**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**BÙI NGÔ TÔN BÁCH**

**DỰ ĐOÁN GIÁ CỔ PHIẾU BẰNG HỌC MÁY**

**LUẬN VĂN THẠC SĨ**

*TP. Hồ Chí Minh, năm 2023*

**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**BÙI NGÔ TÔN BÁCH**

**DỰ ĐOÁN GIÁ CỔ PHIẾU BẰNG HỌC MÁY**

**Ngành: Vật lý Kỹ thuật**

**Mã số Ngành: 8520401**

**NGƯỜI HƯỚNG DẪN KHOA HỌC**

**TS. Nguyễn Anh Huy**

*TP. Hồ Chí Minh, năm 2023*

# LỜI CAM ĐOAN

Tôi cam đoan luận văn thạc sĩ ngành Vật lý Kỹ thuật với đề tài *“Dự đoán giá cổ phiếu bằng học máy”* là công trình khoa học do tôi thực hiện dưới sự hướng dẫn của TS. Nguyễn Anh Huy.

Những kết quả nghiên cứu của luận văn hoàn toàn trung thực và chính xác.

Học viên cao học

(Ký tên, ghi họ tên)

Bùi Ngô Tôn Bách

# LỜI CẢM ƠN

Để có thể hoàn thành được khóa luận, lời đầu tiên tôi xin gửi lời cảm ơn chân thành nhất tới TS. Nguyễn Anh Huy và anh Huỳnh Quốc Việt, người đã trực tiếp hướng dẫn tôi trong suốt quá trình hình thành ý tưởng và thực hiện đề tài luận văn này.

Xin chân thành cảm ơn quý thầy cô đã tận tình giảng dạy, truyền đạt kiến thức cho tôi trong gần hai năm qua. Tôi xin chân thành cảm ơn gia đình, bạn bè, đồng nghiệp đã có sự động viên, hỗ trợ và đóng góp ý kiến để tôi có thể hoàn thành công trình nghiên cứu.

Tuy nhiên, trong quá trình thực hiện luận văn không tránh khỏi những sai sót do kiến thức còn hạn chế, tôi rất mong nhận được sự đóng góp từ quý thầy cô và các bạn để tôi hoàn thành luận văn một cách hoàn chỉnh nhất và để trao dồi thêm kiến thức cho những nghiên cứu trong tương lai.

Tôi xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày tháng năm 2023

Bùi Ngô Tôn Bách

**MỤC LỤC**

[LỜI CAM ĐOAN i](#_heading=h.gjdgxs)

[LỜI CẢM ƠN ii](#_heading=h.30j0zll)

[DANH MỤC CÁC HÌNH VẼ VÀ ĐỒ THỊ v](#_heading=h.1fob9te)

[DANH MỤC CÁC BẢNG SỐ LIỆU vii](#_heading=h.3znysh7)

[DANH MỤC CÁC TỪ VIẾT TẮT viii](#_heading=h.2et92p0)

[TRANG THÔNG TIN LUẬN VĂN TIẾNG VIỆT ix](#_heading=h.tyjcwt)

[THESIS INFORMATION xi](#_heading=h.3dy6vkm)

[MỞ ĐẦU 1](#_heading=h.1t3h5sf)

[CHƯƠNG 1 TỔNG QUAN 2](#_heading=h.26in1rg)

[1.1](#_heading=h.4fsjm0b) KHÁI NIỆM CHỨNG KHOÁN 2

[1.2. PHÂN LOẠI CHỨNG KHOÁN 3](#_heading=h.lnxbz9)

[1.2.1 Cổ phiếu 3](#_heading=h.35nkun2)

[1.2.2 Trái phiếu 4](#_heading=h.1ksv4uv)

[1.2.3 Chứng chỉ quỹ 5](#_heading=h.1ksv4uv)

[1.2.4 Phái sinh 5](#_heading=h.1ksv4uv)

[1.3 MỤC TIÊU NGHIÊN CỨU 6](#_heading=h.44sinio)

[1.4 GIỚI HẠN CỦA LUẬN VĂN 6](#_heading=h.2jxsxqh)

[CHƯƠNG 2 CƠ SỞ LÝ THUYẾT 8](#_heading=h.3j2qqm3)

[2.1 TỔNG QUAN VỀ HỌC MÁY VÀ HỌC SÂU 8](#_heading=h.1y810tw)

[2.2 MẠNG NƠ-RON 9](#_heading=h.2xcytpi)

[2.2.1](#_heading=h.1ci93xb) Mạng nơ-ron truyền thẳng (ANN) 9

[2.2.2 Mạng hồi quy (RNN) 11](#_heading=h.qsh70q)

[2.2.3 Mạng bộ nhớ dài-ngắn hạn (LSTM) 13](#_heading=h.49x2ik5)

[2.2.4 Mạng nơ-ron tích chập (CNN) 16](#_heading=h.3o7alnk)

[2.3 CÁC PHƯƠNG PHÁP HUẤN LUYỆN MẠNG NƠ-RON 22](#_heading=h.3tbugp1)

[2.3.1 Phương pháp học có giám sát 22](#_heading=h.28h4qwu)

[2.3.2 Phương pháp học không giám sát 23](#_heading=h.1mrcu09)

[2.3.3 Phương pháp học tăng cường 24](#_heading=h.111kx3o)

[2.4](#_heading=h.206ipza) PHƯƠNG PHÁP ĐÁNH GIÁ 25

[CHƯƠNG 3 MÔ HÌNH ĐỀ XUẤT 27](#_heading=h.3q5sasy)

[3.1 TỔNG QUAN MÔ HÌNH 27](#_heading=h.25b2l0r)

[3.2 TRÍCH XUẤT TÍNH NĂNG 28](#_heading=h.25b2l0r)

[3.3 THIẾT KẾ MẠNG BỘ NHỚ DÀI-NGẮN (LSTM) 29](#_heading=h.25b2l0r)

[CHƯƠNG 4 THỰC NGHIỆM 31](#_heading=h.pkwqa1)

[4.1 DỮ LIỆU 31](#_heading=h.39kk8xu)

[4.2 XỬ LÝ DỮ LIỆU 34](#_heading=h.48pi1tg)

[4.2.1 Phân chia tập dữ liệu 34](#_heading=h.111kx3o)

[4.2.2 Tiền xử lý dữ liệu 34](#_heading=h.111kx3o)

[4.3 HUẤN LUYỆN 34](#_heading=h.1302m92)

[4.4 ĐÁNH GIÁ 38](#_heading=h.haapch)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 40](#_heading=h.1idq7dh)

[TÀI LIỆU THAM KHẢO 41](#_heading=h.1vsw3ci)

# DANH MỤC CÁC HÌNH VẼ VÀ ĐỒ THỊ

[**Hình 1.1:** Bảng điện chứng khoán 3](#_heading=h.2uxtw84)

[**Hình 2.1:** Cấu trúc của một nút 10](#_heading=h.3whwml4)

[**Hình 2.2:** Cấu trúc của một mạng ANN 11](#_heading=h.2bn6wsx)

[**Hình 2.3:** Cấu trúc của một mạng hồi quy 12](#_heading=h.3as4poj)

[**Hình 2.4:** Mô hình chi tiết của một lớp ẩn 13](#_heading=h.1pxezwc)

[**Hình 2.5:** Cấu trúc của một tế bào trong mạng LSTM 14](#_heading=h.2p2csry)

[**Hình 2.6:** Phương pháp nhận diện của máy tính với một hình ảnh 17](#_heading=h.23ckvvd)

[**Hình 2.7:** Các bước tính toán của phương pháp tích chập 18](#_heading=h.32hioqz)

[**Hình 2.8:** Ma trận chập với độ trượt bằng một 18](#_heading=h.1hmsyys)

[**Hình 2.9:** Ma trận chập với độ trượt bằng hai 19](#_heading=h.41mghml)

[**Hình 2.10:** Ma trận chập bị trượt ra ngoài 19](#_heading=h.2grqrue)

[**Hình 2.11:** Lớp padding phủ xung quanh hỗ trợ ma trận chập bị trượt ra ngoài có thể tính toán được các tính năng 20](#_heading=h.vx1227)

[**Hình 2.12:** Quá trình max pooling 20](#_heading=h.1v1yuxt)

[**Hình 2.13:** Quá trình global average polling 21](#_heading=h.4f1mdlm)

[**Hình 2.14:** Mô hình quyết định của lớp kết nối đầy đủ 22](#_heading=h.19c6y18)

[**Hình 2.15:** Mô hình huấn luyện học có giám sát 23](#_heading=h.nmf14n)

[**Hình** **2.16:** Mô hình kiểm tra phương pháp học có giám sát 23](#_heading=h.37m2jsg)

[**Hình 2.17:** Phương pháp học không giám sát 24](#_heading=h.2lwamvv)

[**Hình 2.18:** Mô tả phương pháp học tăng cường 24](#_heading=h.3l18frh)

[**Hình 3.1:** Mô hình tổng quan đề xuất 27](#_heading=h.2iq8gzs)

[**Hình 3.2:** Tính năng trước và sau trích xuất 29](#_heading=h.xvir7l)

[**Hình 3.3:** Mô hình tổng quan của mạng LSTM 30](#_heading=h.xvir7l)

[**Hình 4.1:** Biểu đồ giá cổ phiếu Vietcombank 35](#_heading=h.1opuj5n)

[**Hình 4.2:** Biểu đồ giá cổ phiếu Vingroup 36](#_heading=h.2nusc19)

[**Hình 4.3:** Biểu đồ giá cổ phiếu Hoà Phát 36](#_heading=h.3mzq4wv)

[**Hình 4.4:** Biểu đồ giá cổ phiếu Vinamilk 37](#_heading=h.2250f4o)

[**Hình 4.5:** Biểu đồ giá cổ phiếu Thế Giới Di Động 38](#_heading=h.319y80a)

# DANH MỤC CÁC BẢNG SỐ LIỆU

[**Bảng 4.1:** Dữ liệu cổ phiếu Vietcombank 31](#_heading=h.184mhaj)

[**Bảng 4.2:** Dữ liệu cổ phiếu Vingroup 32](#_heading=h.3s49zyc)

[**Bảng 4.3:**Dữ liệu cổ phiếu Hoà Phát 32](#_heading=h.meukdy)

[**Bảng 4.4:** Dữ liệu cổ phiếu Vinamilk 33](#_heading=h.36ei31r)

[**Bảng 4.5:** Dữ liệu cổ phiếu Thế Giới Di Động 33](#_heading=h.45jfvxd)

[**Bảng 4.6:** Kết quả đánh giá cổ phiếu Vietcombank 35](#_heading=h.2koq656)

[**Bảng 4.7:** Kết quảđánh giá cổ phiếu Vingroup 35](#_heading=h.zu0gcz)

[**Bảng 4.8:** Kết quả](#_heading=h.3jtnz0s) đánh giá cổ phiếu Hoà Phát [36](#_heading=h.3jtnz0s)

[**Bảng 4.9:** Kết quả đánh giá cổ phiếu Vinamilk 37](#_heading=h.1yyy98l)

[**Bảng 4.10:** Kết quả đánh giá cổ phiếu Thế Giới Di Động 37](#_heading=h.2ce457m)

[**Bảng 4.11:** So sánh kết quả đánh giá cổ phiếu Vietcombank giữa 2 đề tài 39](#_heading=h.184mhaj)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **VIẾT TẮT** | **TÊN TIẾNG ANH** | **TÊN TIẾNG VIỆT** |
| RMSE | Root Mean Square Error | Sai số trung bình bình phương gốc |
| MAE | Mean Absolute Error | Sai số trung bình tuyệt đối |
| MAPE | Mean Absolute Percentage Error | Phầm tram sai số trung bình tuyệt đối |
| AR | Average Return | Lợi nhuận trung bình |
| RAR | Real Average Return | Lợi nhuận trung bình thực tế |
| 1D | One-Dimensional | Một chiều |
| 2D | Two-Dimensional | Hai chiều |
| 3D | Three-Dimensional | Ba chiều |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| ML | Machine Learning | Học máy |
| DL | Deep Learning | Học sâu |
| CNN | Convolution Neural Network | Mạng nơ-ron tích chập |
| RNN | Recurrent neural network | Mạng nơ-ron hồi quy |
| ANN | Artificial | Mạng nơ-ron truyền thẳng |
| LSTM | Long Short Term Memory | Bộ nhớ dài ngắn hạn |

# TRANG THÔNG TIN LUẬN VĂN TIẾNG VIỆT

Tên đề tài luận văn: DỰ ĐOÁN GIÁ CỔ PHIẾU BẰNG HỌC MÁY.

Ngành: Vật lý Kỹ thuật

Mã số ngành: 8520401

Họ tên học viên cao học: Bùi Ngô Tôn Bách

Khóa đào tạo: 30

Người hướng dẫn khoa học: TS. Nguyễn Anh Huy

Cơ sở đào tạo: Trường Đại học Khoa học Tự nhiên, ĐHQG.HCM

1. **TÓM TẮT NỘI DUNG LUẬN VĂN**:

Nội dung của luận văn gồm 4 chương:

**Chương 1: Tổng quan**

Trình bày một số kiến thức cơ bản về chứng khoán và mục tiêu nghiên cứu của đề tài.

**Chương 2: Cơ sở lý thuyết**

Trình bày tóm lược về học máy, mạng nơ-ron nhân tạo, các loại cấu trúc thiết kế mạng nơ-ron, các phương pháp huấn luyện cũng như các phương pháp đánh giá.

**Chương 3: Mô hình đề xuất**

Xây dựng mô hình LSTM giải quyết bài toán dự đoán giá cổ phiếu.

**Chương 4: Thực nghiệm**

Huấn huyện mô hình LSTM trên tập dữ liệu có sẵn, đánh giá hiệu suất và kết luận.

**2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN VĂN**:

Luận văn đã đề xuất mô hình LSTM trong việc dự đoán giá cổ phiếu. Kết quả đánh giá cho thấy mô hình có hiệu suất tương đối tốt hơn mô hình trong một vài đề tài tương tự và có tiềm năng để phát triển tiếp.

