**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

=========\*\*\*=========

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

NGÀNH HỆ THỐNG THÔNG TIN



NGHIÊN CỨU VỀ DEEP LEARNING VÀ ỨNG DỤNG GIẢI BÀI TOÁN DỰ ĐOÁN QUAN HỆ

THUỐC- BỆNH

**CBHD: TS. Nguyễn Văn Tỉnh**

**Sinh viên: Phạm Ngọc Huy**

**Mã số sinh viên: 1141260130**

**Hà Nội, 2023**

LỜI CẢM ƠN

Với tình cảm sâu sắc và chân thành nhất, em xin được phép bày tỏ lòng biết ơn đến tất cả các cá nhân và tổ chức đã tạo điều kiện hỗ trợ, giúp đỡ em trong suốt quá trình học tập và nghiên cứu đề tài “Nghiên cứu về Deep learning và ứng dụng giải bài toán dự đoán quan hệ Thuốc - Bệnh”. Trong suốt thời gian từ khi bắt đầu học tập tại trường đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý thầy cô và bạn bè. Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý thầy cô ở khoa công nghệ thông tin đã truyền đạt vốn kiến thức quý báu cho em trong suốt thời gian học tập tại trường. Nhờ có những lời hướng dẫn, dạy bảo của các thầy, các cô nên đề tài tốt nghiệp của em mới có thể hoàn thiện. Một lần nữa, em xin chân thành cảm ơn thầy Nguyễn Văn Tỉnh – người đã trực tiếp giúp đỡ, quan tâm, hướng dẫn em hoàn thành tốt bài báo cáo này trong thời gian qua. Bước đầu đi vào thực tế của em còn hạn chế và còn nhiều bỡ ngỡ nên không tránh khỏi những thiếu sót, em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để kiến thức của em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao kỹ năng nghề nghiệp của em.

Em xin chân thành cảm ơn!

Hà Nội, ngày tháng 09 năm 2023

*Sinh viên thực hiện*

**Huy**

Phạm Ngọc Huy

MỞ ĐẦU

**Lý do chọn đề tài**

Trong những năm gần đây, Deep Learning đã đạt được những thành tựu ấn tượng trong nhiều lĩnh vực, từ thị giác máy tính đến xử lý ngôn ngữ tự nhiên. Khả năng của Deep Learning trong việc học và trích xuất thông tin từ dữ liệu lớn đã đặt nền móng cho việc áp dụng nó vào các bài toán trong lĩnh vực y học và dược phẩm.

Mối quan hệ giữa thuốc và bệnh là một phần quan trọng trong lĩnh vực y học, có thể ảnh hưởng đến quá trình điều trị và quản lý bệnh. Dự đoán quan hệ này có thể giúp cải thiện việc chẩn đoán bệnh, đề xuất phương pháp điều trị và dự đoán tác dụng phụ của thuốc. Việc nắm vững mối quan hệ giữa thuốc và bệnh có thể hỗ trợ các quyết định lâm sàng, tối ưu hóa phác đồ điều trị, và cung cấp tri thức mới cho ngành y học. Ứng dụng có thể trải dài từ hỗ trợ bác sĩ trong việc đưa ra quyết định chẩn đoán đến việc phát triển mới các phương pháp điều trị hiệu quả.

Lĩnh vực dự đoán quan hệ Thuốc - Bệnh vẫn đang được khám phá và phát triển, cung cấp cơ hội cho các nghiên cứu mới và sáng tạo. Sự kết hợp giữa Deep Learning và y học đang tạo ra những khám phá thú vị và có tiềm năng thay đổi cách chúng ta tiếp cận y tế. Do đó, em xin lựa chọn đề tài “Nghiên cứu về Deep learning và ứng dụng giải bài toán dự đoán quan hệ Thuốc - Bệnh (Drug-disease)”. Nghiên cứu đề tài này sẽ giúp em phát triển kỹ năng trong việc áp dụng Deep Learning vào dữ liệu y học, từ việc xử lý dữ liệu y tế phức tạp đến xây dựng và đánh giá mô hình dự đoán. Điều này không chỉ giúp em có kiến thức về cả hai lĩnh vực mà còn mở rộng khả năng giải quyết vấn đề và nghiên cứu trong tương lai.

Nội dung của báo cáo gồm 3 chương:

Chương 1: Tổng quan về Deep Learning và ứng dụng trong các bài toán sinh học.

Chương 2: Một mô hình Deep Learning trong giải bài toán dự đoán quan hệ thuốc – bệnh.

Chương 3: Thực nghiệm và kết quả.

MỤC LỤC

[DANH MỤC TỪ VIẾT TẮT 1](#_Toc145916644)

[DANH MỤC HÌNH VẼ, ĐỒ THỊ 2](#_Toc145916645)

[DANH MỤC BẢNG 3](#_Toc145916646)

[CHƯƠNG 1: TỔNG QUAN VỀ DEEP LEARNING VÀ ỨNG DỤNG TRONG CÁC BÀI TOÁN SINH HỌC 4](#_Toc145916647)

[1.1. Tổng quan về Deep Learning 4](#_Toc145916648)

[1.1.1. Giới thiệu về Deep Learning 4](#_Toc145916649)

[1.1.2. Đặc điểm của Deep Learning 4](#_Toc145916650)

[1.1.3. Các mô hình Deep Learning tiêu biểu 6](#_Toc145916651)

[1.1.4. Một số ứng dụng của Deep Learning 10](#_Toc145916652)

[1.1.5. Một số đại lượng đánh giá mô hình 11](#_Toc145916653)

[1.2. Ứng dụng của Deep Learning trong các bài toán sinh học 16](#_Toc145916654)

[1.2.1. Một số đối tượng sinh học đáng quan tâm 16](#_Toc145916655)

[1.2.2. Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh và các đối tượng sinh học 21](#_Toc145916656)

[1.3. Tổng kết chương 23](#_Toc145916657)

[CHƯƠNG 2. MỘT MÔ HÌNH ỨNG DỤNG DEEP LEARNING DỰ ĐOÁN QUAN HỆ THUỐC – BỆNH 25](#_Toc145916658)

[2.1. Giới thiệu chung về bài toán 25](#_Toc145916659)

[2.2. Một mô hình ứng dụng Deep Learning trong bài toán nghiên cứu và ứng dụng Deep Learning trong dự đoán mối quan hệ thuốc – bệnh 26](#_Toc145916660)

[2.2.1. Mô hình của bài toán 26](#_Toc145916661)

[2.2.2. Đầu vào và đầu ra của bài toán 27](#_Toc145916662)

[2.2.3. Các bước thực hiện 28](#_Toc145916663)

[CHƯƠNG 3. THỰC NGHIỆM VÀ KẾT QUẢ 38](#_Toc145916664)

[3.1. Cài đặt và thực nghiệm 38](#_Toc145916665)

[3.1.1. Môi trường cài đặt 38](#_Toc145916666)

[3.1.2. Tham số của mô hình DRWBNCF 38](#_Toc145916667)

[3.1.3. Hướng dẫn chạy chương trình 39](#_Toc145916668)

[3.1.4. Thực nghiệm 41](#_Toc145916669)

[3.2. Kết quả thu được và so sánh với một số phương pháp khác 42](#_Toc145916670)

[3.2.1. Kết quả thu được 42](#_Toc145916671)

[3.2.2. So sánh với một số phương pháp khác 50](#_Toc145916672)

[3.2.3. Tổng kết chương 52](#_Toc145916673)

[KẾT LUẬN 53](#_Toc145916674)

[TÀI LIỆU THAM KHẢO 54](#_Toc145916675)

# DANH MỤC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **STT** | **Ký hiệu chữ viết tắt** | **Chữ viết đầy đủ** |
| 1 | GPU | Graphics Processing Unit |
| 2 | TPU | Tensor Processing Unit |
| 3 | D-FFNN | Deep Feedforward Neural Networks |
| 4 | CNNs | Convolutional Neural Networks |
| 5 | RNN | Recurrent Neural Networks |
| 6 | LSTM | Long Short-Term Memory |
| 7 | TP | True Positive |
| 8 | FP | False Positive |
| 9 | AUC | Area Under the Curve |
| 10 | ROC | Receiver Operating Characteristic |
| 11 | TPR | True Positive Rate |
| 12 | FPR | False Positive Rate |
| 13 | AUPR | Area Under the Precision-Recall Curve |
| 14 | RNA | Ribonucleic Acid |
| 15 | DRWBNCF | Drug Reposition Weighted Bilinear Neural Collaborative Filtering |

