for fingerprint liveness detection. *Pattern Recognition Letters*, 77, 58-65.

Kontschieder, P., Fiterau, M., Criminisi, A., & Buló, S. R. (2015). Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision* (pp. 1467-1475).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lai, L., & Liu, J. (2014). Support vector machine and least square support vector machine stock forecasting models. *Computer Science and Information Technology*, 2(1), 30-39.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5), 421-436.

Li, M., Zhu, Y., Shen, Y., & Angelova, M. (2023). Clustering-enhanced stock price prediction using Deep Learning. *World Wide Web*, 26(1), 207–232.

Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1805-1812).

- Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14 (pp. 21-37). Springer International Publishing.
- Luu, Q., Nguyen, S., & Pham, U. (2020). Time series prediction: A combination of Long Short-Term Memory and structural time series models. *Science & Technology Development Journal: Economics- Law & Management*, 4(1), 500-515.
- Masoud, N. (2014). Predicting direction of stock prices index movement using artificial neural networks: The case of Libyan financial market. *British Journal of Economics, Management & Trade*, 4(4), 597-619.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115-133.
- Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics* (pp. 63-67). Ieee.
- Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, 72(2011), 1-19.
- Ngoc, C. D., Huy, T. P., Thi Cam, T. T., Thi My, T. L., Thi Thuy, H. N., & Hai, M. N. (2023). Political affiliate clustering with machine learning in Vietnam Stock Exchange Market. *Journal of International Commerce, Economics and Policy*, 14(03).
- O'Connor, N., & Madden, M. G. (2005, December). A neural network approach to predicting stock exchange movements using external factors. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 64-77). London: Springer London.
- Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
- Qing He, Junyi Liu, Sizhu Wang & Jishuang Yu (2020) The impact of COVID-19 on stock markets, *Economic and Political Studies*, 8:3, 275-288

- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- Shibao, S., & Keyun, Q. (2007). Research on modified k-means data cluster algorithm. IS Jacobs and CP Bean, "Fine particles, thin films and exchange anisotropy," *Computer Engineering*, 33(13), 200-201.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srvinay, Manujakshi, B., Kabadi, M., & Naik, N. (2022). A hybrid stock price prediction model based on pre and Deep Neural Network. *Data*, 7(5), 51.
- Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2892-2900).
- Sureshkumar, K. K., & Elango, N. M. (2011). An efficient approach to forecast Indian stock market price and their performance analysis. *International Journal of Computer Applications*, 34(5), 44-49.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- Ta, V.-D., & Liu, C.-M. (2016). Stock market analysis using clustering techniques. *Proceedings of the Seventh Symposium on Information and Communication Technology - SoICT '16*.
- Tilakaratne, C. D., Mammadov, M., & Morris, S. A. (2009). Modified neural network algorithms for predicting trading signals of stock market indices.
- Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103).

Wang, Y. (2014). Stock price direction prediction by directly using prices data: an empirical study on the KOSPI and HSI. *International Journal of Business Intelligence and Data Mining*, 9(2), 145-160.

Wong, H. L., Tu, Y. H., & Wang, C. C. (2010). Application of fuzzy time series models for forecasting the amount of Taiwan export. *Expert Systems with Applications*, 37(2), 1465-1470.

Xiong, Y., & Zuo, R. (2016). Recognition of geochemical anomalies using a deep autoencoder network. *Computers & Geosciences*, 86, 75-82.

Yamashita, T., Hirasawa, K., & Hu, J. (2005). Multi-branch neural networks and its application to stock price prediction. In *Knowledge-Based Intelligent Information and Engineering Systems: 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part I 9* (pp. 1-7). Springer Berlin Heidelberg.

Yildiz, B., Yalama, A., & Coskun, M. (2008). Forecasting the Istanbul stock exchange national 100 index using an artificial neural network. *An International Journal of Science, Engineering and Technology*, 46, 36-39.

Yuan, F., Meng, Z. H., Zhang, H. X., & Dong, C. R. (2004, August). A new algorithm to get the initial centroids. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)* (Vol. 2, pp. 1191-1193). IEEE.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833). Springer International Publishing.