**3.** **CÁC ỨNG DỤNG/ KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU**

* Thu thập bộ dữ liệu có kích thước lớn và đủ để phục vụ cho việc huấn luyện.
* Kết hợp nhiều thuật toán khác nhau để tăng hiệu quả của việc dự đoán.
* Ứng dụng của việc dự đoán trong giao dịch thực tế.

|  |  |
| --- | --- |
| **TẬP THỂ CÁN BỘ HƯỚNG DẪN**  (Ký tên, họ tên) | **HỌC VIÊN CAO HỌC**  (Ký tên, họ tên) |

**XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO**

**HIỆU TRƯỞNG**

# THESIS INFORMATION

Thesis title: PREDICTING STOCK PRICES USING MACHINE LEARNING.

Speciality: Engineering Physics

Code: 8520401

Name of Master Student: Bui Ngo Ton Bach

Academic year: 30

Supervisor: PhD. Nguyen Anh Huy

At: VNUHCM - University of Science

1. **SUMMARY:**

The content of the thesis consists of 4 chapters:

**Chapter 1: Overview**

This chapter presents some basic knowledge about securities and the research objectives of the topic.

**Chapter 2: Theoretical basis**

This chapter summarize about machine learning, artificial neural networks, types of neural network design structures, training methods and evaluation methods.

**Chapter 3: Proposed model**

Building the LSTM model to solve the problem of stock price prediction.

**Chapter 4: Designing an Artificial Neural Network Model**

Training the LSTM model on available data, evaluating performance and drawing conclusions.

**2. NOVELTY OF THESIS:**

This thesis proposed the LSTM model for predicting stock prices. The evaluation results showed that the model has relatively good performance compared to similar studies and has the potential for further development.

**3. APPLICATIONS/ APPLICABILITY/ PERSPECTIVE**

* Collecting a large dataset to serve for training.
* Combining multiple algorithms to increase the effectiveness of the prediction.
* Application of the prediction in real trading

|  |  |
| --- | --- |
| **SUPERVISOR** | **Master STUDENT** |

**CERTIFICATION**

**UNIVERSITY OF SCIENCE**

**PRESIDENT**

# MỞ ĐẦU

Thị trường chứng khoán là một trong những thị trường quan trọng nhất trên thế giới, cung cấp nền tảng để các công ty và cá nhân đầu tư vào cổ phiếu, giúp cho việc tăng trưởng và phát triển kinh doanh của công ty. Thị trường chứng khoán cổ phiếu là thị trường tiềm năng, với nhiều yếu tố bên trong và bên ngoài tác động đến giá cổ phiếu. Để thành công trong đầu tư chứng khoán, nhà đầu tư cần phải hiểu rõ các yếu tố này và cách chúng tác động đến giá cổ phiếu. Một trong những yếu tố quan trọng nhất là tình hình kinh tế, bao gồm tỷ lệ lãi suất, tỷ lệ thất nghiệp, và tỷ giá. Các sự kiện quốc tế cũng có thể tác động đến thị trường chứng khoán, như chiến tranh thương mại hoặc sự kiện chính trị.

Dự đoán giá cổ phiếu là một khía cạnh quan trọng trong quản lý tài chính và đầu tư. Việc dự đoán giá cổ phiếu có thể giúp nhà đầu tư đưa ra quyết định mua bán cổ phiếu một cách hợp lý và hiệu quả. Hiện nay, thông tin về giá cổ phiếu được cập nhật liên tục và nhanh chóng, điều này đòi hỏi các nhà đầu tư phải có khả năng dự đoán giá cổ phiếu một cách chính xác và đồng thời quản lý rủi ro một cách hiệu quả.

Trong luận văn này, tôi sẽ trình bày một phương pháp dự đoán giá cổ phiếu hiệu quả là sử dụng mô hình LSTM. LSTM là một loại mô hình trí tuệ nhân tạo được sử dụng rộng rãi trong các bài toán dự đoán thời gian liên tục, có khả năng học và ghi nhớ các quan hệ giữa các sự kiện trong quá khứ và hiện tại, giúp cho việc dự đoán trong tương lai trở nên chính xác hơn. Tôi sử dụng dữ liệu giá cổ phiếu thực tế để tiến hành thực nghiệm, đánh giá hiệu năng của thuật toán học máy, xem xét các yếu tố ảnh hưởng đến độ chính xác của dự đoán và cách cải thiện độ chính xác.

Nội dung của luận văn gồm 4 chương:

Chương 1: Tổng quan

Chương 2: Cơ sở lý thuyết

Chương 3: Mô hình đề xuất

Chương 4: Thực nghiệm

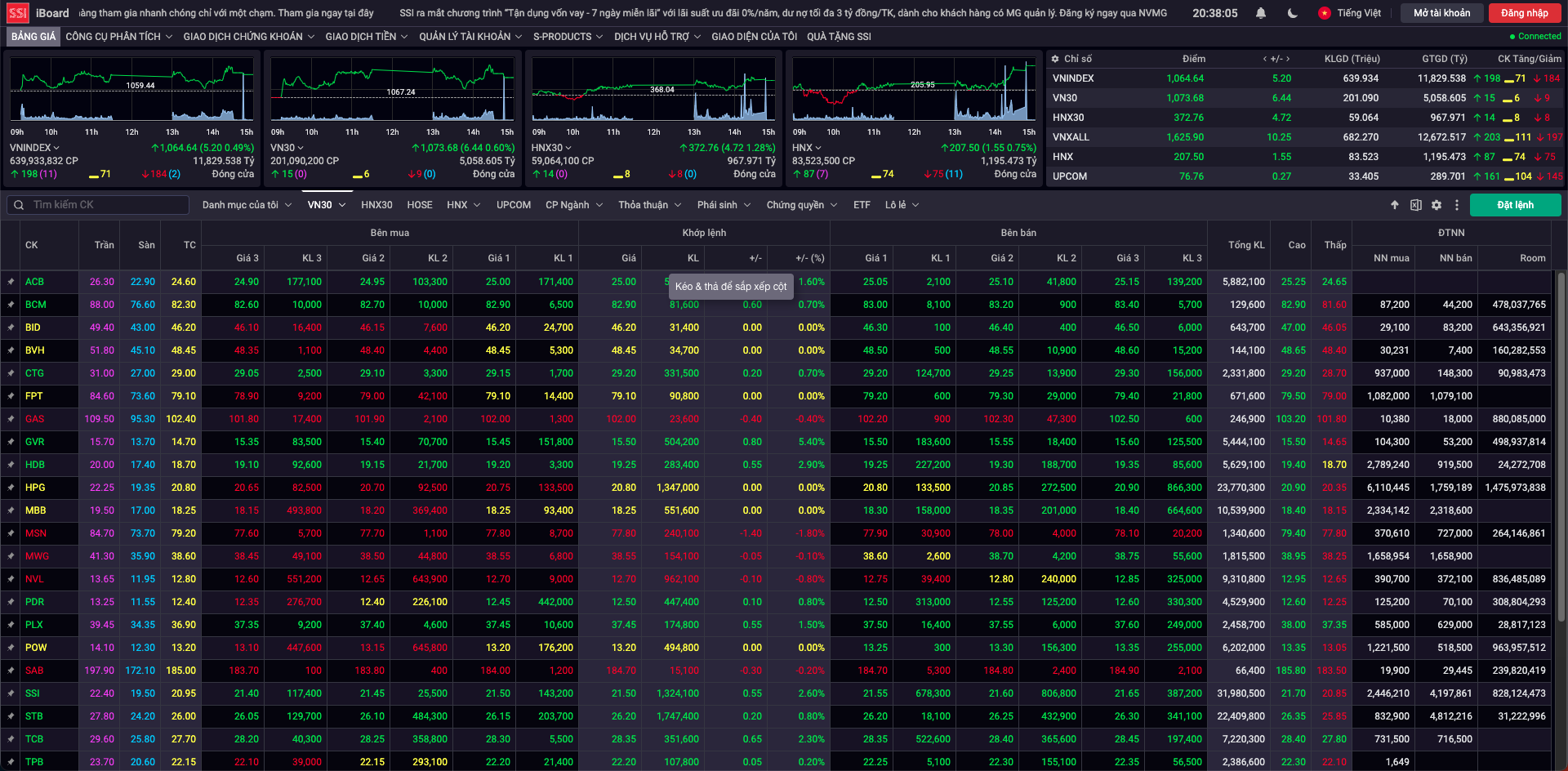
# CHƯƠNG 1 TỔNG QUAN

## KHÁI NIỆM CHỨNG KHOÁN

Chứng khoán là một loại tài sản quan trọng trong giao dịch tài chính, xác nhận việc góp vốn cho hoạt động kinh doanh của một công ty hoặc tổ chức. Khi một công ty hoặc tổ chức muốn tăng vốn, chứng khoán có thể được phát hành để bán cho người mua. Ngoài ra, chúng còn chứa thông tin về sở hữu vốn của công ty hoặc tổ chức và quyền lợi có liên quan đến.

Chứng khoán có thể được chuyển nhượng trên thị trường chứng khoán, cho phép các nhà đầu tư mua và bán chứng khoán của công ty hoặc tổ chức. Giá chứng khoán thường được xác định bởi sự cân bằng giữa cung và cầu trên thị trường chứng khoán. Những chứng khoán giá tăng thường được xem là có giá trị hơn và có thể tạo ra lợi nhuận cao hơn cho nhà đầu tư. Tuy nhiên, chứng khoán cũng có rủi ro liên quan đến sự biến động giá trên thị trường và hoạt động kinh doanh của công ty hoặc tổ chức.

Có nhiều yếu tố quan trọng ảnh hướng đến giá chứng khoán có thể được liệt kê như bên dưới:

* Tình hình kinh tế: sự thay đổi tình hình kinh tế như như tỉ lệ thất nghiệp, tăng trưởng GDP và lãi suất.
* Tình hình tài chính của công ty hoặc tổ chức: các thông số như lợi nhuận, dòng tiền và tình hình dư nợ của công ty hoặc tổ chức.
* Hoạt động kinh doanh: các hoạt động kinh doanh mới của công ty hoặc tổ chức như mở rộng vào các thị trường mới hoặc làm việc với các đối tác mới.
* Tình hình thị trường chứng khoán: xu hướng giảm hoặc tăng của thị trường chứng khoán (downtrend hoặc uptrend).
* Tin tức và sự kiện: sự kiện và tin tức liên quan đến công ty hoặc tổ chức như thông báo tài sản, quản lý thay đổi hoặc sự kiện chính trị.
* Đánh giá của chuyên gia: các đánh giá của chuyên gia về công ty hoặc tổ chức như xếp hạng của tổ chức tín dụng hoặc các chuyên gia tài chính.

##### Hình 1.1: Bảng điện chứng khoán

## 1.2. PHÂN LOẠI CHỨNG KHOÁN

### 1.2.1 Cổ phiếu

Cổ phiếu là loại chứng khoán xác nhận quyền sở hữu và lợi ích hợp pháp đối với thu nhập và tài sản của công ty cổ phần. Số vốn đóng góp để thành lập công ty được chia ra thành nhiều phần nhỏ bằng nhau gọi là cổ phần. Người mua cổ phần gọi là cổ đông. Cổ phiếu có thể được phát hành dưới dạng chứng chỉ vật chất hoặc bút toán ghi sổ, chỉ có công ty cổ phần mới có cổ phiếu. Giá trị ban đầu ghi trên cổ phiếu là mệnh giá của cổ phiếu. Mệnh giá là giá trị danh nghĩa. Số tiền nhận được từ khoản góp vốn gọi là cổ tức. Giá cổ phiếu dao động qua các phiên giao dịch trên thị trường chứng khoán và tách rời so với mệnh giá. Cổ phiếu được chia thành 2 loại:

* Cổ phiếu thường: là loại cổ phiếu không có kì hạn, tồn tại cùng với sự tồn tại của công ty phát hành ra nó, không có lãi suất cố định, số lãi được chia vào cuối mỗi kì tất toán. Cổ đông nắm giữ cổ phiếu này có quyền bỏ phiếu, mua thêm cổ phiếu mới, tham gia họp đại hội cổ đông.
* Cổ phiếu ưu đãi: gồm 2 loại là cổ phiếu ưu đãi biểu quyết và cổ phiếu ưu đãi tài chính. Cổ phiếu ưu đãi biểu quyết là cổ phiếu dành cho các cổ đông sáng lập. Cổ đông nắm giữ loại cổ phiếu này phải nắm giữ trong một khoảng thời gian nhất định, không được chuyển nhượng hay trao đổi. Cổ phiếu ưu đãi tài chính tương tự cổ phiếu thường nhưng có một số hạn chế như cổ đông nắm giữ cổ phiếu này không được tham gia bầu cử, ứng cử vào hội đồng quản trị, ban kiểm soát của công ty. Bù lại, họ được hưởng ưu đãi về tài chính theo một mức cổ tức riêng biệt, có tính cố định hằng năm, được ưu tiên chia cổ tức và phân chia tài sản còn lại của công ty sau khi thanh lý, giải thể trước cổ phiếu thường.