# DANH MỤC HÌNH VẼ, ĐỒ THỊ

[Hình 1‑1 Cấu trúc của D-FFNN [1] 7](#_Toc146568691)

[Hình 1‑2 Cấu trúc mô hình CNN [2] 8](#_Toc146568692)

[Hình 1‑3 Sơ đồ đo lường hiệu suất [3] 12](#_Toc146568693)

[Hình 1‑4 Biểu đồ ROC [4] 14](#_Toc146568694)

[Hình 1‑5 Một ví dụ về đường cong ROC sau khi huấn luyện mô hình 15](#_Toc146568695)

[Hình 1‑6 Một ví dụ về đường cong AUPR sau khi huấn luyện mô hình 16](#_Toc146568696)

[Hình 1‑7 RNA [8] 17](#_Toc146568697)

[Hình 1‑8 miRNA [9] 18](#_Toc146568698)

[Hình 1‑9 Mô hình PANDA áp dụng cho bài toán toán dự đoán quan hệ giữa bệnh – Long non-coding RNAs [5] 22](#_Toc146568699)

[Hình 1‑10 Mô hình VAEMDA áp dụng cho bài toán toán dự đoán quan hệ giữa bệnh – mRNAs [6] 23](#_Toc146568700)

[Hình 2‑1 Cấu trúc của mô hình DRWBNCF [7] 27](#_Toc146568701)

[Hình 2‑2 Tiền xử lý dữ liệu của DRWBNCF [7] 31](#_Toc146568702)

[Hình 2‑3 Tiền xử lý dữ liệu của DRWBNCF [7] 33](#_Toc146568703)

[Hình 2‑4 Dữ liệu đầu vào của mạng nơ-ron thuộc mô hình DRWBNCF [7] 35](#_Toc146568704)

[Hình 2‑5 Sơ đồ mạng nơ-ron của DRWBNCF [7] 36](#_Toc146568705)

[Hình 3‑1 PyCharm Edu 2022.2.2 39](#_Toc146568706)

[Hình 3‑2 Anaconda Navigator – quản lý môi trường ảo 40](#_Toc146568707)

[Hình 3‑3 Mã nguồn của mô hình DRWBNCF 40](#_Toc146568708)

[Hình 3‑4 Sử dụng file demo.sh để chạy mô hình 41](#_Toc146568709)

[Hình 3‑5 Mã nguồn các tham số 41](#_Toc146568710)

[Hình 3‑6 Mã nguồn các tham số 42](#_Toc146568711)

[Hình 3‑7 Biểu đồ AUROC cho lần huấn luyện thứ nhất 43](#_Toc146568712)

[Hình 3‑8 Biểu đồ AUROC cho lần huấn luyện thứ hai 44](#_Toc146568713)

[Hình 3‑9 Biểu đồ AUROC cho lần huấn luyện thứ ba 44](#_Toc146568714)

[Hình 3‑10 Biểu đồ AUROC cho lần huấn luyện thứ tư 44](#_Toc146568715)

[Hình 3‑11 Biểu đồ AUROC cho lần huấn luyện thứ năm 45](#_Toc146568716)

[Hình 3‑12 Biểu đồ AUROC cho lần huấn luyện thứ sáu 45](#_Toc146568717)

[Hình 3‑13 Biểu đồ AUROC cho lần huấn luyện thứ bảy 45](#_Toc146568718)

[Hình 3‑14 Biểu đồ AUROC cho lần huấn luyện thứ tám 46](#_Toc146568719)

[Hình 3‑15 Biểu đồ AUROC cho lần huấn luyện thứ chín 46](#_Toc146568720)

[Hình 3‑16 Biểu đồ AUROC cho lần huấn luyện thứ mười 46](#_Toc146568721)

[Hình 3‑17 Biểu đồ AUPR cho lần huấn luyện thứ nhất 47](#_Toc146568722)

[Hình 3‑18 Biểu đồ AUPR cho lần huấn luyện thứ hai 47](#_Toc146568723)

[Hình 3‑19 Biểu đồ AUPR cho lần huấn luyện thứ ba 48](#_Toc146568724)

[Hình 3‑20 Biểu đồ AUPR cho lần huấn luyện thứ tư 48](#_Toc146568725)

[Hình 3‑21 Biểu đồ AUPR cho lần huấn luyện thứ năm 48](#_Toc146568726)

[Hình 3‑22 Biểu đồ AUPR cho lần huấn luyện thứ sáu 49](#_Toc146568727)

[Hình 3‑23 Biểu đồ AUPR cho lần huấn luyện thứ bảy 49](#_Toc146568728)

[Hình 3‑24 Biểu đồ AUPR cho lần huấn luyện thứ tám 49](#_Toc146568729)

[Hình 3‑25 Biểu đồ AUPR cho lần huấn luyện thứ chín 50](#_Toc146568730)

[Hình 3‑26 Biểu đồ AUPR cho lần huấn luyện thứ mười 50](#_Toc146568731)

# DANH MỤC BẢNG

[Bảng 3.1 Kết quả sau khi huấn luyện mô hình DRWBNCF trên 3 tập dữ liệu so sánh với kết quả của các phương pháp khác đã được công bố trước đó [7] 49](#_Toc145855602)

[Bảng 3.2 Kết quả sau khi huấn luyện mô hình DRWBNCF trên 3 tập dữ liệu so sánh với kết quả của các phương pháp khác đã được công bố trước đó [7] 49](#_Toc145855603)

# TỔNG QUAN VỀ DEEP LEARNING VÀ ỨNG DỤNG TRONG CÁC BÀI TOÁN SINH HỌC

## Tổng quan về Deep Learning

### Giới thiệu về Deep Learning

Deep Learning là một lĩnh vực quan trọng trong trí tuệ nhân tạo (Artificial Intelligent), cho phép máy tính học hỏi và hiểu các đặc trưng phức tạp từ dữ liệu. Điều đó cho phép chúng ta xây dựng các mô hình máy tính có khả năng giải quyết các vấn đề phức tạp mà trước đây chỉ có con người mới có thể thực hiện. Deep Learning sử dụng mạng nơ-ron nhân tạo (neural networks) có nhiều lớp ẩn (hidden layers) để biểu diễn và học các đặc trưng phức tạp từ dữ liệu đầu vào (ảnh, video, văn bản, số học, v.v.). Hiện nay, các mô hình deep learning có thể được huấn luyện bằng cách sử dụng các thuật toán như Backpropagation và Stochastic Gradient Descent.[11]

Deep learning đã chứng minh được hiệu quả của mình trong nhiều tác vụ AI quan trọng như phân loại ảnh, nhận dạng giọng nói, dịch máy, xác định đối tượng trong video, tự động lái xe và nhiều ứng dụng khác. Tuy Deep learning mang lại nhiều lợi ích, nhưng cũng đòi hỏi một lượng dữ liệu lớn và tài nguyên tính toán cao. Việc huấn luyện mô hình sẽ tốn nhiều thời gian và có thể yêu cầu sử dụng GPU hoặc TPU để đạt được hiệu suất tối ưu.

Tóm lại, Deep learning là một lĩnh vực hứa hẹn của trí tuệ nhân tạo, đem lại những tiềm năng tuyệt vời trong việc giải quyết các bài toán phức tạp. Tuy còn gặp nhiều khó khăn, thách thức nhưng với sự phát triển của công nghệ và nguồn lực, deep learning sẽ tiếp tục định hình tương lai của trí tuệ nhân tạo và tạo ra những ứng dụng đột phá trong cuộc sống hàng ngày của chúng ta.

### Đặc điểm của Deep Learning

* Học tự động

Deep Learning cho phép máy tính học tự động thông qua việc điều chỉnh trọng số và hệ số tự do trong mạng nơ-ron. Không cần phải xác định và lập trình cụ thể các đặc trưng hay quy tắc giải quyết vấn đề, điều này giúp nó phù hợp với nhiều loại dữ liệu và tác vụ phức tạp.

* Biểu diễn hiệu quả thông tin phức tạp

Deep Learning có khả năng biểu diễn thông tin từ dữ liệu đầu vào một cách hiệu quả và tự động tìm ra các đặc trưng ẩn (hidden features) có ý nghĩa trong quá trình xử lý thông tin. Điều này cho phép nó xử lý các vấn đề phức tạp và đa dạng như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, và nhiều ứng dụng khác.

* Cấu trúc của mạng nơ-ron sâu

Deep Learning sử dụng mạng nơ-ron với nhiều lớp (layers), bao gồm các lớp đầu vào (input layer), các lớp ẩn (hidden layers), và lớp đầu ra (output layer). Các lớp ẩn giúp mô hình học các biểu diễn phức tạp từ dữ liệu đầu vào, và các lớp đầu ra đưa ra kết quả dự đoán của mô hình.

* Sử dụng hàm kích hoạt (activation function) phi tuyến tính

Các hàm kích hoạt được áp dụng cho mỗi nơ-ron trong mạng để giới hạn đầu ra của nó và tạo tính phi tuyến (non-linearity) cho mô hình. Các hàm kích hoạt phổ biến bao gồm Sigmoid, Tanh, ReLU (Rectified Linear Unit) và các biến thể của chúng.

* Đào tạo (training) và tối ưu hóa

Để học từ dữ liệu, Deep Learning sử dụng các thuật toán tối ưu hóa như Gradient Descent để điều chỉnh trọng số và hệ số tự do sao cho mô hình có thể dự đoán chính xác nhất. Quá trình này được thực hiện thông qua quá trình lan truyền thuận (forward propagation) và lan truyền ngược (backward propagation) trong mạng.

* Khối lượng tính toán lớn

Để đạt được hiệu suất cao, Deep Learning đòi hỏi sử dụng tài nguyên tính toán mạnh mẽ như GPU và TPU, đặc biệt khi xử lý các mô hình có hàng triệu hoặc thậm chí hàng tỷ tham số.

* Ứng dụng đa dạng

Deep Learning đã có những tiến bộ đáng kể trong nhiều lĩnh vực như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, xe tự lái, robotica, tài chính, y tế và nhiều ứng dụng khác.

### Các mô hình Deep Learning tiêu biểu

#### Mô hình mạng nơ-ron tiến thẳng sâu (Deep Feedforward Neural Networks)

D-FFNN là một dạng cơ bản của mạng nơ-ron sâu, không có sự kết nối ngược (recurrent connections) trong các lớp, điều này làm cho nó thích hợp cho việc xử lý dữ liệu có cấu trúc đơn giản. [11] Cấu trúc của D-FFNN bao gồm ít nhất ba loại lớp:

* Lớp Đầu Vào (Input Layer): Lớp này chứa các nơ-ron tương ứng với các đặc trưng hoặc thuộc tính của dữ liệu đầu vào. Mỗi nơ-ron trong lớp này tương ứng với một đặc trưng riêng biệt và chấp nhận dữ liệu đầu vào tương ứng.
* Các Lớp Ẩn (Hidden Layers): D-FFNN có thể bao gồm một hoặc nhiều lớp ẩn giữa lớp đầu vào và lớp đầu ra. Mỗi lớp ẩn có thể chứa nhiều nơ-ron và thực hiện việc trích xuất và biểu diễn thông tin từ dữ liệu đầu vào. Các lớp ẩn thường sử dụng các hàm kích hoạt phi tuyến để tạo ra các biểu diễn phi tuyến của dữ liệu.
* Lớp Đầu Ra (Output Layer): Lớp này chứa các nơ-ron đại diện cho đầu ra dự đoán của mô hình. Số lượng nơ-ron trong lớp này thường phụ thuộc vào loại bài toán bạn đang giải quyết, có thể là phân loại, hồi quy hoặc các bài toán khác.



Hình ‑ Cấu trúc của D-FFNN [1]

Quá trình huấn luyện D-FFNN bao gồm việc cập nhật trọng số của các kết nối nơ-ron dựa trên giá trị sai khác giữa đầu ra dự đoán và đầu ra thực tế trong tập dữ liệu đào tạo. Thuật toán backpropagation và các thuật toán tối ưu hóa như stochastic gradient descent thường được sử dụng để điều chỉnh trọng số này. D-FFNN đã được áp dụng rộng rãi trong nhiều lĩnh vực như phân loại hình ảnh, nhận dạng văn bản, và dự báo.

#### Mô hình mạng nơ-ron tích chập Convolutional Neural Networks (CNNs)

CNNs là một loại mạng nơ-ron sử dụng trong Deep Learning, được thiết kế đặc biệt để xử lý dữ liệu hình ảnh và video. Các lớp của CNNs bao gồm

* Các lớp Convolutional: Các lớp Convolutional trong CNNs sử dụng các bộ lọc (filters) để trích xuất các đặc trưng của hình ảnh. Các bộ lọc này được áp dụng trên toàn bộ hình ảnh để tạo ra các feature maps. Các feature maps này được đưa vào các lớp Activation để tạo ra các feature maps kích hoạt.
* Các lớp Pooling: Các lớp Pooling trong CNNs được sử dụng để giảm kích thước của đầu ra từ các lớp Convolutional. Các lớp Pooling thường sử dụng các phép toán như Max Pooling hoặc Average Pooling để lấy giá trị lớn nhất hoặc trung bình của các pixel trong một vùng nhất định của feature maps.
* Các lớp Activation: Các lớp Activation trong CNNs được sử dụng để áp dụng các hàm kích hoạt (activation functions) lên các feature maps kích hoạt. Các hàm kích hoạt này giúp tạo ra các feature maps phi tuyến tính, giúp mô hình có khả năng học được các đặc trưng phức tạp của hình ảnh.
* Các lớp Fully Connected: Các lớp Fully Connected trong CNNs được sử dụng để kết nối các feature maps với nhau và tạo ra đầu ra cuối cùng của mô hình. Các lớp Fully Connected thường được sử dụng ở cuối của mô hình để phân loại các đối tượng trong hình ảnh.



Hình ‑ Cấu trúc mô hình CNN [2]

CNNs được sử dụng rộng rãi trong các tác vụ như nhận dạng khuôn mặt, phân loại hình ảnh, nhận dạng vật thể, phân tích video và nhiều ứng dụng khác trong lĩnh vực thị giác máy tính.

#### Mạng nơ-ron hồi quy (Recurrent Neural Networks)

Mạng nơ-ron hồi quy (RNNs) là một loại mạng nơ-ron sử dụng trong Deep Learning, được thiết kế đặc biệt để xử lý dữ liệu tuần tự như văn bản, âm thanh và video. Các lớp Recurrent trong RNNs được sử dụng để lưu trữ thông tin về các phần tử trước đó trong dãy dữ liệu và thường sử dụng các hàm kích hoạt như Tanh hoặc ReLU để tính toán giá trị của các hidden state. [11]

Có nhiều kiến trúc RNNs khác nhau được sử dụng trong Deep Learning, bao gồm Simple RNNs, LSTM và GRU. [11]

* Simple RNNs là kiến trúc RNNs đơn giản nhất, sử dụng các lớp Recurrent đơn giản để lưu trữ thông tin về các phần tử trước đó trong dãy dữ liệu. Tuy nhiên, Simple RNNs có vấn đề về vanishing gradient khi huấn luyện trên các dãy dữ liệu dài.
* LSTM (Long Short-Term Memory) là một kiến trúc RNNs được thiết kế để giải quyết vấn đề vanishing gradient của Simple RNNs. LSTM sử dụng các cổng (gates) để kiểm soát thông tin được lưu trữ và truyền qua các lớp Recurrent. Các cổng này bao gồm cổng quên (forget gate), cổng đầu vào (input gate) và cổng đầu ra (output gate), giúp LSTM có khả năng lưu trữ thông tin trong thời gian dài và tránh vanishing gradient.
* GRU (Gated Recurrent Unit) là một kiến trúc RNNs tương tự như LSTM, nhưng sử dụng ít cổng hơn. GRU có cổng quên và cổng đầu vào kết hợp thành một cổng cập nhật (update gate), giúp GRU có khả năng lưu trữ thông tin trong thời gian dài và tránh vanishing gradient.