### 1.2.2 Trái phiếu

Trái phiếu là loại chứng khoán quy định nghĩa vụ của người phát hành phải trả cho người nắm giữ chứng khoán đó một khoản tiền xác định vào những thời hạn cụ thể và theo những điều kiện nhất định. Đây là những chứng khoán nợ, được phát hành dưới dạng chứng chỉ vật chất hoặc bút toán ghi sổ. Trái phiếu được chia thành các dạng sau:

* Trái phiếu vô danh: là loại trái phiếu không ghi tên trái chủ trên cả chứng chỉ và sổ sách của các tổ chức phát hành. Việc chuyển nhượng loại trái phiếu này rất dễ dàng nên thường được giao dịch trên thị trường chứng khoán.
* Trái phiếu ghi danh: là loại trái phiếu ghi tên, địa chỉ của trái chủ trên chứng chỉ và sổ sách của tổ chức phát hành. Loại trái phiếu này ít được giao dịch trên thị trường.
* Trái phiếu chính phủ: là loại trái phiếu do chính phủ phát hành nhằm bù đắp cho việc thâm hụt ngân sách. Đây là loại chứng khoán mà các nhà đầu tư an toàn rất thích vì nó hầu như không có rủi ro thanh toán.
* Trái phiếu công trình: là loại trái phiếu được phát hành để huy động vốn xây dựng các công trình cơ sở hạ tầng hay công trình phúc lợi công cộng.
* Trái phiếu công ty: là loại trái phiếu do công ty phát hành để vay vốn trung và dài hạn. Khi công ty bán trái phiếu thì công ty là người đi vay của người mua, người mua là chủ nợ của công ty. Công ty phải cam kết trả cả lãi và gốc cho trái chủ như đã nêu trong hợp đồng. Trái phiếu công ty bao gồm trái phiếu có đảm bảo, trái phiếu không có đảm bảo, trái phiếu có thể mua lại, …

### 1.2.3 Chứng chỉ quỹ

Chứng chỉ quỹ là chứng khoán được phát hành bởi công ty quản lý quỹ để huy động vốn từ các nhà đầu tư. Vốn được dung để mua bán, kinh doanh các loại chứng khoán khác để kiếm lời, sau đó chia tiền lời đó cho các nhà đầu tư. Có thể phân loại chứng chỉ quỹ đầu tư thành:

* Quỹ đầu tư chung: là loại quỹ mà tất cả mọi người đều có thể tham gia.
* Quỹ đầu tư riêng: là loại quỹ chỉ giới hạn ở một nhóm người nhất định.
* Quỹ đầu tư dạng mở: là loại quỹ mà nhà đầu tư có quyền bán lại chứng chỉ quỹ cho quỹ phát hành chứng khoán liên tục để huy động vốn và sẵn sàng mua lại số chứng khoán mà quỹ đã phát hành ra.
* Quỹ đầu tư cổ phiếu: là loại quỹ chuyên đầu tư vào cổ phiếu.
* Quỹ đầu tư trái phiếu: là loại quỹ chuyên đầu tư vào trái phiếu.
* Quỹ đầu tư hỗn hợp: là loại quỹ đầu tư vào bất kì loại chứng khoán nào nếu chúng có hiệu quả.

### 1.2.4 Phái sinh

Phái sinh là chứng khoán dưới dạng một hợp đồng mà trong đó giá trị phụ thuộc vào một hay nhiều loại tài sản cơ sở. Cụ thể, hợp đồng xác nhận quyền và nghĩa vụ của các bên đối với việc thanh toán tiền, chuyển giao số lượng tài sản cơ sở vào thời điểm xác định trong tương lai với mức giá nhất định được thoả thuận trước. Tài sản cơ sở của phái sinh cũng được quy định là chứng khoán, các chỉ số chứng khoán hoặc tài sản khác (thực phẩm, nông sản, kim loại, …). Các sản phẩm phái sinh cho phép nhà đầu tư *“đặt cược”* vào sự *“tăng”* hoặc *“giảm”* của tài sản cơ sở trong tương lai. Nếu sự thay đổi diễn ra đúng như dự đoán, nhà đầu tư sẽ có lời và ngược lại. Phái sinh gồm 4 loại chính:

* Hợp đồng kì hạn: là thoả thuận pháp lý giữa hai bên tham gia về việc mua và bán một loại tài sản tại một thời điểm nhất định trong tương lai với mức giá được xác định trước tại ngày thực hiện giao dịch.
* Hợp đồng tương lai: là một dạng hợp đồng kì hạn được chuẩn hoá, niêm yết và giao dịch tại thị trường tập trung.
* Hợp đồng quyền chọn: là thoả thuận pháp lý mà trong đó một bên có quyền yêu cầu thực hiện và bên kia có nghĩa vụ phải mua hoặc bán một lượng tài sản cơ sở theo mức giá đã được xác định trước tại hợp đồng trong một khoảng thời gian hoặc tại một thời điểm nhất định trong tương lai.
* Hợp đồng hoán đổi: là thoả thuận pháp lý mà trong đó có hai bên cam kết hoá đổi dòng tiền của công cụ tài chính của một bên với dòng tiền của công cụ tài chính của bên còn lại trong một khoảng thời gian nhất định.

## 1.3 MỤC TIÊU NGHIÊN CỨU

Trong luận văn này, tôi đã đề xuất và tập trung nghiên cứu những yếu tố sau:

* Nghiên cứu về giá cổ phiếu thuộc VN30.
* Nghiên cứu các phương pháp xử lý dữ liệu và các phương pháp đánh giá hiệu suất.
* Nghiên cứu các mô hình phân loại như LSTM, SVM, BAYES, ANN và GAN.
* Xử lý, tính toán, đánh giá các kết quả đạt được của mô hình LSTM với các bộ dữ liệu khác nhau.
* Đề xuất mô hình phân loại hiệu quả cho việc dự đoán giá cổ phiếu và hướng cải thiện hiệu suất mô hình LSTM.

## 1.4 GIỚI HẠN CỦA LUẬN VĂN

Trong luận văn, tôi chỉ sử dụng bộ dữ liệu của một vài cổ phiếu thuộc VN30, là 30 cổ phiếu được niêm yết tại sàn HoSE, có giá trị vốn hoá thị trường và tính thanh khoản cao nhất (tính đến 12/2022) (bộ dữ liệu đã được thu thập trên các trang về chứng khoán) nhằm phục vụ quá trình khảo sát. Những mô hình đã nghiên cứu hoặc đề xuất chỉ được thiết kế phù hợp với bộ dữ liệu này. Tuy nhiên, để có thể sử dụng bộ dữ liệu mới hoặc tự thu thập cho các mô hình phân loại này thì tôi cần phải tìm hiểu sâu về các thuật toán học máy cũng như các phương pháp xử lý dữ liệu khác chuyên biệt. Ngoài ra, luận văn chỉ tập trung xây dựng mô hình và phương pháp cho phù hợp với một tập dữ liệu nhằm tạo tiền đề và cơ sở cho các nghiên cứu chuyên sâu hơn sau này.

# CHƯƠNG 2 CƠ SỞ LÝ THUYẾT

## 2.1 TỔNG QUAN VỀ HỌC MÁY VÀ HỌC SÂU

Cách mạng công nghiệp 4.0 là một khái niệm để miêu tả sự ứng dụng cao của công nghệ thông tin vào các quá trình sản xuất và kinh doanh của doanh nghiệp. Điều này bao gồm việc sử dụng công nghệ như Internet vạn vật, điều khiển tự động, trí tuệ nhân tạo và tương tác trực tuyến để tăng năng suất và hiệu quả của quá trình sản xuất và kinh doanh. Trong đó, học máy đang trở thành một phần không thể thiếu trong lĩnh vực tài chính.

Học máy (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo (Artificial Intelligence) liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống *“học”* tự động từ dữ liệu để giải quyết những vấn đề cụ thể như các máy có thể *“học”* cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp thư vào thư mục tương ứng. Học máy hiện nay được áp dụng rộng rãi như máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động robot.

Học máy cũng đóng vai trò quan trọng trong lĩnh vực tài chính như:

* Giúp tự động hoá các quá trình tài chính như xử lý giao dịch, tạo hoá đơn, thanh toán và rút tiền. Điều này giúp giảm thiểu những lỗi do con người làm và tăng năng suất.
* Phân tích dữ liệu và dự đoán xu hướng giá của các mã chứng khoán hoặc tín dụng của khách hàng. Điều này giúp các nhà đầu tư và các công ty tài chính quản lý rủi ro và tìm kiếm cơ hội đầu tư tốt hơn.
* Xác thực và kiểm soát rủi ro cho các giao dịch tài chính. Nó có thể được sử dụng để phát hiện các giao dịch giả mạo hoặc không đúng, giúp ngăn chặn các rủi ro tội phạm.
* Tạo ra sản phẩm và dịch vụ tài chính mới và tiên tiến hơn như tài khoản ngân hàng tự động hoá, thẻ tín dụng điện tử hoặc dịch vụ chuyển tiền qua điện thoại.
* Chăm sóc khách hàng bằng các chatbot hoặc hệ thống tự động trả lời hỗ trợ khách hàng. Điều này giúp cho các công ty tài chính cung cấp dịch vụ khách hàng tốt hơn và giảm chi phí cho công ty.
* Đánh giá và quản lý rủi ro cho các khoản vay hoặc giao dịch tài chính khác và đưa ra quyết định hợp lý.

Học sâu (Deep Learning) là một phần của học máy và cũng được ứng dụng trong nhiều lĩnh vực khác nhau như nhận dạng ảnh, giọng nói, văn bản và các lĩnh vực như tài chính. Với sự phát triển của tốc độ máy tính, đặc biệt là khả năng tính toán trên GPU và sự tăng nhanh của dữ liệu cùng với các thư viện như TensorFlow hay Pytorch, việc *“huấn luyện”* các máy *“học”* trở nên dễ dàng hơn.

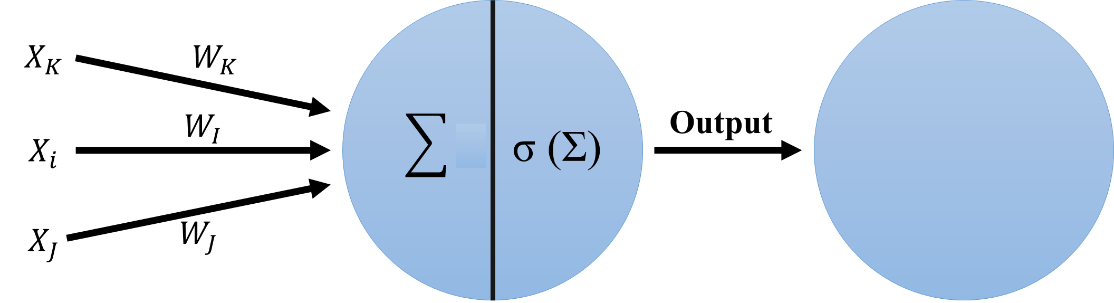
Học sâu sử dụng mạng nơ-ron nhân tạo để học và dự đoán. Mạng nơ-ron nhân tạo giống như một phiên bản mô hình hoá của các quá trình tình cảm và suy nghĩ của con người. Nó bao gồm các tầng được gọi là *“lớp”* và mỗi lớp là một tập hợp các *“nơ-ron”* được kết nối với nhau. Mỗi nơ-ron nhận vào dữ liệu từ các nơ-ron trong lớp trước và sau đó tính toán một giá trị đầu ra. Mạng nơ-ron được huấn luyện bằng cách sử dụng một tập dữ liệu huấn luyện và thay đổi các trọng số của các nơ-ron để tối ưu hoá đầu ra. Khi học sâu, mạng nơ-ron có thể học được các mẫu phức tạp và tìm ra các quan hệ giữa các đặc trưng và nhãn.

Học sâu nói riêng và học máy nói chung cung cấp nhiều loại mô hình và phương pháp học để giải quyết các vấn đề phức tạp khác nhau.

## MẠNG NƠ-RON

### 2.2.1 Mạng nơ-ron truyền thẳng (ANN)

Mạng nơ-ron nhân tạo được xây dựng dựa trên mô hình của mạng nơ-ron sinh học, trong đó các nút được kết nối với nhau qua các kết nối giống như các tế bào thần kinh. Mỗi kết nối có một giá trị trọng số đặc trưng, đại diện cho mức độ quan trọng của kết nối đó giữa các nút rơ-ron. Các trọng số này là thông số cơ bản cho bộ nhớ dài hạn của mô hình ANN. Quá trình học của mạng nơ-ron được thực hiện thông qua việc điều chỉnh lặp lại các trọng số sau mỗi chu kì huấn luyện để đạt được kết quả tối ưu nhất. Ba thành phần quan trọng của mạng nơ-ron bao gồm: thông số nút, cấu trúc của mạng và quy tắc huấn luyện [18].



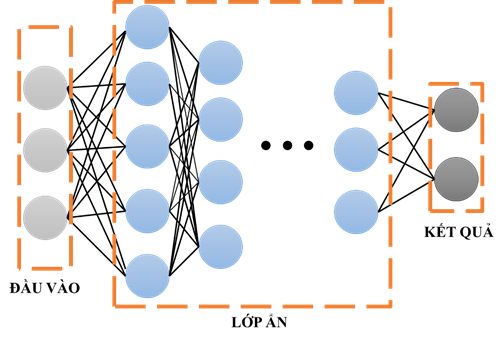
##### Hình 2.1: Cấu trúc của một nút

* Thông số nút: Đây là một thành phần quan trọng trong mạng nơ-ron, bao gồm việc xác định các xử lý tín hiệu trong các nút. Mỗi nút trong mạng nơ-ron có các thông số như số lượng đầu vào và đầu ra, sự liên kết với các nút khác, trọng số của từng mỗi nút và các hàm hoạt động. Các nút trong mạng nơ-ron nhận đầu vào từ những nút khác, mỗi đầu vào giữa các nút là các kết nối kèm theo trọng số liên quan. Khi tổng số đầu vào trọng số vượt quá giá trị ngưỡng của nút, các nơ-ron sẽ được kích hoạt và truyền tín hiệu thông qua các liên kết. Quá trình này có thể được biểu diễn bằng công thức toán học:

|  |  |
| --- | --- |
|  | (2.1) |

Trong đó, *y* là đầu ra của nút, là hàm hoạt động, *wi* là trọng số của đầu vào *xi* và *T* là giá trị ngưỡng. Hàm hoạt động có công thức khác nhau, tuy nhiên, hàm hoạt động phi tuyến sẽ hữu ích hơn so với những hàm hoạt động tuyến tính trong quá trình huấn luyện [26].

* Cấu trúc mạng: định hình cách tổ chức và kết nối giữa các nút trong mạng. Thông thường, các nút được tổ chức thành các lớp tuyến tính, gọi là lớp ẩn, tuy nhiên, có một số mạng nơ-ron không sử dụng lớp ẩn. Thiết kế cấu trúc mạng bao gồm việc quyết định số lượng nút trong mỗi lớp, số lượng lớp trong mạng và các liên kết giữa các nút, mặc định là các kiến trúc mạng truyền thẳng. Ban đầu, các thông số thường được thiết lập ngẫu nhiên và sau đó tối ưu hóa trong quá trình huấn luyện với các dữ liệu đầu vào khác nhau. Một số kiến trúc mạng truyền thẳng nổi tiếng như perceptron đã được sử dụng rộng rãi, cũng như một số mạng phản hồi nổi tiếng như mạng Hopfield và bản đồ tự tổ chức Kohonen [26].