RNNs được sử dụng rộng rãi trong các tác vụ như dịch máy, nhận dạng giọng nói, phân tích tín hiệu thị giác, phân tích dữ liệu chuỗi thời gian và nhiều ứng dụng khác trong lĩnh vực xử lý ngôn ngữ tự nhiên và thị giác máy tính

### Một số ứng dụng của Deep Learning

Deep learning đã có ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ xử lý ngôn ngữ tự nhiên đến nhận dạng hình ảnh và thậm chí là y học.

* Xử lý Ngôn Ngữ Tự Nhiên (NLP):
  + Dịch máy: Deep learning đã cải thiện đáng kể khả năng dịch máy, giúp dịch tự động và hiểu ngôn ngữ giữa các quốc gia và văn hóa khác nhau.
  + Phân loại văn bản: Deep learning có thể phân loại văn bản thành các chủ đề, cảm xúc hoặc định dạng.
  + Phân tích ý kiến: Deep learning giúp xác định ý kiến, cảm xúc của người dùng từ dữ liệu mạng xã hội, đánh giá sản phẩm, bài viết trên trang web.
* Nhận Dạng Hình Ảnh:
  + Nhận dạng đối tượng và vật thể: Deep learning có khả năng nhận dạng đối tượng, vật thể trong hình ảnh, hỗ trợ việc phân loại và nhận biết.
  + Xử lý ảnh y khoa: Deep learning giúp dự đoán, phát hiện và phân loại các vấn đề y khoa trong hình ảnh chụp từ máy MRI, CT scan và ảnh siêu âm.
  + Xử lý hình ảnh tự động: Deep learning có thể tự động tạo ra mô tả hoặc mở rộng ảnh, làm mượt hình ảnh và dự đoán hình ảnh tiếp theo.
* Nhận Dạng Âm Thanh và Ngôn Ngữ Giọng Nói:
  + Nhận dạng giọng nói: Deep learning giúp chuyển đổi giọng nói thành văn bản, dự đoán người nói, và thậm chí phân biệt các giọng địa phương.
  + Tạo âm thanh: Deep learning có thể tạo âm thanh, nhạc, giọng nói tự nhiên, hỗ trợ trong sản xuất âm thanh.
* Ô tô tự hành và Robot:
  + Xe tự hành: Deep learning hỗ trợ trong việc xử lý dữ liệu từ cảm biến, nhận dạng biển số, xác định vị trí, và quyết định lái xe.
  + Robot hợp tác: Deep learning giúp các robot hợp tác với con người và môi trường xung quanh một cách thông minh hơn.
* Y Học và Dược:
  + Chẩn đoán y khoa: Deep learning có thể phát hiện bệnh qua hình ảnh y khoa, nhận biết ung thư, dự đoán bệnh tật, và hỗ trợ bác sĩ trong việc ra quyết định.
  + Phát triển thuốc: Deep learning giúp tạo ra mô hình dự đoán tương tác giữa các phân tử để phát triển các loại thuốc mới.
* Tài chính và Kinh doanh:
  + Dự đoán thị trường tài chính: Deep learning có thể dự đoán biến động giá cổ phiếu, thay đổi thị trường, hỗ trợ trong quản lý rủi ro tài chính.
  + Phân loại giao dịch gian lận: Deep learning giúp phát hiện giao dịch gian lận trong các giao dịch tài chính.

### Một số đại lượng đánh giá mô hình

Trong lĩnh vực học máy, việc tính toán hiệu suất của mô hình là một điều rất quan trọng. Việc đánh giá mô hình giúp chúng ta giải quyết những vấn đề sau:

* + Mô hình đã được huấn luyện thành công hay chưa?
  + Mức độ thành công của mô hình tốt đến đâu?
  + Khi nào nên dừng quá trình huấn luyện?
  + Khi nào nên cập nhật mô hình?

Đánh giá một mô hình có tốt hay không thường được thực hiện trên dữ liệu mà mô hình chưa được huấn luyện. Tỷ lệ thường thấy của một tập dữ liệu huấn luyện so với tập dữ liệu thử nghiệm là 70% và 30%. Những bài toán khác nhau sẽ có những tiêu chí đánh giá khác nhau, vì vậy cần phải xác định rõ thứ tự ưu tiên của các tiêu chí cho việc đánh giá mô hình.

Một số đại lượng đánh giá mô hình được đề cập dưới đây

#### Accuracy

Accuracy (độ chính xác) chỉ đơn giản đánh giá mô hình thường xuyên dự đoán đúng đến mức nào. Độ chính xác là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu.

Tuy nhiên, accuracy không phải lúc nào cũng phản ánh đầy đủ khả năng của mô hình, đặc biệt khi tỷ lệ phân phối giữa các lớp là không đồng đều.

#### Confusion Matrix

Confusion matrix là một kỹ thuật đánh giá hiệu năng của mô hình cho các bài toán phân lớp. Confusion matrix là một ma trận thể hiện số lượng điểm dữ liệu thuộc vào một class và được dự đoán thuộc vào class. Ngoài ra, nó còn cung cấp thêm thông tin về tỉ lệ phân lớp đúng giữa các lớp, hay giúp phát hiện các lớp có tỉ lệ phân lớp nhầm cao nhờ vào các khái niệm True (False) Positive (Negative).[3]



Hình ‑ Sơ đồ đo lường hiệu suất [3]

* + True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)
  + True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)
  + False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error
  + False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error

Confusion Matrix giúp hiểu rõ hơn về khả năng của mô hình trong việc phân loại đúng và sai, từ đó đưa ra quyết định cải thiện hoặc điều chỉnh mô hình để tăng hiệu suất.

#### Precision và Recall

Với những thông tin có được từ Confusion matrix, chúng ta có thể định lượng độ hiệu quả của mô hình qua nhiều thang đo khác nhau. Precision và Recall là hai thang đo quan trọng trong số đó.

Precision cho ta biết được rằng trong số các điểm dữ liệu được mô hình phân loại vào lớp Positive, có bao nhiêu điểm dữ liệu thực sự thuộc về lớp Positive. Mặt khác, Recall giúp ta biết được có bao nhiêu điểm dữ liệu thực sự ở lớp Positive được mô hình phân lớp đúng trong mọi điểm dữ liệu thực sự ở lớp Positive.

Precision và Recall có giá trị trong [0,1], hai giá trị này càng gần với 1 thì mô hình càng chính xác. Precision càng cao đồng nghĩa với các điểm được phân loại càng chính xác. Recall càng cao cho thể hiện cho việc ít bỏ sót các điểm dữ liệu đúng.

#### AUC – ROC

AUC - ROC là một phương pháp tính toán hiệu suất của một mô hình phân loại theo các ngưỡng phân loại khác nhau.

Đường cong ROC là một biểu đồ thể hiện hiệu suất của một mô hình phân loại ở các ngưỡng khác nhau. Nó được tạo ra bằng cách vẽ đường cong giữa tỷ lệ dương tính thực sự (True Positive Rate - TPR) và tỷ lệ giả dương tính (False Positive Rate - FPR) trên trục y và trục x, tương ứng.

AUC là một chỉ số đánh giá hiệu suất của mô hình phân loại. Nó đo diện tích phía dưới đường cong ROC, biểu thị tỷ lệ giữa tỷ lệ dương tính giả và tỷ lệ dương tính thực sự.

A diagram of a curve

Description automatically generated

Hình ‑ Biểu đồ ROC [4]

* TPR (True Positive Rate/Sentivity/Recall): Biểu diễn tỷ lệ phân loại chính xác các mẫu dương tính trên tất cả các mẫu dương tính

TPR càng cao thì các mẫu dương tính càng được phân loại chính xác.

* Specificity: Biểu diễn tỷ lệ phân loại chính xác các mẫu âm tính trên tất cả các mâu âm tính, được tính theo công thức:
* FPR (False Positive Rate/Fall-out): Biểu diễn tỷ lệ gắn nhãn sai các mẫu âm tính thành dương tính trên tất cả các mẫu âm tính FPR càng cao thì Specificity càng giảm và số lượng các mẫu âm tính bị gắn nhãn sai càng lớn. FPR được tính theo công thức như sau:

Đánh giá mô hình qua chỉ số AUC: Chỉ số AUC càng gần 1 thì mô hình càng phân loại chính xác. AUC càng gần 0.5 thì hiệu suất phân loại càng tệ còn nếu gần 0 thì mô hình sẽ phân loại ngược kết quả (phân loại dương tính thành âm tính và ngược lại).

A graph with a line

Description automatically generated

Hình ‑ Một ví dụ về đường cong ROC sau khi huấn luyện mô hình

#### AUPR

AUPR là viết tắt của "Area Under the Precision-Recall Curve", là một phương pháp đánh giá hiệu suất của một mô hình phân loại. Nó được sử dụng để đo lường khả năng của một mô hình phân loại trong việc phát hiện các trường hợp dương tính (positive cases) trong một tập dữ liệu.

AUPR được tính bằng cách tính toán diện tích dưới đường cong Precision-Recall Curve, trong đó Precision là tỷ lệ giữa số lượng trường hợp dương tính được phân loại đúng và tổng số trường hợp được phân loại là dương tính, còn Recall là tỷ lệ giữa số lượng trường hợp dương tính được phân loại đúng và tổng số trường hợp dương tính trong tập dữ liệu.

AUPR là một phương pháp đánh giá hiệu suất phổ biến trong các bài toán phân loại mà tập dữ liệu có sự mất cân bằng giữa các lớp, tức là số lượng trường hợp dương tính và âm tính không đồng đều. Một mô hình phân loại tốt sẽ có AUPR cao, tức là có Precision và Recall cao.

Tuy nhiên, việc đánh giá mô hình chỉ dựa trên AUPR không đủ để đánh giá toàn diện hiệu suất của mô hình. Nên kết hợp AUPR với các phương pháp đánh giá khác như AUC (Area Under the ROC Curve) để đánh giá toàn diện hiệu suất của mô hình.

A graph showing a line

Description automatically generated with medium confidence

Hình ‑ Một ví dụ về đường cong AUPR sau khi huấn luyện mô hình

## Ứng dụng của Deep Learning trong các bài toán sinh học

### Một số đối tượng sinh học đáng quan tâm

#### Long non-coding RNAs

RNA (ribonucleic acid) là một loại phân tử sinh học có vai trò quan trọng trong quá trình di truyền thông tin di truyền và tổ chức sinh học trong tế bào của tất cả các hệ thống sống. RNA thường được tạo ra từ DNA thông qua quá trình gọi là transcription (chuyển mã), và nó có nhiều chức năng khác nhau trong tế bào.[8] RNA được chia làm 2 dạng coding RNA và non-coding RNA.



Hình ‑ RNA [8]

IncRNA (long non-coding RNA) là một dạng của non-coding RNA có độ dài lớn hơn 200 nucleotide. Các lncRNA có thể có nhiều chức năng khác nhau, bao gồm tham gia vào việc chỉ đạo các phức hợp protein lớn đến DNA dẫn đến ảnh hưởng đến việc biểu hiện gen cụ thể, và trong việc sửa đổi biểu hiện hoặc số lượng các chuỗi mRNA tương ứng. lncRNA cũng có thể cung cấp một phương tiện để điều chỉnh một hệ thống tế bào ở nhiều cấp độ và hoạt động nhanh chóng đáp ứng với các tín hiệu bên ngoài như các tín hiệu chỉ đạo axon hoặc môi trường xung quanh.

Các nghiên cứu gần đây đã tổng hợp các báo cáo về lncRNA trong hệ thống sinh học, bao gồm cả vai trò của chúng trong quá trình phát triển thần kinh, ung thư, bệnh tim mạch và các bệnh lý khác. Một số lncRNA cũng được biết đến là có vai trò trong việc điều chỉnh quá trình biểu hiện gen và sự phát triển của các tế bào.

#### Micro RNAs

MiRNAs là một loại RNA nhỏ, có độ dài khoảng 21-25 nucleotide có vai trò quan trọng trong việc nhận dạng và tương tác với các mục tiêu gen.[9]



Hình ‑ miRNA [9]

MiRNAs đóng vai trò quan trọng trong nhiều quá trình sinh học như phát triển và chuyển hoá tế bào, chia tách tế bào, tự tử tế bào, và điều chỉnh quá trình đáp ứng miễn dịch. Chúng có thể ảnh hưởng đến hàng trăm hoặc thậm chí hàng ngàn gen khác nhau, tạo ra một mạng lưới phức tạp của sự tương tác gen. Sự thay đổi trong sự biểu hiện miRNA có thể góp phần vào nhiều bệnh tật khác nhau. Sự tăng cường hoặc giảm thiểu miRNA cụ thể có thể liên quan đến ung thư, bệnh tim mạch, bệnh thần kinh, và nhiều bệnh khác. Vì vậy, miRNAs đã trở thành một lĩnh vực nghiên cứu quan trọng trong y học phân tử và tạo ra triển vọng trong việc phát triển các phương pháp chẩn đoán và điều trị.

Nghiên cứu về miRNAs đã đóng góp vào sự hiểu biết sâu hơn về cơ chế điều chỉnh gen và tạo ra cơ hội cho việc phát triển các phương pháp mới để chẩn đoán và điều trị bệnh. Các ứng dụng tiềm năng của miRNAs bao gồm việc sử dụng chúng như biomarkers (chỉ số sinh học) cho các bệnh, phát triển thuốc mục tiêu miRNA, và thậm chí chỉnh sửa sự biểu hiện miRNA để điều trị bệnh.

#### Bệnh

Bệnh tật được định nghĩa là một tình trạng bất thường hoặc không bình thường trong cơ thể hoặc tế bào, dẫn đến sự suy yếu hoặc thay đổi chức năng của cơ thể. Các bệnh có thể phát triển do nhiều nguyên nhân, bao gồm các yếu tố di truyền, môi trường, lối sống và tương tác phức tạp giữa chúng. Phân loại bệnh, bao gồm:

* Bệnh nhiễm trùng (Infectious Diseases): Là các bệnh được gây ra bởi vi khuẩn, virus, nấm hoặc sinh vật khác xâm nhập vào cơ thể và gây ra sự nhiễm trùng. Ví dụ: cúm, HIV/AIDS, bệnh lao, ...
* Bệnh không nhiễm trùng (Non-infectious Diseases): Là các bệnh không liên quan đến tác động của vi sinh vật, thường xuất phát từ các yếu tố như di truyền, môi trường, lối sống, hoặc tương tác giữa chúng. Ví dụ: bệnh tim mạch, ung thư, tiểu đường, ...
* Bệnh di truyền (Genetic Diseases): Là các bệnh do lỗi trong gene hoặc DNA của người bệnh. Ví dụ: bệnh bạch cầu, hội chứng Down, bệnh thalassemia, ...
* Bệnh tự miễn (Autoimmune Diseases): Là các bệnh do hệ miễn dịch tấn công các cơ tế bào và mô của chính cơ thể. Ví dụ: bệnh lupus, bệnh tự miễn tăng viêm khớp, ...
* Bệnh dự phòng (Preventive Diseases): Là các bệnh có thể được ngăn ngừa bằng cách duy trì lối sống lành mạnh và thực hiện các biện pháp phòng ngừa như tiêm phòng, kiểm tra định kỳ. Ví dụ: bệnh tim mạch do lối sống không lành mạnh, ...

Hiện nay, y học dự đoán và tiến triển nghiên cứu để cải thiện việc chẩn đoán, điều trị và dự phòng bệnh. Công nghệ như dịch vụ y tế trực tuyến, phân tích dữ liệu y học, và phát triển thuốc mới đều đóng góp vào sự tiến bộ trong lĩnh vực này.

#### Thuốc

Thuốc chỉ những hợp chất hóa học hoặc dược phẩm được sử dụng để chẩn đoán, điều trị, ngăn ngừa hoặc kiểm soát bệnh tật ở con người và động vật. Các loại thuốc bao gồm thuốc tây y, dược liệu thiên nhiên, thuốc thảo dược và các loại sản phẩm dược phẩm khác.

Thuốc thường được phát triển thông qua quá trình nghiên cứu và thử nghiệm kỹ thuật, bao gồm các giai đoạn như nghiên cứu phân tử, thử nghiệm trên tế bào, thử nghiệm trên động vật và cuối cùng thử nghiệm lâm sàng trên con người. Quá trình này đảm bảo rằng thuốc an toàn và hiệu quả trước khi được phê chuẩn sử dụng trong thực tế.