##### Hình 2.2: Cấu trúc của một mạng ANN

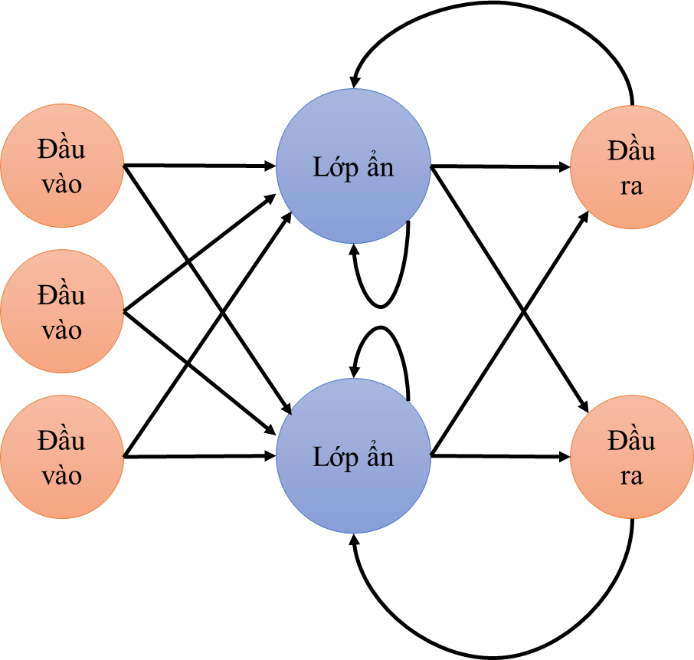
* Quy tắc huấn luyện được sử dụng để điều chỉnh các trọng số ban đầu của mạng nơ-ron thông qua các dữ liệu đầu vào trong quá trình huyến luyện.

Tuy nhiên, hầu hết các mạng ANN hiện nay không phù hợp cho việc xử lý dữ liệu dạng tuần tự (như ngôn ngữ, văn phạm, vv), loại dữ liệu này rất phổ biến trong đời sống con người. Điều này bởi vì các mạng ANN chỉ có khả năng truyền thẳng mà không có khả năng phản hồi trở lại của các nút, vì vậy các thông tin đã đi trước sẽ bị bỏ qua. Điều này dẫn đến mô hình ANN không thể nhận ra sự phụ thuộc vào thứ tự và mối liên hệ giữa các tín hiệu chuỗi khi được đưa vào mô hình

### 2.2.2 Mạng nơ-ron hồi quy (RNN)

Để giải quyết vấn đề của việc xử lý dữ liệu tuần tự trong mạng nơ-ron, các nhà nghiên cứu đã đưa ra phương pháp giải quyết bằng cách thêm các kết nối phản hồi giữa các lớp và nút của mạng. Phương pháp này cho phép mô hình học được mối tương quan giữa dữ liệu đầu vào trước đó và các dữ liệu hiện tại. Kết quả là một kiến trúc được gọi là mạng nơ-ron hồi quy (RNN). Mạng RNN có kiến trúc phản hồi, với một vài nút hoạt động giống như kiến trúc truyền thẳng, trong khi các nút khác có khả năng nhận giá trị từ các nút phía trước cũng như các nút lân cận hoặc phía sau. Kiến trúc của mạng RNN được mô tả chi tiết trong Hình 2.3

.



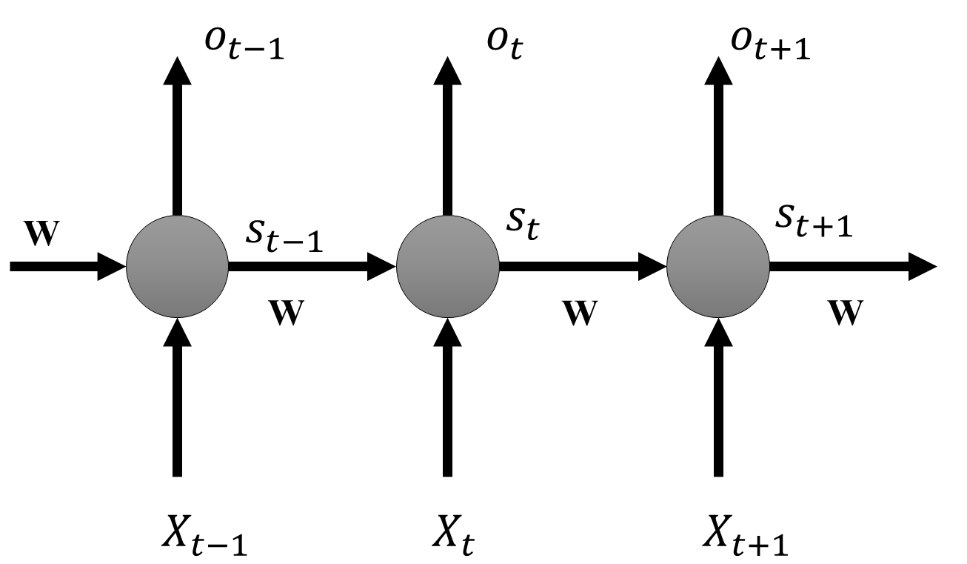
##### Hình 2.3: Cấu trúc của một mạng hồi quy

Hình 2.4 cho thấy cách triển khai một mạng nơ-ron hồi quy dưới dạng nhiều mạng nơ-ron kết nối với nhau để tạo thành một chuỗi tuần tự. Nguyên tắc hoạt động nơ-ron này có thể được mô tả như sau: nếu dữ liệu đầu vào gồm 5 thành phần, thì mạng nơ-ron sẽ được triển khai thành 5 tầng nơ-ron tương ứng với mỗi thành phần. Sau đó, các tính toán bên trong mạng nơ-ron hồi quy được thực hiện theo cách sau:

* là đầu vào tại thời điểm t hiện tại
* là bộ nhớ của mạng hồi quy tại thời điểm t. sẽ được tính toán thông qua các trạng thái ẩn trước đó và giá trị đầu vào hiện tại thông qua công thức:

|  |  |
| --- | --- |
|  | (2.2) |

* là giá trị đầu ra tại thời điểm t, một mô hình mạng hồi quy có thể có đưa ra một hoặc nhiều giá trị kết quả, với mỗi kết quả tương đương với từng thời điểm.



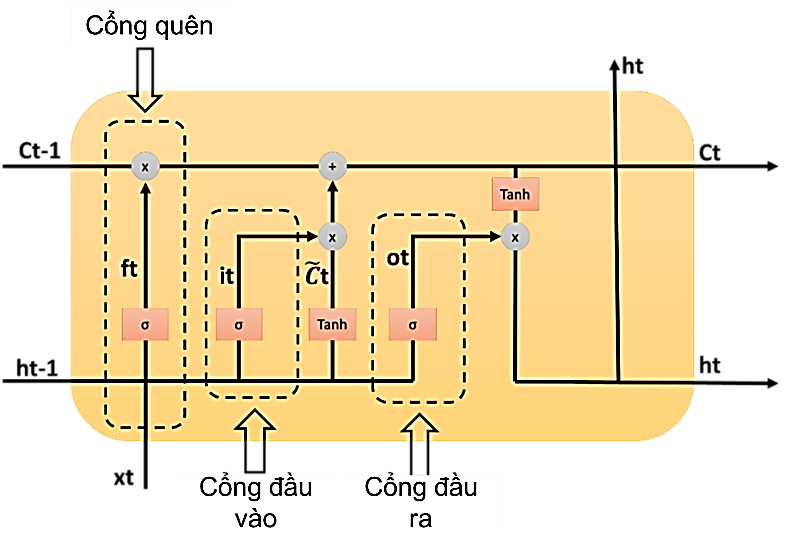
##### Hình 2.4: Mô hình chi tiết của một lớp ẩn

Mạng hồi quy có khả năng học các giá trị dữ liệu tuần tự, nhưng đối với các bộ dữ liệu có độ dài lớn, việc huấn luyện mô hình có thể gặp phải vấn đề mất mát dữ liệu giữa các lớp ẩn. Điều này có thể dẫn đến kết quả dự đoán không chính xác ở các nút phía sau, cho đến nút cuối cùng. Mặc dù lý thuyết cho thấy RNN có thể xử lý được các dữ liệu dài và lớn bằng cách tinh chỉnh các tham số phù hợp, trên thực tế, mô hình RNN thường không thể làm được điều này [27]. Tuy nhiên, vào năm 1997, mô hình LSTM được phát triển để khắc phục những vấn đề này mà mạng hồi quy không thể giải quyết được.

### 2.2.3 Mạng bộ nhớ dài-ngắn hạn (LSTM)

Mạng LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997) và đã được cải tiến và phổ biến bởi nhiều nhà nghiên cứu khác nhau, bao gồm Felix Gers, Fred Cummins, Santiago Fernandez, Justin Bayer, Daan Wierstra, Julian Togelius, Faustino Gomez, Matteo Gagliolo và Alex Graves [28]. Mạng này đã chứng tỏ sự hiệu quả của nó trên nhiều vấn đề và hiện đang được sử dụng rộng rãi.

Mạng LSTM được thiết kế để khắc phục vấn đề mất mát thông tin trong mạng hồi quy khi tín hiệu được truyền qua một lượng lớn các tế bào có trong mạng hồi quy. Điều này giúp mô hình LSTM trở thành một loại mạng hồi quy có khả năng xử lý các chuỗi có chiều dài tùy ý và khắc phục được nhược điểm của mô hình RNN. Bên cạnh đó, cấu trúc của một tế bào trong mạng như Hình 2.5 được sử dụng để giải thích cách hoạt động của mạng LSTM:



##### Hình 2.5: Cấu trúc của một tế bào trong mạng LSTM

Cấu trúc của một tế bào LSTM được chia làm ba cổng: cổng quên, cổng đầu vào và cổng đầu ra, trong đó các kí hiệu được mô tả như Bảng 2.2.

###### Bảng 2.2: Thông số các giá trị trong tế bào LSTM

|  |  |
| --- | --- |
| **Ký hiệu** | **Thông số** |
|  | Tín hiệu đầu vào |
|  | Tín hiệu đầu ra |
|  | Giá trị lưu trữ cho tế bào tiếp theo |
|  | Giá trị sẽ được lưu trữ |
|  | Cổng đầu vào |
|  | Cổng quên |
|  | Cổng đầu ra |

Các thông tin sẽ được đi qua cổng đầu tiên chính là “*cổng quên*”, chức năng của “*cổng quên”* sẽ xác định những thông tin nào từ tế bào trước nên bị loại bỏ và không cần thiết cho quá trình huấn luyện. Nếu giá trị kết quả là '0', thông tin của tế bào trước đó sẽ bị xóa. Ngược lại, nếu giá trị là '1' thì thông tin đó sẽ được cho phép đi qua cổng quên. Đầu ra của cổng được mô tả tổng quan thông qua phương trình sau:

|  |  |  |
| --- | --- | --- |
|  |  | (2.3) |

Tiếp đó sẽ là “*cổng đầu vào*”, đây là cổng hoạt động với sự hợp tác của hai hàm hoạt động song song với nhau. Hàm *Tanh* sẽ đưa ra các giá trị được chọn, trong khi hàm *Sigmoid* sẽ hoạt động tương tự như cổng quên đã đề cập bên trên. Do đó, trạng thái của giá trị tại vị trí “*cổng đầu vào*” là sự kết hợp các giá trị tính năng của các giá trị trong các tế bào trước và các giá trị được cập nhật thêm của tế bào hiện tại, công thức của “*cổng đầu vào*” có thể được mô tả theo phương trình:

|  |  |  |
| --- | --- | --- |
|  |  | (2.4) |

Cổng cuối cùng là “*cổng đầu ra*”. Tổng quát nhất, “*cổng đầu ra*” sẽ quyết định giá trị ở đầu ra của các tế bào hiện tại, các giá trị này sẽ được đưa đến các tế bào tiếp theo, đồng thời cũng sẽ được hiển thị ra ngoài như là một kết quả cuối cùng, công thức cho “*cổng đầu ra*” được hiển thị trong phương trình 2.5 và 2.6 tương ứng bên dưới:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  | | --- | |  | |  | | (2.5)  (2.6) |

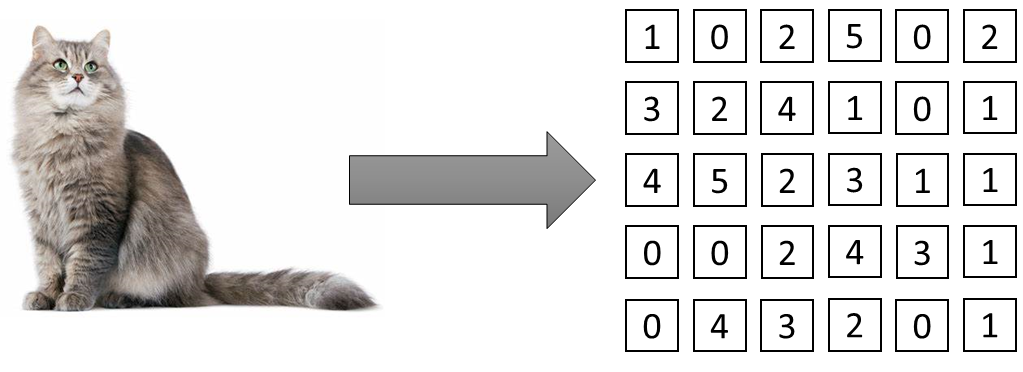
Các cổng trong mạng LSTM cho phép mạng này khắc phục vấn đề của mạng hồi quy và tránh xung đột giữa trọng số đầu vào và đầu ra. Mạng LSTM sử dụng thông tin của tín hiệu đầu vào trước đó để quyết định khi nào tín hiệu nên được lưu lại hoặc bị ghi đè trong bộ nhớ của tế bào, và khi nào các giá trị đã lưu trữ nên được truy cập. Các cổng này giúp mạng LSTM ngăn các đơn vị khác bị xáo trộn bởi các giá trị lưu trữ. Tuy nhiên, khi các tín hiệu lỗi đi vào các tế bào vào các thời điểm khác nhau, cổng đầu ra phải xác định lỗi nào cần được để lại trong cổng quên và cổng đầu vào phải loại bỏ các lỗi bằng cách thay đổi tỉ lệ độ lớn giá trị. Việc sử dụng các loại cổng này không phải lúc nào cũng cần thiết và đôi khi chỉ cần sử dụng cổng đầu vào là đủ cho quá trình huấn luyện. Tuy nhiên, việc sử dụng các cổng đầu ra cũng có nhiều lợi ích, ví dụ như ngăn chặn mạng lưu trữ các thông tin có độ trễ thời gian dài, góp phần tăng hiệu suất của mạng.