Thuốc có thể hoạt động theo nhiều cách khác nhau. Một số loại thuốc tác động trực tiếp lên các cơ quan hoặc tế bào cụ thể để điều chỉnh chức năng của chúng. Các loại thuốc khác có thể ngăn chặn hoặc can thiệp vào quá trình sinh học trong cơ thể để ngăn ngừa sự phát triển của bệnh hoặc giảm triệu chứng.

Thuốc được phân loại theo nhiều tiêu chí khác nhau, bao gồm cơ chế hoạt động, mục tiêu tác động, phân loại hóa học và cách chúng được sử dụng. Ví dụ, các antiviral drugs được thiết kế để ngăn chặn sự phát triển của virus trong cơ thể, trong khi các analgesic drugs giúp giảm đau.

Trong y học, việc sử dụng thuốc phải dựa trên các dẫn chứng lâm sàng và tuân thủ các quy tắc an toàn. Các y bác sĩ và nhà nghiên cứu y học cần đánh giá kỹ càng các lợi ích và rủi ro của việc sử dụng thuốc cho từng trường hợp cụ thể.

Tóm lại, thuốc là các hợp chất hóa học hoặc dược phẩm được sử dụng trong y học để chẩn đoán, điều trị, ngăn ngừa hoặc kiểm soát bệnh tật. Thuốc được phát triển qua quá trình nghiên cứu và thử nghiệm kỹ thuật, với mục tiêu đảm bảo tính an toàn và hiệu quả của chúng trong thực tế y học.

### Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh và các đối tượng sinh học

#### Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh – Long non-coding RNAs

Việc dự đoán cách mà lncRNA tương tác với các bệnh là một nhiệm vụ quan trọng, vì điều này có thể giúp ta hiểu rõ hơn về cơ chế trong việc phát triển và tiến triển của các bệnh. LncRNA có vai trò quan trọng trong nhiều quá trình trong cơ thể, và nhiều nghiên cứu đã liên kết lncRNA với nhiều bệnh khác nhau, như ung thư vú, bệnh bạch cầu và bệnh tim mạch. Hiểu rõ hơn về mối quan hệ giữa lncRNA và các bệnh có thể giúp ta xác định được các chỉ số có thể dùng để chẩn đoán bệnh, cũng như xác định được những mục tiêu mới cho việc phát triển các loại thuốc và liệu pháp điều trị.

Tuy nhiên, việc xác định những mối quan hệ này không dễ dàng do tính chất phức tạp của quá trình sinh học, cùng với đó là lượng dữ liệu lớn. Vì vậy, việc tạo ra các phương pháp để dự đoán những mối quan hệ mới giữa lncRNA và bệnh có thể giúp nhanh chóng khám phá ra những chỉ số mới và các phương pháp điều trị tiềm năng.

Một trong số đó là phương pháp PANDA.[5] Phương pháp PANDA hoạt động bằng cách xây dựng một đồ thị đa dạng chứa các mối quan hệ đã biết giữa lncRNA và bệnh, mức biểu hiện gen và triệu chứng của các bệnh. Sau đó, các tác giả sử dụng một mô hình Graph Auto-encoder để học cách biểu diễn các đặc trưng của các nút và mối quan hệ trên đồ thị, và sử dụng mạng Neural Network để dự đoán các mối quan hệ tiềm năng.

A diagram of a gene editing process

Description automatically generated with medium confidence

Hình ‑ Mô hình PANDA áp dụng cho bài toán toán dự đoán quan hệ giữa bệnh – Long non-coding RNAs [5]

MiRNAs, hoặc MicroRNAs, là một dạng đặc biệt của RNA có khả năng tác động lên các quá trình sinh học cơ bản. Điều này bao gồm sự kiểm soát của miRNAs đối với quá trình như sự phân chia tế bào, sự phát triển của tế bào, quá trình khác biệt giữa các loại tế bào, sự tử vong của tế bào, quá trình tự tử, quản lý chuyển hóa, quá trình lão hóa của tế bào, truyền tín hiệu trong cơ thể, và thậm chí cả quá trình nhiễm virut. Do miRNAs đóng vai trò quan trọng trong nhiều khía cạnh của sinh học và có tiềm năng lớn để nghiên cứu và ứng dụng trong lĩnh vực y tế và sinh học phân tử nên sự quan tâm đối với miRNAs đang gia tăng đáng kể trong cộng đồng nghiên cứu khoa học.

Một trong số đó là việc nghiên cứu mối quan hệ giữa miRNAs – bệnh. Nhiều công trình nghiên cứu đã xác minh một số mối quan hệ này, và dấu vết đã được tìm thấy giữa sự thay đổi trong miRNAs và sự xuất hiện của nhiều bệnh, bao gồm các bệnh ung thư, bệnh tim mạch, bệnh tiểu đường và các bệnh khác. Việc này không chỉ giúp cải thiện sự hiểu biết về cơ chế phân tử và nguyên nhân gây bệnh, mà còn có lợi cho việc chẩn đoán và điều trị lâm sàng. Nó mang lại cơ hội để phát triển phương pháp chẩn đoán và điều trị tốt hơn dựa trên miRNAs và làm tăng khả năng kiểm soát và xử lý các bệnh trên một mức độ phân tử.

Một số mô hình Deep Learning đã được phát triển để dự đoán mối quan hệ miRNAs – bệnh như RLSMDA (Regularized Least Squares for MiRNA-Disease Association), IMCMDA (Inductive Matrix Completion for MiRNA-Disease Association), RKNNMDA(Ranking-based KNN for MiRNA-Disease Association Prediction), VAEMDA (variational autoencoder for MiRNA–disease association) [6]

A diagram of a dna sequence

Description automatically generated with medium confidence

Hình ‑ Mô hình VAEMDA áp dụng cho bài toán toán dự đoán quan hệ giữa bệnh – mRNAs [6]

## Tổng kết chương

Như vậy, Deep Learning là một lĩnh vực quan trọng trong máy tính và trí tuệ nhân tạo đã đem lại nhiều đột phá trong nhiều lĩnh vực, trong đó có Sinh học. Deep Learning đã trải qua một sự phát triển đáng kể, đặc biệt là sau sự ra đời của các mạng nơ-ron sâu như Convolutional Neural Networks (CNNs) và Recurrent Neural Networks (RNNs). Cùng với đó, sự phát triển của phần cứng, sự gia tăng về dữ liệu, và các kỹ thuật mới đã tạo điều kiện thuận lợi cho Deep Learning.

Hiện tại, Deep Learning đã được ứng dụng rộng rãi trong nhiều bài toán sinh học, bao gồm chẩn đoán bệnh, phát hiện gen, tạo hình ảnh y học, và dự đoán quan hệ giữa bệnh và thuốc. Mặc dù Deep Learning có nhiều ứng dụng tiềm năng, nó cũng đối mặt với thách thức về việc thu thập dữ liệu lâm sàng và đảm bảo tính riêng tư của dữ liệu y tế. Tương lai của Deep Learning trong sinh học sẽ chứng kiến sự phát triển tiếp tục của các mô hình học sâu và phương pháp tiền xử lý dữ liệu phức tạp.

# MỘT MÔ HÌNH ỨNG DỤNG DEEP LEARNING DỰ ĐOÁN QUAN HỆ THUỐC – BỆNH

## Giới thiệu chung về bài toán

Việc phát triển các loại thuốc mới là một quá trình lâu dài với tốc độ chậm, tỷ lệ tiêu hao cao và chi phí đáng kể, thậm chí việc đưa chúng vào thị trường thành công cần nhiều nỗ lực hơn. Phần lớn các ứng cử viên thuốc bị loại bỏ trong quá trình thử nghiệm lâm sàng giai đoạn I. Do đó, việc khám phá những cách hiệu quả để cải thiện tỷ lệ thành công của nghiên cứu và phát hiện thuốc là rất cấp bách và có ý nghĩa quan trọng. Trong những năm gần đây, một số đề xuất đáng chú ý trong lĩnh vực phát triển thuốc và tái định vị thuốc đang thu hút sự quan tâm ngày càng tăng từ cả ngành dược phẩm và cộng đồng nghiên cứu.