### 2.2.4 Mạng nơ-ron tích chập (CNN)

Mạng nơ-ron tích chập (CNN) là một mô hình mạng nơ-ron truyền thẳng được giới thiệu vào năm 1990 và được tạo ra để giải quyết các vấn đề liên quan đến hình ảnh. Không chỉ có thể giải quyết các vấn đề về hình ảnh, mô hình CNN còn có thể được sử dụng để giải quyết các vấn đề trong lĩnh vực máy học và khai thác dữ liệu, trong đó các tín hiệu đầu vào có thể được biểu thị bằng một hình ảnh hoặc một tập hợp các hình ảnh. Mô hình này có thể được xem như là một phiên bản cải tiến của mô hình mạng nơ-ron đa perceptron [30].

Mạng nơ-ron tích chập có thể huấn luyện trực tiếp với các tín hiệu thô mà không cần trích xuất tính năng của từng tín hiệu. Tuy nhiên, để huấn luyện mạng CNN, cần có bộ dữ liệu huấn luyện lớn để có thể học được hàng triệu tính năng từ một dữ liệu cụ thể. Tính toán trong quá trình huấn luyện sẽ rất phức tạp và thời gian hoàn thành huấn luyện sẽ rất dài. Để giải quyết vấn đề này, cần sử dụng các bộ xử lý đồ họa đặc thù (GPU) để huấn luyện mạng nhanh hơn và trong thời gian chấp nhận được.

Một mạng nơ-ron tích chập bao gồm các thành phần chính: lớp chập, lớp pooling và lớp fully-connect.



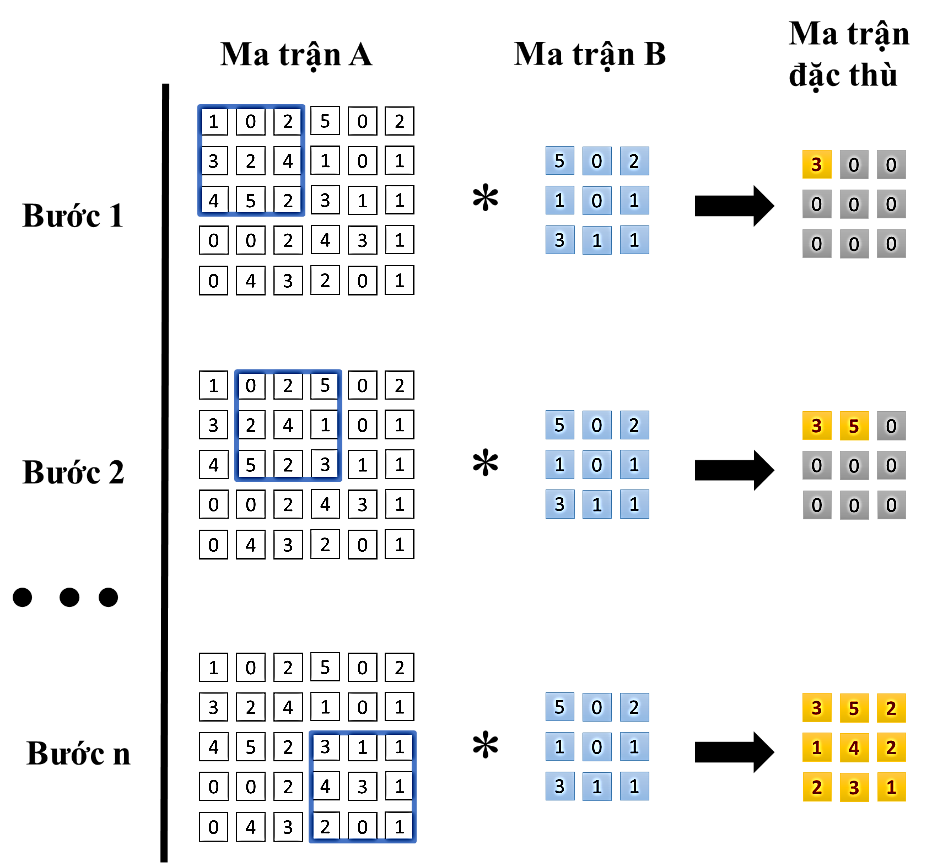
##### Hình 2.6: Phương pháp nhận diện của máy tính với một hình ảnh

#### 2.2.4.1 Lớp chập

Trong lớp chập của mạng CNN, các thông số như ma trận chập, độ trượt (stride) và lớp phủ (padding) sẽ cần phải được chú ý đến trong quá trình thiết kế cho một mạng nơ-ron tích chập.

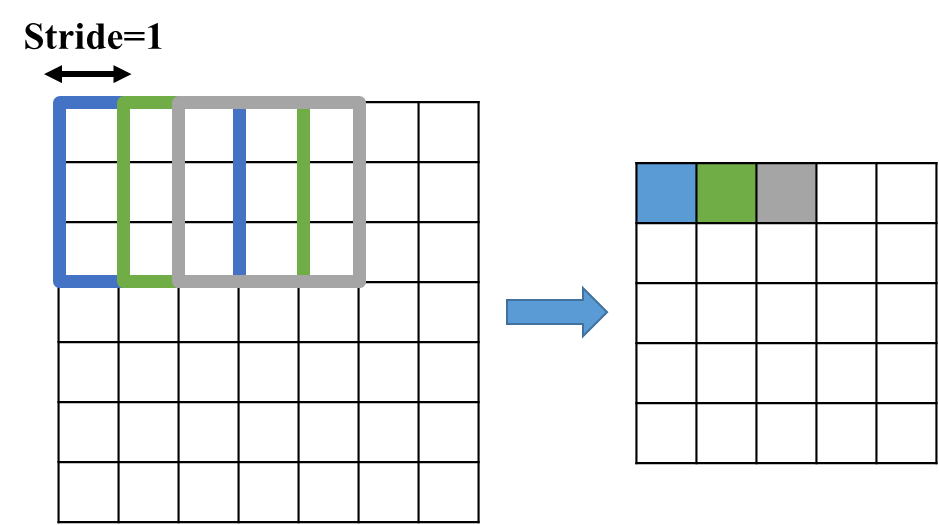
Đầu tiên là ma trận chập, khác với các mạng đã đề cập ở các mô hình mạng truyền thẳng và mạng hồi quy, các trọng số của mạng sẽ tương đương các giá trị bên trong mỗi ma trận chập (Kernel). Nguyên lý hoạt động chủ yếu của các lớp chập sẽ dựa trên các tính toán của tích chập.

Tích chập là một loại phép toán tuyến tính chuyên biệt được sử dụng để trích xuất các tính năng đặc trưng. Để thực hiện tích chập, một ma trận A sẽ được chập bởi một ma trận B để ra một ma trận mới chứa các đặc thù của ma trận A, phương pháp thực hiện được mô tả như Hình 2.7.

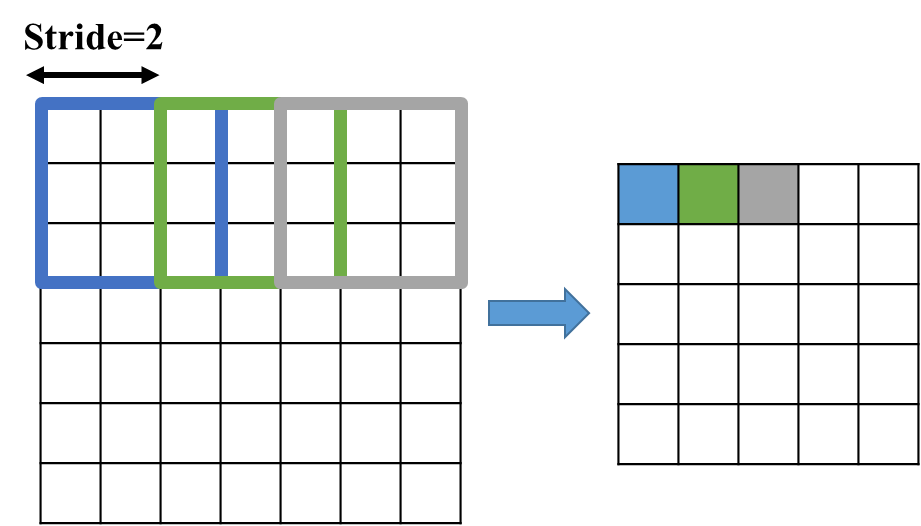


##### Hình 2.7: Các bước tính toán của phương pháp tích chập

Vùng mà ma trận B quét tới được gọi là vùng tiếp nhận. Ma trận đó sẽ trượt qua các ma trận A (các nút nơ-ron của mạng), trong quá trình trượt, hai ma trận sẽ nhân các phần tử của ma trận B với các phần tử trong khu vực tiếp nhận của ma trận A. Quá trình trượt bộ lọc được gọi là “*stride*”. Những ma trận kết quả được tạo thành sẽ được gọi là bản đồ tính năng. Trong quá trình chập sẽ có một số lượng lớn bản đồ tính năng được tạo thành bằng phương pháp chập với các ma trận chập khác nhau. Các bản đồ tính năng đấy sẽ đại diện cho một đặc thù như cạnh, đường cong [31].

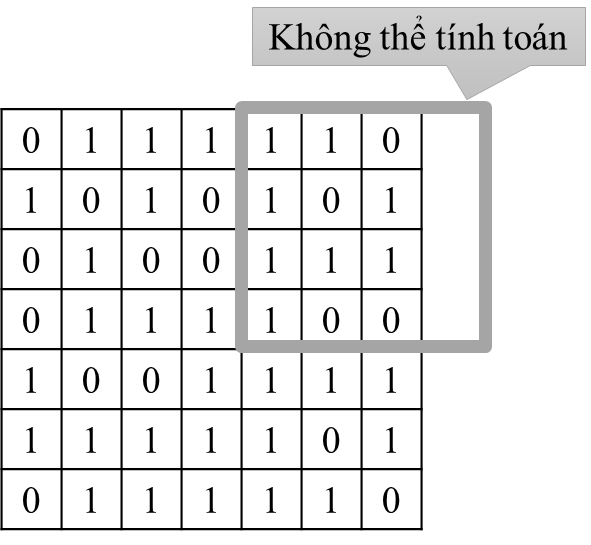


##### Hình 2.8: Ma trận chập với độ trượt bằng một



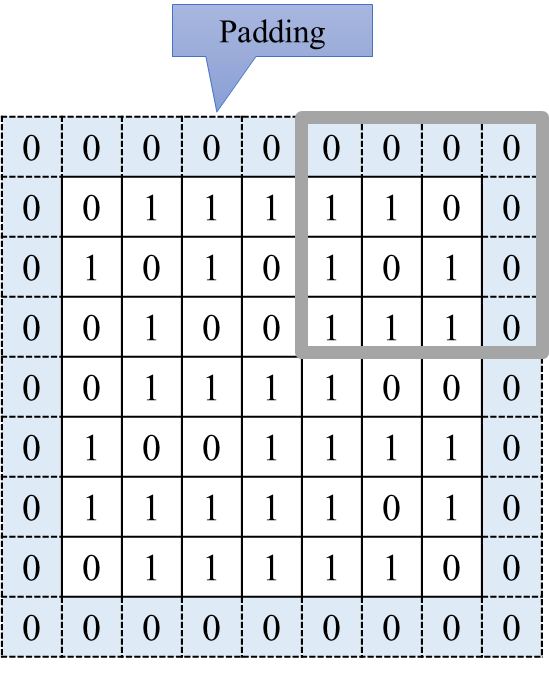
##### Hình 2.9: Ma trận chập với độ trượt bằng hai

Thông số tiếp theo là độ trượt (Stride). Stride là khoảng cách giữa các vị trí mà bộ lọc trượt trên ma trận đầu vào. Thông số này sẽ điều khiển việc trượt của ma trận chập trên toàn bộ ma trận và kết quả của bản đồ tính năng cũng sẽ thay đổi tùy thuộc vào giá trị stride. Hình 2.8 và 2.9 minh họa quá trình trượt của ma trận chập với stride bằng một và hai. Trong quá trình trượt với các giá trị stride lớn hoặc khi các bản đồ tính năng đã được tính toán qua nhiều lớp chập, kích thước của các bản đồ tính năng sẽ rất nhỏ. Khi tính toán tiếp trong lớp chập, các ma trận chập có thể bị trượt ra ngoài (Hình 2.10) và rìa của ma trận chập sẽ không có giá trị để có thể tính toán được.



##### Hình 2.10: Ma trận chập bị trượt ra ngoài

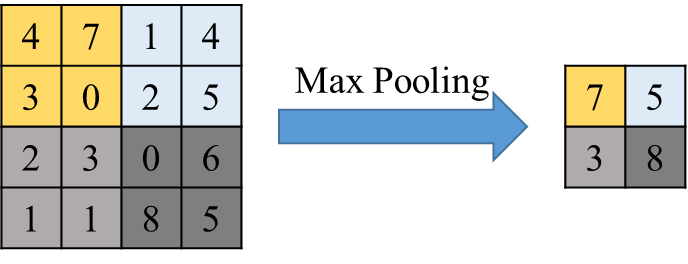
Để giải quyết vấn đề kích thước ma trận bị giảm trong quá trình chập, ta có thể áp dụng phương pháp "Zero padding" bằng cách thêm các giá trị '0' xung quanh ma trận ban đầu. Phương pháp này giúp duy trì kích thước đầu vào và loại bỏ việc giảm kích thước đầu ra. Đặc biệt, khi sử dụng trong các mạng nơ-ron với nhiều lớp, phương pháp này giúp tránh tình trạng kích thước đầu ra bị thu nhỏ quá nhiều [31].



##### Hình 2.11: Lớp padding phủ xung quanh hỗ trợ ma trận chập bị trượt ra ngoài có thể tính toán được các tính năng

#### 2.2.4.2 Lớp Pooling

Sau khi tín hiệu được xử lý qua lớp tích chập, thường sẽ sử dụng các lớp pooling để đơn giản hóa thông tin được trích xuất. Các lớp này giúp tạo ra một bản đồ tính năng rút gọn từ bản đồ tính năng ban đầu. Các phương pháp pooling phổ biến bao gồm max pooling và global average pooling. Max pooling chọn giá trị lớn nhất trong mỗi vùng của bản đồ tính năng, trong khi global average pooling tính giá trị trung bình của toàn bộ bản đồ tính năng.

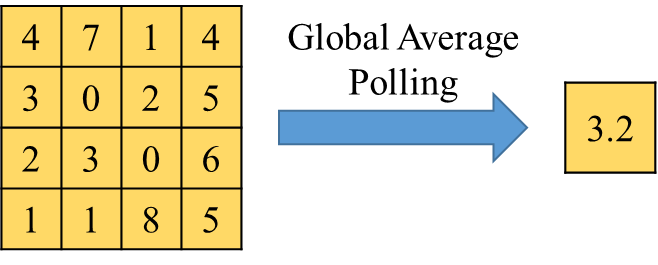


##### Hình 2.12: Quá trình max pooling

Phương pháp gộp thông tin phổ biến nhất là max pooling, nó tách các vùng của bản đồ tính năng đầu vào và trích xuất giá trị lớn nhất trong mỗi vùng, loại bỏ các giá trị còn lại (Hình 2.12). Max pooling với ma trận kích thước 2 × 2 và độ trượt bằng hai thường được sử dụng trong thực tế để giảm kích thước của bản đồ tính năng theo hệ số 2 trong mặt phẳng. Khác với chiều cao và chiều rộng, kích thước chiều sâu của bản đồ tính năng không thay đổi [32].

Một phương pháp gộp khác đáng chú ý là phép gộp trung bình toàn cục. Phương pháp này sẽ chuyển đổi bản đồ tính năng có kích thước chiều cao x chiều rộng được lấy mẫu thành một mảng có kích thước 1 x 1 bằng cách lấy giá trị trung bình của tất cả các phần tử trong mỗi bản đồ tính năng, tuy nhiên độ sâu của bản đồ đối tượng không thay đổi. Phương pháp này thường chỉ được áp dụng một lần trước khi các bản đồ tính năng được truyền vào lớp kết nối đầy đủ. Việc thực hiện phép gộp trung bình toàn cục mang lại nhiều ưu điểm, ví dụ:

* Giảm số lượng các tham số có thể được huấn luyện
* Cho phép CNN chấp nhận các tín hiệu đầu vào có kích thước khác nhau.

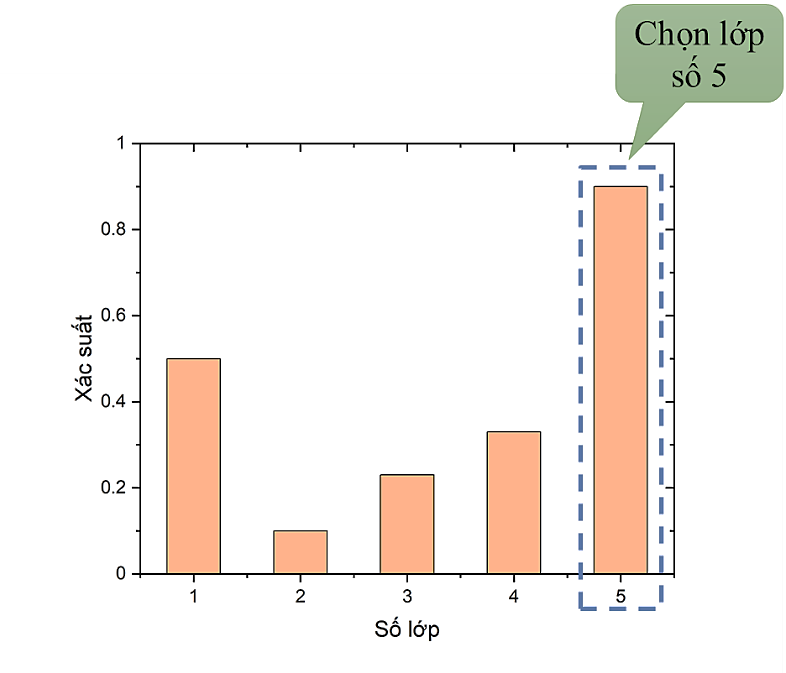


##### Hình 2.13: Quá trình global average polling

#### 2.2.4.3 Lớp kết nối đầy đủ (Fully-Connect)

Ở lớp kết nối đầy đủ, bản đồ tính năng từ lớp tích chập hoặc lớp gộp sẽ được phẳng thành một mảng một chiều (1D), sau đó kết nối với một hoặc nhiều lớp kết nối đầy đủ. Các tính năng đầu vào được liên kết với đầu ra thông qua các trọng số.

Ví dụ, giả sử ta có một vectơ 1 chiều có 5 giá trị là: 0.5, 0.1, 0.23, 0.33 và 0.9. Sau khi truyền qua lớp kết nối đầy đủ, ta sẽ có 5 lớp được trích xuất ra. Mỗi giá trị trong vectơ tương ứng với xác suất của đầu vào được phân loại vào từng lớp. Ví dụ, giá trị 0.5 là xác suất của đầu vào được phân loại vào lớp 1, giá trị 0.1 là xác suất của nó thuộc lớp 2 và tương tự cho các giá trị còn lại. Vì giá trị xác suất tương ứng với lớp 5 là cao nhất (0.9), mô hình CNN sẽ quyết định tín hiệu đầu vào được phân loại vào lớp thứ 5 [31]. Để thực hiện việc phân loại đầu vào, ta cần sử dụng hàm softmax để chuẩn hóa các giá trị thực đầu ra từ lớp kết nối đầy đủ thành xác suất của từng lớp mục tiêu. Kết quả trả về sẽ là một vectơ 1 chiều chứa các giá trị xác suất trong khoảng từ 0 đến 1, tổng các giá trị bằng 1. Cuối cùng, ta sẽ quyết định lớp phù hợp nhất cho đầu vào dựa trên giá trị xác suất cao nhất trong vectơ xác suất này.



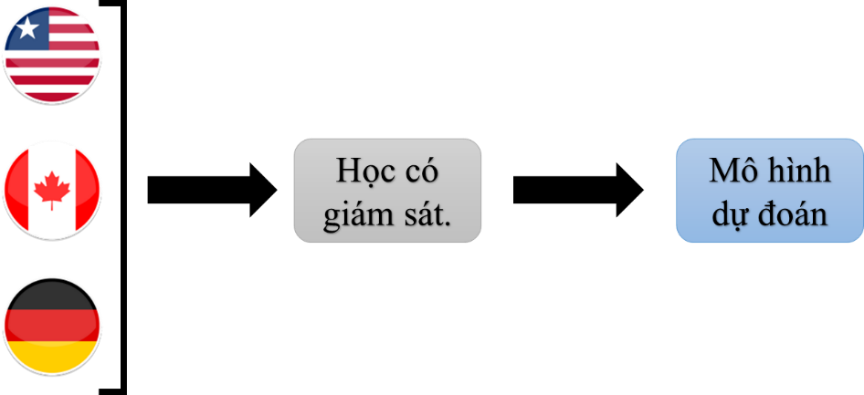
##### Hình 2.14: Mô hình quyết định của lớp kết nối đầy đủ

## 2.3 CÁC PHƯƠNG PHÁP HUẤN LUYỆN MẠNG NƠ-RON

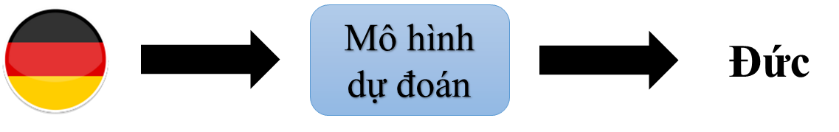
Trong quá trình huấn luyện, các trọng số được điều chỉnh để đạt được các giá trị mong muốn, để có thể thực hiện quá trình đó các mạng nơ-ron cần phải có phải được huấn luyện. Quá trình huấn luyện có thể được phân thành hai loại chính: học giám sát và học không giám sát, học tăng cường.

### 2.3.1 Phương pháp học có giám sát

Trong phương pháp học có giám sát, tập dữ liệu huấn luyện bao gồm các cặp đầu vào và đầu ra tương ứng được cung cấp trước, vì vậy việc sở hữu một tập dữ liệu chính xác là vô cùng quan trọng để đảm bảo tính chính xác của quá trình huấn luyện. Trọng số sẽ được điều chỉnh để giảm thiểu sai số giữa đầu ra của mạng và đầu ra mong muốn. Một mô hình mạng nơ-ron được huấn luyện một cách lý tưởng phải có sai số giữa kết quả đầu ra và các kết quả mong muốn ở mức thấp nhất. Đối với các mô hình sử dụng học có giám sát, trước khi sử dụng, mô hình phải được huấn luyện và tạo ra các kết quả cho một loạt các đầu vào. Sau đó, các trọng số được cố định khi sai số là nhỏ nhất hoặc khi quá trình huấn luyện kết thúc. Khi đó, mô hình mạng có thể được sử dụng để dự đoán kết quả cho các đầu vào mới [26].



##### Hình 2.15: Mô hình huấn luyện học có giám sát

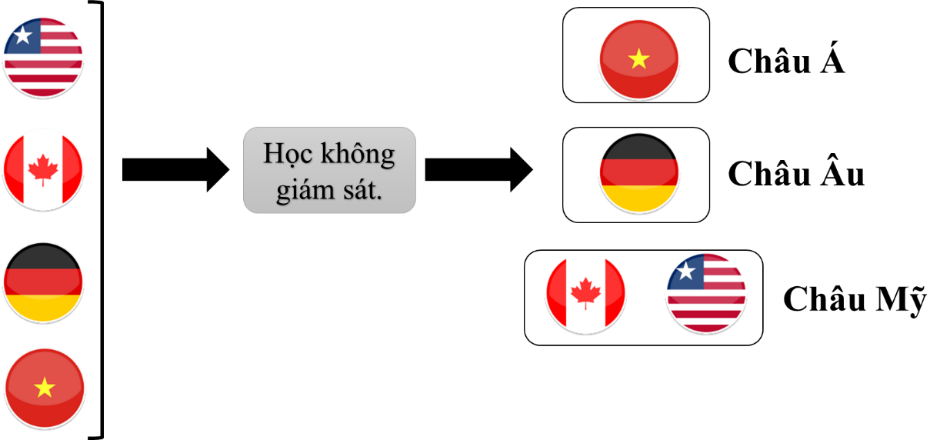


##### Hình 2.16: Mô hình kiểm tra phương pháp học có giám sát

### 2.3.2 Phương pháp học không giám sát

Trong học không giám sát, không sử dụng các giá trị đầu ra mục tiêu từ bộ dữ liệu. Thay vào đó, mạng nơ-ron sẽ tự động học và xây dựng các đại diện của đầu vào, có thể được sử dụng để dự đoán đầu ra trong tương lai hoặc phát hiện các cấu trúc hoặc mẫu tiềm ẩn trong dữ liệu đầu vào. Tóm lại, mạng nơ-ron sẽ cố gắng tạo ra các mô hình phù hợp cho bộ dữ liệu đó hoặc tìm ra các xu hướng, cấu trúc ẩn trong dữ liệu đầu vào [33].

Trong quá trình học này, các dữ liệu đầu vào thường được giả định là các mẫu độc lập. Ví dụ, trong trường hợp hình ảnh, các đối tượng có thể được chiếu sáng từ các góc độ khác nhau, dẫn đến tạo ra các hình ảnh khác nhau. Trong số đó, có một số hình ảnh tốt hơn cho quá trình muốn thực hiện. Mô hình sẽ cố gắng phân cụm các hình ảnh này lại với nhau để tạo ra những mô hình đại diện cho các đối tượng trong hình ảnh [33].



##### Hình 2.17: Phương pháp học không giám sát

### 2.3.3 Phương pháp học tăng cường



##### Hình 2.18: Mô tả phương pháp học tăng cường

Trong quá trình học tăng cường, các mô hình mạng tương tác với dữ liệu huấn luyện bằng cách áp dụng các phương pháp giải quyết khác nhau: A1, A2, ... Các hành động này sẽ có ảnh hưởng tuần tự đến bộ dữ liệu, và với mỗi hành động, mô hình sẽ nhận được một phần thưởng (hoặc một khoản phạt) tương ứng với A1, A2 và kết quả được gọi là R1, R2, ... (Hình 2.18). Mục tiêu của phương pháp học tăng cường là học cách thực hiện các hành động để tối đa hóa phần thưởng (hoặc giảm thiểu khoản phạt) trong suốt quá trình huấn luyện. Học tăng cường thường được sử dụng trong các lĩnh vực liên quan đến lý thuyết quyết định (trong thống kê và khoa học quản lý) và lý thuyết điều khiển (trong kỹ thuật) [34].

## 2.4 PHƯƠNG PHÁP ĐÁNH GIÁ

Với các bài toán dự đoán giá tiền ảo, chứng khoán nói chung, để đánh giá hiệu quả của mô hình và phương pháp thường dựa trên các tiêu chí thông kê là các độ đo:

* Sai số trung bình bình phương gốc RMSE (Root Mean Square Error):

|  |  |
| --- | --- |
|  | (2.7) |

* Sai số trung bình tuyệt đối MAE (Mean Absolute Error):

|  |  |  |
| --- | --- | --- |
|  |  | (2.8) |

* Phần trăm sai số trung bình tuyệt đối MAPE (Mean Absolute Percentage Error):

|  |  |
| --- | --- |
|  | (2.9) |

* Lợi nhuận trung bình AR (Average Return):

|  |  |  |
| --- | --- | --- |
|  |  | (2.10) |

Trong đó:

là dữ liệu cổ phiếu thực tế

là dữ liệu cổ phiếu dự đoán

là tổng số ngày giao dịch cổ phiếu (số mẫu)

Các giá trị RMSE, MAE, MAPE càng thấp có nghĩa là dự đoán càng chính xác, giá dự đoán càng sát với giá thật.