Tái định vị thuốc, hoặc tái sử dụng thuốc, nhằm mục đích xác định các cơ hội điều trị mới cho các loại thuốc hiện có, đồng thời giảm thời gian, chi phí và rủi ro của việc phát triển thuốc thông thường. Cách đơn giản nhất để tìm ra các chỉ định mới cho các loại thuốc hiện có là thông qua các thí nghiệm sinh học để thực hiện sàng lọc dựa trên mục tiêu hoặc tế bào cho hàng nghìn loại thuốc. Mặc dù chiến lược dựa trên hoạt động này có thể trực tiếp phát hiện các khả năng tiềm ẩn đối với thuốc, nhưng đây vẫn là một quá trình tốn nhiều thời gian và công sức, đồng thời việc thử nghiệm thuốc trong các xét nghiệm dựa trên một số cơ sở dữ liệu hợp chất lâm sàng toàn diện có sẵn cũng vô cùng khó khăn. Tuy nhiên, nhờ những tiến bộ nhanh chóng trong các phương pháp học sâu đã mang đến cơ hội tuyệt vời để khai thác quá trình tái định vị thuốc bằng các phương pháp tính toán với quy trình tái định vị nhanh hơn nhiều với chi phí thấp hơn. Cho đến nay đã có nhiều phương pháp hiệu quả như DRIMC (Drug Repositioning by Integrating Multiple Data Sources via Compressive Matrix Completion), BNNR (bounded nuclear norm regularization), LAGCN (Layer Attention Graph Convolutional Network method for the drug–disease association prediction)...[7] được phát triển để dự đoán cho bài toán tái sử dụng thuốc, đặc biệt cho dự đoán mối quan hệ thuốc – bệnh. Có thể kể đến phương pháp DRWBNCF được thiết kế để dự đoán các tương tác giữa các thuốc và bệnh dựa trên dữ liệu đã biết, nhằm giúp trong việc tối ưu hóa quá trình phát hiện thuốc mới.

DRWBNCF (Weighted Bilinear Neural Collaborative Filtering), còn được gọi là phương pháp lọc cộng tác tuyến tính hai chiều có trọng số bằng mạng thần kinh, là một phương pháp mới trong lĩnh vực học máy và khai phá dữ liệu, đặc biệt liên quan đến việc giải quyết bài toán dự đoán tương tác giữa các cặp thuốc và bệnh. Phương pháp này tập trung vào việc tạo ra dự đoán về tương tác thuốc - bệnh thông qua việc xây dựng mạng lưới cộng tác neural bilinear với trọng số. Đầu tiên, phương pháp xây dựng ba mạng lưới riêng biệt: mạng lưới tương tác thuốc - bệnh đã biết, mạng lưới tương tự thuốc - thuốc và mạng lưới tương tự bệnh - bệnh. Các mạng lưới này được sử dụng để mô hình hóa thông tin về các hàng xóm gần kề và tương tác giữa chúng.

DRWBNCF giúp tạo ra dự đoán tương tác thuốc - bệnh chính xác hơn bằng cách sử dụng thông tin từ các hàng xóm gần kề và tương tác giữa chúng. Phương pháp này có thể ứng dụng trong việc cải thiện chất lượng chẩn đoán và điều trị trong lĩnh vực y học và dược học.

## Một mô hình ứng dụng Deep Learning trong bài toán nghiên cứu và ứng dụng Deep Learning trong dự đoán mối quan hệ thuốc – bệnh

### Mô hình của bài toán

Mô hình DRWBNCF được công bố trên tạp chí Briefings in Bioinformatics vào năm 2022. Mô hình này được phát triển bởi các nhà nghiên cứu thuộc Đại học Hồ Nam, Trung Quốc và Geneis Beijing Co. Ltd. Các nhà nghiên cứu được đề cập tới đó là Yajie Meng, Chang Cheng Lu, Min Jin, Junlin Xu, Xiangxiang Zeng và Jialiang Yang [7].

A diagram of a algorithm

Description automatically generated

Hình ‑ Cấu trúc của mô hình DRWBNCF [7]

### Đầu vào và đầu ra của bài toán

Trong đồ án này, tôi sử dụng ba bộ dữ liệu thực tế đã được sử dụng trong các nghiên cứu trước để đánh giá hiệu quả của DRWBNCF.

* Bộ dữ liệu Fdataset: Bộ dữ liệu chứa 1933 tương tác thuốc - bệnh đã biết giữa 593 loại thuốc được thu thập từ cơ sở dữ liệu DrugBank và 313 bệnh thu thập từ cơ sở dữ liệu OMIM.
* Bộ dữ liệu Cdataset: Bộ dữ liệu này chứa 2532 tương tác thuốc - bệnh đã biết giữa 663 loại thuốc được trích xuất từ cơ sở dữ liệu DrugBank và 409 bệnh được liệt kê trong cơ sở dữ liệu OMIM.
* Bộ dữ liệu LRSSL: Bộ dữ liệu này bao gồm 763 loại thuốc được thu thập từ cơ sở dữ liệu DrugBank, 681 bệnh thu thập từ cơ sở dữ liệu MeSH và tổng cộng 3051 tương tác thuốc - bệnh đã được xác nhận.

Kết quả đầu ra của phương pháp DRWBNCF là một tập hợp các xác suất dự đoán liên kết giữa các cặp thuốc-bệnh chưa được quan sát trước đó. Đối với mỗi cặp thuốc - bệnh, mô hình ước tính xác suất rằng cặp này có khả năng liên kết. Cụ thể của DRWBNCF là một ma trận, trong đó mỗi phần tử (i, j) đại diện cho xác suất mà cặp thuốc i và bệnh j có liên kết. Các giá trị của ma trận nằm trong khoảng từ 0 đến 1, thể hiện xác suất dự đoán của mô hình

Kết quả này có thể được sử dụng để xây dựng các đồ thị AUROC và AUPR, từ đó đánh giá hiệu suất dự đoán của phương pháp. Nó cung cấp thông tin về khả năng dự đoán của mô hình trong việc xác định các mẫu dương (các cặp thuốc-bệnh có khả năng liên kết) và mẫu âm (các cặp thuốc-bệnh không có khả năng liên kết).

### Các bước thực hiện

#### Cấu trúc của ba mạng dữ liệu

* Ta thiết lập ba mạng dữ liệu, bao gồm
  + Mạng liên kết thuốc – bênh đã biết từ kết quả của các nghiên cứu trước
  + Mạng tương đồng thuốc – thuốc đã biết từ kết quả của các nghiên cứu trước
  + Mạng tương đồng bệnh – bệnh đã biết từ kết quả của các nghiên cứu trước.
* Ta biểu diễn mạng liên kết thuốc – bệnh bởi ma trận A có kích thước n\*m, trong đó n là số lượng thuốc và m là số lượng bệnh. Nếu thuốc r được xác nhận có liên kết với bệnh d thì Aij = 1, ngược lại thì Aij = 0 (0 ≤ i ≤ n, 0 ≤ j ≤ m).
* Ví dụ: Giả sử chúng ta có các số liệu sau:
  + Số lượng thuốc n = 4
  + Số lượng bệnh m = 3
  + Ma trận A cho mạng liên kết thuốc - bệnh đã biết
* Mạng liên kết tương đồng thuốc – thuốc được biểu diễn bằng ma trận A­r có kích thước n\*n, trong đó n là số thuốc.
  + Mỗi phần tử của Ar được tính bằng mức độ tương tự giữa các cặp thuốc. Các độ tương tự này được biểu thị bởi ma trận Sr có kích thước n\*n, trong đó phần tử Sr (i, j) tượng trưng cho độ tương tự giữa thuốc ri và thuốc rj.
  + Để đảm bảo kết quả chính xác hơn bằng cách tránh thông tin nhiễu, mô hình sẽ chỉ sử dụng k - láng giềng gần nhất (k – nearest) thay vì tất cả láng giềng tương tự như trong các nghiên cứu trước đó. Phần tử Ar (i, j) của ma trận Ar được định nghĩa như sau:

Ar (i, j) = (1)

Trong đó: Nk (ri)là tập hợp gồm k thuốc có mức độ tương tự cao nhất với thuốc r­i. Như vậy

Ví dụ: Giả sử chúng ta có các số liệu sau:

* + Số lượng thuốc n = 4
  + Ma trận tương đồng thuốc - thuốc (Sr)