Tuy nhiên, do đặc điểm giao dịch của thị trường chứng khoán Việt Nam, công thức tính lợi nhuận trung bình không phù hợp với thực tế vì nhà đầu tư còn phải chịu thuế TNCN và phí giao dịch cổ phiếu. Giả sử thị trường Việt Nam cho phép mua bán cổ phiếu ngay trong ngày (T0), mức phí giao dịch mua bán thường được các công ty chứng khoán áp dụng là 0.015% giá trị mua bán và thuế TNCN là 0.01% giá trị bán. Khi đó công thức tính là:

* Lợi nhuận trung bình thực tế RAR (Real Average Return):

|  |  |
| --- | --- |
|  | (2.11) |

**CHƯƠNG 3  
MÔ HÌNH ĐỀ XUẤT**

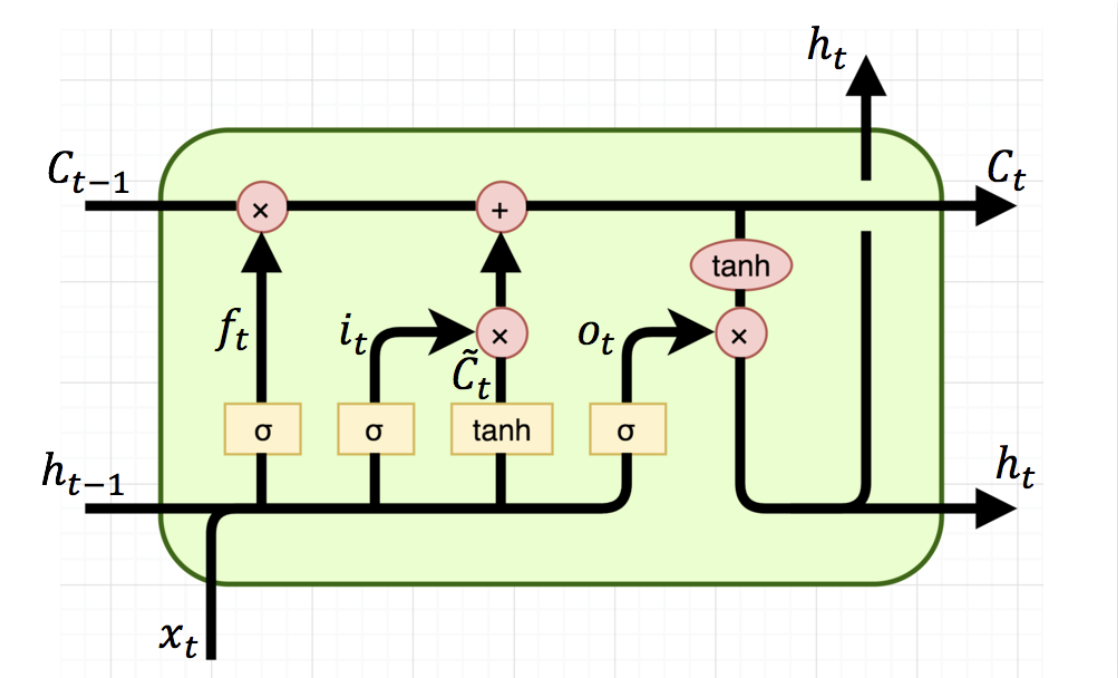
## 3.1 TỔNG QUAN MÔ HÌNH

## Mô hình tổng quan đề xuất được trình bày ở hình 3.1 bên dưới.

Data *(1)*

Preprocessing *(2)*

Features *(3)*

Training Model *(4)*

Predicting Price *(5)*

##### Hình 3.1: Mô hình tổng quan đề xuất

* *(1)*: Tôi thu thập dữ liệu thô bằng API của Investpy và lưu dữ liệu dưới dạng tệp đuôi .csv. Sau khi đã có đủ dữ liệu cổ phiếu từ năm 2015 đến năm 2022 (khoảng 1800 mẫu), tôi bắt đầu tiến hành tách tập dữ liệu thành tập huấn luyện và tập kiểm thử với tỉ lệ 80:20 (khoảng 1500 mẫu và 300 mẫu).
* *(2)* và *(3)*: Tôi tiền xử lý bằng cách chuẩn hoá dữ liệu và trích xuất tính năng nhằm lấy các đặc trưng của dữ liệu. Tôi loại bỏ những dữ liệu thừa, chỉ giữ lại những tính năng cần thiết cho việc huấn luyện bao gồm Open, High, Low, Close và Volume.
* *(4)*: Tôi đưa tập dữ liệu sau khi được tiền xử lý vào huấn luyện mô hình. Bằng phương pháp vét cạn, tôi chọn được epochs (số lần lặp lại quá trình huấn luyện) là 100, kích thước batch (số lượng mẫu được sử dụng trong mỗi lần cập nhật tham số) là 32, bộ tối ưu hoá (optimizer) Adam, hàm mất mát (loss) Mean Squared Error.
* *(5)*: Cuối cùng, tôi sử dụng mô hình vừa huấn luyện để dự đoán giá bằng việc chạy trên tập dữ liệu kiểm thử. Tôi tính các độ đo để đánh giá hiệu suất mô hình và vẽ đồ thị giá dự đoán, giá thực tế của cổ phiếu để biểu diễn trực quan.

## 3.2 TRÍCH XUẤT TÍNH NĂNG

Việc trích xuất các tính năng của giá cổ phiếu là quá trình cực kì quan trọng và cần thiết. Thông qua các tính năng được trích xuất, giá cổ phiếu sẽ có cơ sở để dự đoán trong tương lai, từ đó giúp giảm thiểu chi phí tính toán cũng như tăng cường độ chính xác của kết quả cuối cùng. Nói cách khác, quá trình trích xuất tính năng sẽ làm tăng tốc độ huấn luyện cho các dữ liệu khi sử dụng các thuật toán học máy có giám sát nhằm phục vụ việc dự đoán, ở đây là dự đoán giá cổ phiếu. Tuy nhiên để có thể thực hiện quá trình này mà hiệu quả đạt được là cao nhất, thì cần phải có một bộ dữ liệu mẫu chuẩn xác và uy tín. Chính vì thế, sau tôi sẽ lấy dữ liệu từ API Investpy, từ dữ liệu lấy được dưới dạng .csv gồm các cột tính năng, tôi sử dụng các hàm của thư viện pandas để trích xuất các cột Open, High, Low, Close và Volume.

*Hình 3.2: Tính năng trước và sau trích xuất*

**Tính năng ban đầu**

*Open, High, Low, Close, Volume,*

*Total bid, Total ask, Fbuy, Fsell, Room, …*

**Tính năng trích xuất**

*Open*

*High,*

*Low*

*Close*

*Volume*

## 3.3 THIẾT KẾ MẠNG BỘ NHỚ DÀI-NGẮN (LSTM)

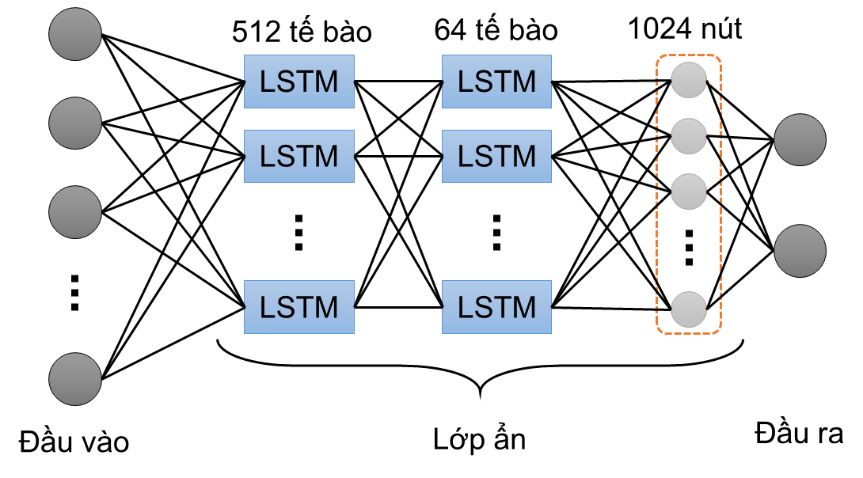
Mạng LSTM của luận văn được xây dựng bằng cách sử dụng Sequential từ thư viện Keras. Đây là một loại mô hình mạng nơ ron được xây dựng theo kiểu tuyến tính từ trên xuống dưới, mỗi lớp được thêm vào sau lớp trước đó.

Cấu trúc mạng LSTM bao gồm các lớp LSTM được xếp chồng lên nhau, mỗi lớp với số đơn vị (units) là 50. Lớp đầu tiên cho phép đầu ra của nó là một chuỗi, đá ứng yêu cầu của đầu tiếp theo. Các lớp sau đó không cần vì chúng chỉ cần kết quả của lớp trước đó.

Sau lỗi lớp LSTM, tôi có lớp Dropout với giá trị 0.2, được sử dụng để tránh overfit trong quá trình huấn luyện. Lớp Dropout giúp bỏ ngẫu nhiên một số đơn vị nơ ron trong quá trình huấn luyện, từ đó giảm thiểu overfit.

Cuối cùng, tôi có lớp Dense với một đơn vị đầu ra để dự đoán giá trị trong bài toán dự đoán chuỗi thời gian.

Sau khi xây dựng xong mạng LSTM, tôi sử dụng phương thức biên dịch *compile* để cấu hình mô hình. Tôi chọn bộ tối ưu hoá (optimizer) là Adam và hàm mất mát (loss) là Mean Squared Error với 100 epochs và kích thước batch là 32. Sau khi mô hình đã được biên dịch, tôi sử dụng phương thức *fit* để huấn luyện mạng LSTM. Dữ liệu huấn luyện *X\_train* và *y\_train* được đưa vào cùng với các tham số khác như số epoch và kích thước batch. Trong quá trình huấn luyện, mạng LSTM sẽ được cập nhật dần dần để tối ưu hoá mô hình và giảm thiểu sai số giữa kết quả dự đoán và giá trị thực tế. Quá trình này sẽ tiếp tục cho đến khi đạt được số epoch đã chỉ định.



##### Hình 3.2: Mô hình tổng quan của mạng LSTM

# CHƯƠNG 4 THỰC NGHIỆM

## 4.1 DỮ LIỆU

Tôi đánh giá mô hình dựa trên dữ liệu của 5 cổ phiếu chọn ngẫu nhiên thuộc nhiều ngành nghề khác nhau trong nhóm VN30 (top 30 cổ phiếu có giá trị vốn hoá thị trường và tính thanh khoản cao nhất) là ngân hàng TMCP Vietcombank (VCB), tập đoàn Vingroup (VIC), tập đoàn Hoà Phát (HPG), CTCP Vinamilk (VNM), CTCP Thế Giới Di Động (MWG). Tất cả dữ liệu cổ phiếu được lấy bằng API Investpy từ trang Investing.com, là một trang thông tin tài chính toàn cầu thuộc sở hữu của Fusion Media Ltd. Tôi chọn khoảng thời gian giao dịch trong khoảng 7 năm (từ năm 2015 đến năm 2022), gồm khoảng 1800 mẫu với giá mở phiên (open), giá cao nhất phiên (high), giá thấp nhất phiên (low), giá kết phiên (close) và khối lượng giao dịch (volume). Dữ liệu giao dịch không liên tục vì thị trường nghỉ vào cuối tuần và ngày lễ. Ngoài ra, một số mẫu bị mất mát trong quá trình thu thập và nguồn dữ liệu thiếu hụt.

Cổ phiếu của Vietcombank với giá mở phiên, cao nhất, thấp nhất, kết phiên và khối lượng với khoảng 1800 mẫu.

##### Bảng 4.1:Dữ liệu cổ phiếu Vietcombank

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| VCB | 2015-02-24 | 16716 | 16927 | 16558 | 16822 | 534023 |
| 2015-02-25 | 16822 | 17929 | 16822 | 17929 | 2901891 |
| 2015-02-26 | 17982 | 18245 | 17454 | 17718 | 2463903 |
| 2015-02-27 | 17771 | 18456 | 17718 | 18245 | 1557643 |
| 2015-03-02 | 18404 | 19511 | 18404 | 19458 | 3939034 |

Cổ phiếu của Vingroup với giá mở phiên, cao nhất, thấp nhất, kết phiên và khối lượng với khoảng 1800 mẫu.

##### Bảng 4.2:Dữ liệu cổ phiếu Vingroup

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| VIC | 2015-02-24 | 27886 | 27886 | 27652 | 27827 | 548730 |
| 2015-02-25 | 27710 | 27944 | 27652 | 27944 | 790900 |
| 2015-02-26 | 27944 | 28061 | 27769 | 27886 | 546700 |
| 2015-02-27 | 27886 | 28295 | 27886 | 28061 | 1348640 |
| 2015-03-02 | 28061 | 28412 | 27944 | 27944 | 1658100 |

Cổ phiếu của Hoà Phát với giá mở phiên, cao nhất, thấp nhất, kết phiên và khối lượng với khoảng 1800 mẫu.

##### Bảng 4.3:Dữ liệu cổ phiếu Hoà Phát

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| HPG | 2015-01-05 | 4566.4 | 4652.6 | 4523.3 | 4566.4 | 2193042 |
| 2015-01-06 | 4480.3 | 4609.5 | 4480.3 | 4609.5 | 2758305 |
| 2015-01-07 | 4609.5 | 4652.6 | 4566.4 | 4566.4 | 2465822 |
| 2015-01-08 | 4566.4 | 4609.5 | 4523.3 | 4566.4 | 1137517 |
| 2015-01-09 | 4566.4 | 4652.6 | 4523.3 | 4652.6 | 4335130 |

Cổ phiếu của Vinamilk với giá mở phiên, cao nhất, thấp nhất, kết phiên và khối lượng với khoảng 1800 mẫu.

##### Bảng 4.4:Dữ liệu cổ phiếu Vinamilk

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| VNM | 2015-02-24 | 66667 | 67014 | 66319 | 66667 | 101218 |
| 2015-02-25 | 66667 | 67361 | 65972 | 66667 | 145282 |
| 2015-02-26 | 66667 | 68403 | 66667 | 68056 | 279547 |
| 2015-02-27 | 68056 | 68750 | 67708 | 68403 | 101736 |
| 2015-03-02 | 69097 | 71528 | 68750 | 71528 | 418162 |

Cổ phiếu của Thế Giới Di Động với giá mở phiên, cao nhất, thấp nhất, kết phiên và khối lượng với khoảng 1800 mẫu.

##### Bảng 4.5:Dữ liệu cổ phiếu Thế Giới Di Động

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| MWG | 2015-02-24 | 20415 | 20415 | 19858 | 20044 | 90150 |
| 2015-02-25 | 20044 | 20044 | 19673 | 20044 | 189850 |
| 2015-02-26 | 19858 | 20044 | 19673 | 19673 | 52950 |
| 2015-02-27 | 19858 | 20044 | 19673 | 19858 | 235200 |
| 2015-03-02 | 19673 | 20230 | 19673 | 20230 | 230850 |

## XỬ LÝ DỮ LIỆU

### 4.2.1 Phân chia tập dữ liệu

Mỗi mã cổ phiếu là một tập dữ liệu, có 6 cột bao gồm Date, Open, High,

Low, Close và Volume và khoảng hơn 1800 dòng dữ liệu giá.

Dữ liệu được chia làm hai phần cho Traning và Testing. Tôi sử dụng 80-85% dữ liệu cổ phiếu (khoảng 1500 mẫu) cho Traning và sử dụng 15-20% dữ liệu cổ phiếu (khoảng 300 mẫu) cho Testing.

### Tiền xử lý dữ liệu

Tôi tiến hành chuẩn hoá bằng việc thay đổi giá trị các cột thành một tỉ lệ chung nhằm tăng hiệu suất của mô hình. Để chuẩn hoá tập dữ liệu huấn luyện, tôi dùng MinMaxScaler của Scikit-Learn với khoảng 0 và 1.

Tiếp theo, tôi tạo dữ liệu trong 60 bước thời gian rồi dùng numpy để chuyển nó thành mảng. Cuối cùng, tôi chuyển dữ liệu thành mảng 3 chiều với mẫu *X\_train*, 60 bước thời gian và một tính năng ở mỗi bước.

## 4.3 HUẤN LUYỆN

Tôi chọn ngẫu nhiên 5 cổ phiếu trong top 30 cổ phiếu lớn nhất của thị trường chứng khoán Việt Nam gồm ngân hàng Vietcombank, tập đoàn Vingroup, tập đoàn Hoà Phát, CTCP Vinamilk, CTCP Thế Giới Di Động, với khoảng thời gian giao dịch là 7 năm (từ năm 2015 đến năm 2022), với khoảng 1800 mẫu (khoảng 1500 mẫu cho Training và 300 mẫu cho Testing).

Môi trường thực nghiệm trên Macbook Pro, sử dụng hệ điều hành macOS Ventura 13.1, với chip Apple M1 Pro và RAM 16 GB. Trong quá trình huấn luyện, tôi sử dụng phần mềm PyCharm, Google Colab và ngôn ngữ Python cùng thư viện mã nguồn mở Keras do Google phát triển.

Tôi chọn thực nghiệm từng cổ phiếu bằng mô hình LSTM với 100 epochs. Qua quá trình thực nghiệm, tôi thu được kết quả như sau:

* Cổ phiếu Vietcombank (VCB):

##### Bảng 4.6: Kết quả đánh giá cổ phiếu Vietcombank

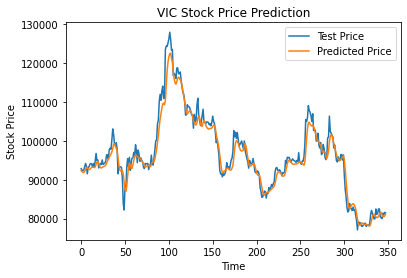
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **RMSE** | **MAE** | **MAPE** | **AR** | **RAR** |
| VCB | 1982.8 | 1400.2 | 0.0172 | 263.3 | 94.1 |

##### Hình 4.1:Biểu đồ giá cổ phiếu Vietcombank

* Cổ phiếu Vingroup (VIC):

##### Bảng 4.7: Kết quả đánh giá cổ phiếu Vingroup

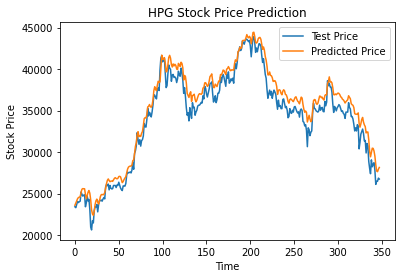
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **RMSE** | **MAE** | **MAPE** | **AR** | **RAR** |
| VIC | 2260.2 | 1557.4 | 0.0158 | 250.2 | 65 |

*Hình 4.2:**Biểu đồ giá cổ phiếu Vingroup*

* Cổ phiếu Hoà Phát (HPG):

##### Bảng 4.8:Kết quả đánh giá cổ phiếu Hoà Phát

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **RMSE** | **MAE** | **MAPE** | **AR** | **RAR** |
| HPG | 1325.6 | 1171.2 | 0.0352 | 207.7 | 136.7 |



##### Hình 4.3:Biểu đồ giá cổ phiếu Hoà Phát

* Cổ phiếu Vinamilk (VNM):

##### Bảng 4.9: Kết quả đánh giá cổ phiếu Vinamilk

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **RMSE** | **MAE** | **MAPE** | **AR** | **RAR** |
| VNM | 1121.5 | 772.7 | 0.0083 | 139.8 | -16.2 |

*Hình 4.4: Biểu đồ giá cổ phiếu Vinamilk*

* Cổ phiếu Thế Giới Di Động (MWG):

##### Bảng 4.10:Kết quả đánh giá cổ phiếu Thế Giới Di Động

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **RMSE** | **MAE** | **MAPE** | **AR** | **RAR** |
| MWG | 4503.4 | 3723 | 0.031 | 573.2 | 319.7 |

##### Hình 4.5:Biểu đồ giá cổ phiếu Thế Giới Di Động

## 4.4 ĐÁNH GIÁ

Tôi tìm RMSE, MAE, MAPE và AR bằng thuật toán dựa trên công thức đã trình bày ở chương 3. Giá trị được trình bày ở trong bảng trên là kết quả sau khi tính toán. Các độ đo RMSE, MAE và MAPE có giá trị thấp nghĩa là dự đoán giá đóng cửa gần đúng với dữ liệu thật và AR chỉ ra lợi nhuận khi giao dịch của những cổ phiếu này.

So sánh với đề tài “Dự đoán giá cổ phiếu bằng phương pháp học sâu không giám sát” của tác giả Nguyễn Minh Lợi [36], dựa vào kết quả thực nghiệm trên cùng cổ phiếu Vietcombank (VCB) nhưng khác thời gian dữ liệu, mô hình của tôi cho độ chính xác cao hơn về cả RMSE, MAE và MAPE. Tôi nhận thấy mô hình LSTM của tôi đạt được kết quả tương đối chính xác và có triển vọng để tiếp tục nghiên cứu và phát triển. Kết quả được so sánh chi tiết trong bảng 4.11 bên dưới.

##### Bảng 4.11:So sánh kết quả đánh giá cổ phiếu Vietcombank giữa 2 đề tài

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tác giả** | **RMSE**  (càng thấp càng tốt) | **MAE**  (càng thấp càng tốt) | **MAPE**  (càng thấp càng tốt | **AR**  (càng cao càng tốt) | **RAR**  (càng cao càng tốt) |
| Bùi Ngô Tôn Bách | 2260.2 | 1557.4 | 0.0158 | 250.2 | 65 |
| Nguyễn Minh Lợi | 5352.6 | 4026.8 | 0.0168 | Không đánh giá | 67.25 |

# 

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

* **KẾT LUẬN**

Trong luận văn này, tôi đã đề xuất các phương pháp dự đoán giá cổ phiếu bằng học máy. Bằng việc sử dụng mô hình LSTM, tôi đã dự đoán giá cổ phiếu, đánh giá hiệu quả mô hình và đưa ra một vài kết luận như sau:

* Phương pháp và mô hình mà tôi sử dụng có kết quả tương đối tốt.
* Độ chính xác có sự khác nhau giữa các cổ phiếu, điều này là do tính chất từng ngành nghề trong từng giai đoạn, cũng như kết quả kinh doanh của công ty tại thời điểm đó.
* Dữ liệu sử dụng còn nhiều hạn chế về sự liên tục, còn thiếu sót và mất mát nhiều.
* Việc giao dịch dựa trên kết quả dự đoán không được khuyến khích vì thuế và phí khiến độ đo AR bị giảm nhiều.
* **HƯỚNG PHÁT TRIỂN**
* Thử nghiệm trên các mô hình khác nhau để so sánh, đánh giá và tìm ra mô hình tối ưu nhất.
* Trích xuất đặc trưng quan trọng và tối ưu mô hình để kết quả dự đoán chính xác hơn.
* Ứng dụng Internet để cập nhật mô hình, xử lý dữ liệu theo thời gian thực, hỗ trợ việc giao dịch ngay trong phiên.
* Áp dụng mô hình lên giao dịch chứng khoán phái sinh.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Li, B., Hoi, S.C.H., 2012. On-line portfolio selection with moving average reversion, in: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012. |
| [2] | Box, G.E.P., Jenkins, G.M., 1976. Time series analysis: Forecasting and control. Journal of Time 31, 238– 242. |
| [3] | Pai, P.F., Lin, C.S., 2005. A hybrid arima and support vector machines model in stock price forecasting. Omega 33, 497–505. |
| [4] | Huang, S., Wang, H., 2006. Combining time-scale feature extractions with svms for stock index forecasting, in: Neural Information Processing, 13th International Conference, ICONIP 2006, Hong Kong, China, October 3-6, 2006, Proceedings, Part III, pp. 390–399. |
| [5] | Areekul, P., Senjyu, T., Toyama, H., Yona, A., 2010. A hybrid arima and neural network model for short-term price forecasting in deregulated market. IEEE Transactions on Power Systems Pwrs. |
| [6] | Chandar, S.K., Sumathi, M., Sivanandam, S.N., 2016. Prediction of stock market price using hybrid of wavelet transform and artificial neural network. Indian Journal of Science & Technology 9. |
| [7] | Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., Iosifidis, A., 2017. Forecasting stock prices from the limit order book using convolutional neural networks, in: 19th IEEE Conference on Business Informatics, CBI 2017, Thessaloniki, Greece, July 24-27, 2017, Volume 1: Conference Papers, pp. 7–12. |
| [8] | Rather, A.M., Agarwal, A., Sastry, V.N., 2015. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst. Appl. 42, 3234–3241. |
| [9] | Saad, E.W., Prokhorov, D.V., II, D.C.W., 1998. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. IEEE Trans. Neural Networks 9, 1456– 1470. |
| [10] | M. Teplan, "Fundamentals of EEG measurement," *Measurement science review ,* vol. 2, no. 2, pp. 1-11, 2002. |
| [11] | Ding, X., Zhang, Y., Liu, T., Duan, J., 2015. Deep learning for event-driven stock prediction, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pp. 2327–2333. |
| [12] | Nevmyvaka, Y., Feng, Y., Kearns, M.J., 2006. Reinforcement learning for optimized trade execution, in: Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, pp. 673–680. |
| [13] | Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., WardeFarley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680. |
| [14] | S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation, vol. 9, pp. 1735–1780, 1997 [14] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. 2015. |
| [15] | S. Hochreiter, “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions", 1997, World Scientific. |
| [16] | S. Hochreiter, J. Schmidhuber, ”Long short-term memory”, 1997. |
| [17] | Denny Britz, “Recurrent Neural Networks Tutorial”, 2015. |
| [18] | Y. H. S.-S. S. Jinming Zou, "Overview of artificial neural networks," *Methods Mol Biol,* vol. 458, no. 15, pp. 15-23, 2008. |
| [19] | Kang Zhang, Guoqiang Zhong, Junyu Dong, Shengke Wang, Yong Wang, “Stock Market Prediction Based on Generative Adversarial Network”, 2018. |
| [20] | Xingyu Zhou, Zhisong Pan, Guyu Hu, Siqi Tang, Cheng Zhao, “Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets”, 2018. |
| [21] | Dương Ngọc Mai Phương, Vũ Thị Phƣơng Anh, Đỗ Thị Trúc Đào & Nguyễn Hữu Tuấn, “Tác động của chính sách tiền tệ đến thị trƣờng chứng khoán: Bằng chứng tại Việt Nam”, 2015. |
| [22] | Đinh Bảo Ngọc, Nguyễn Chí Cường, “Các nhân tố ảnh hưởng đến chính sách cổ tức của các doanh nghiệp niêm yết trên thị trƣờng chứng khoán Việt Nam”, 2014. |
| [23] | Đỗ Văn Thành, Nguyễn Minh Hải, “Phân tích và dự báo chỉ số thị trường chứng khoán bằng sử dụng chỉ số báo trước“, 2016. |
| [24] | Thân Quang Khoát, “Học Máy (Machine Learning)”, 2016. |
| [25] | J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 81, no. 10, pp. 3088-3092, 1984. |
| [26] | M. Minsky, "Overview of Artificial Neural Networks," in *Perceptrons*, MIT Press , 1969, pp. 15-23. |
| [27] | Y. P. S. a. P. F. Bengio, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks,* vol. 5, no. 2, pp. 157-166, 1994. |
| [28] | Y. e. a. Yu, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation,* vol. 31, no. 7, pp. 1235-1270, 2019. |
| [29] | K. a. R. N. O'Shea, An introduction to convolutional neural networks., arXiv , 2015. |
| [30] | P. Bambharolia, "Overview of convolutional neural networks," in *Proceedings of the Conference: International Conference on Academic Research in Engineering and Management*, 2017. |
| [31] | P. Bambharolia, "Overview of convolutional neural networks," in *roceedings of the Conference: International Conference on Academic Research in Engineering and Management*, 2017. |
| [32] | R. e. a. Yamashita, "Convolutional neural networks: an overview and application in radiology.," *Insights into imaging,* vol. 9, no. 4, pp. 611-629, 2018. |
| [33] | P. M. S. a. G. D. Dayan, Unsupervised learning, The MIT encyclopedia of the cognitive sciences , 1999. |
| [34] | Z. Ghahramani, "Unsupervised learning," in *Summer School on Machine Learning*, Springer, 2003. |
| [35] | S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint, 2016. |
| [36] | Nguyễn Minh Lợi, “Dự đoán giá cổ phiếu bằng phương pháp học sâu không giám sát”, 2020. |