Áp dụng công thức (1) với k = 2 ta sẽ được ma trận Ar như sau:

* Tương tư, mạng tương đồng bệnh – bệnh được biểu diễn bằng ma trận A­d có kích thước m\*m, trong đó m là số bệnh.
  + Mỗi phần tử của Ad được tính bằng mức độ tương tự giữa các cặp bệnh. Các độ tương tự này được biểu thị bởi ma trận Sd có kích thước m\*m, trong đó phần tử Sd (i, j) tượng trưng cho độ tương tự giữa bệnh di và bệnh dj.
  + Tương tự như mạng tương đồng thuốc – thuốc, để đảm bảo kết quả chính xác hơn bằng cách tránh thông tin nhiễu, mô hình sẽ chỉ sử dụng k - láng giềng gần nhất (k – nearest) thay vì tất cả láng giềng tương tự như trong các nghiên cứu trước đó. Phần tử Ad (i, j) của ma trận Ad được định nghĩa như sau:

Ad (i, j) = (2)

Trong đó: Nk (di)là tập hợp gồm k thuốc có mức độ tương tự cao nhất với bệnh d­i. Như vậy

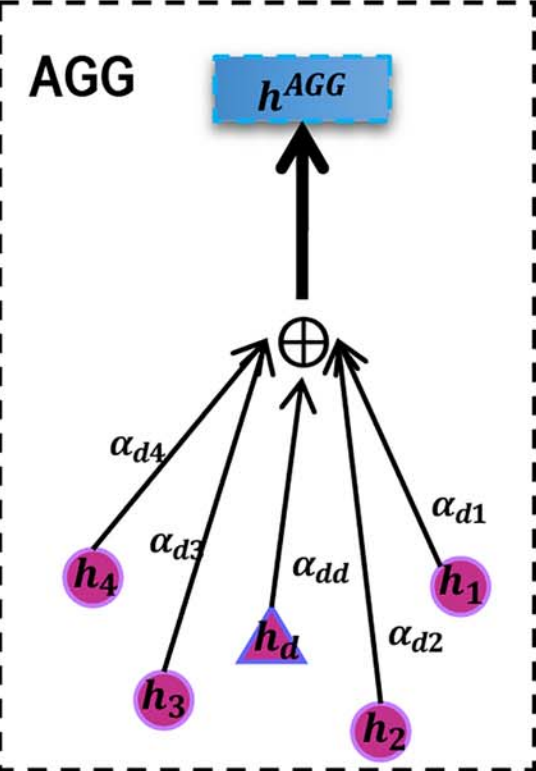
Ví dụ: Giả sử chúng ta có các số liệu sau:

* + Số lượng bệnh (m) = 3
  + Ma trận tương đồng bệnh – bệnh (Sd)

Áp dụng công thức (2) với k = 1 ta sẽ được ma trận Ad như sau:

#### Cấu trúc của mô hình

* Bài toán tập trung vào việc dự đoán mức độ liên kết giữa các cặp thuốc và bệnh (drug-disease association). Mục tiêu là xác định liệu một cặp thuốc và bệnh có liên kết với nhau hay không.
* Các thuốc được biểu diễn bằng tập hợp R, còn các bệnh được biểu diễn bằng tập hợp D. Một cặp thuốc – bệnh <r, d> được gọi là cặp liên kết khi Yr,d = 1, ngược lại Yr,d = 0.
* *Phép tích chập đồ thị thông thường*
  + Mạng nơ-ron đồ thị (Graph Neurons Network) đã định nghĩa phép tích chập đồ thị thông thường như một tổng có trọng số (tức là tổng tuyến tính) của các đặc trưng của các hàng xóm của nút mục tiêu để tạo ra biểu diễn của nút mục tiêu.



Hình ‑ Tiền xử lý dữ liệu của DRWBNCF [7]

* + Để mã hóa thông tin hàng xóm của thuốc r và bệnh d bằng phép tích chập đồ thị thông thường, các biểu diễn của thuốc r và bệnh d được định nghĩa như sau:

(3)

(4)

Trong đó:

* + - i là thuốc (bệnh) thuộc tập mở rộng có mức độ tương tự cao nhất tương ứng của thuốc r (bệnh d).
    - là trọng số của hàng xóm i, được tính bằng , với là số lượng thuốc (bệnh) có liên kết tới thuốc r (bệnh d) và chính thuốc r (bệnh d) đó
    - Wr, Wd là ma trận trọng số, Wr có kích thước |R| \* k2, Wd có kích thước |D| \* k3
* *Mã hoá mối liên hệ đã biết giữa thuốc và bệnh*
  + Trong mạng liên kết thuốc - bệnh, ta đã biết về sự tương quan giữa các cặp thuốc và bệnh, có sẵn trong ma trận nhị phân A kích thước n \* m. Nếu thuốc ri liên quan đến bệnh dj thì Aij = 1, ngược lại thì Aij = 0.
  + Để biểu diễn thông tin về mối quan hệ này, các đặc trưng gốc của thuốc và bệnh được biểu diễn bằng ma trận Sr và Sd. Đặc trưng của thuốc ri được mã hóa trong vector sr với kích thước 1\*|R|, và đặc trưng của bệnh dj được mã hóa trong vector sd có kích thước R1\*|D|
  + Bằng cách áp dụng hai ma trận chiếu P có kích thước |R|\*K3 và Q có kích thước |D|×K3, đặc trưng mã hóa sr và sd sẽ được chuyển thành các vector ẩn pr và qd:

pr = sr · P, qd = sd · Q (5)

Ở đây, pr và qd là các vector có kích thước 1 \* K3 cho thuốc và bệnh tương ứng.

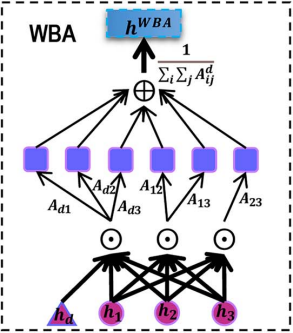
* + Ta kết hợp 2 ma trận thông qua phép nhân các thành phần để tạo ra biểu diễn ẩn cho mối quan hệ giữa thuốc và bệnh cụ thể. Hàm này được biểu thị bằng biểu thức sau:

hr,d = (pr, qd) = pr qd (6)

*Mã hoá mối liên kết lân cận*

* + Để tìm hiểu các đặc điểm đại diện của thuốc r toàn diện hơn, ta xem xét tất cả các tương tác có trọng số theo cặp giữa k-nearest trong tập xác định tập hợp song tuyến tính có trọng số (Weighted bilinear aggregator):

hr = WBA () = (7)



Hình ‑ Tiền xử lý dữ liệu của DRWBNCF [7]

* + Công thức (7) bao gồm các bước như sau
    - Tính tổng trọng số tương tác giữa tất cả các cặp hàng xóm (i, j) trong tập r), với i ≠ j, được thể hiện bởi .
    - Duyệt qua tất cả các cặp hàng xóm i và j, và tính giá trị tương tác giữa i và j dựa trên ma trận trọng số Ur có kích thước (được xem xét như ma trận trọng số tương tác) và các vector ẩn ở công thức (5) biểu diễn của i và j.
    - Tính tổng các giá trị tương tác được tính toán ở trên, và nhân tổng này với tương ứng để có thành phần của vector biểu diễn hr.
  + Tương tự đối với bệnh ta có công thức (8)

hd = WBA () = (8)

* + Để tăng cường việc biểu diễn các đặc tính của thuốc r và bệnh d, ta áp dụng phương pháp tuyến tính sơ đồ nhị phân bao gồm đồ thị thông thường toán tử tích chập (tức là tổng có trọng số) và WBA, chính thức xác định một phép tích chập đồ thị mới như sau:

(9)

(10)

Trong đó

* + - λ là tham số điều chỉnh, đo lường sự ảnh hưởng của WBA so với AGG.
    - σ(·) là hàm kích hoạt phi tuyến, được áp dụng sau khi tính toán tổng.
    - có kích thước 1 \* K1 và có kích thước 1 \* K2 là vector thông tin về hàng xóm và tương tác hàng xóm của loại thuốc r và bệnh d.
  + Áp dụng các công thức (3), (4), (7), (8), công thức (9), (10) được thay đổi như sau

(11)

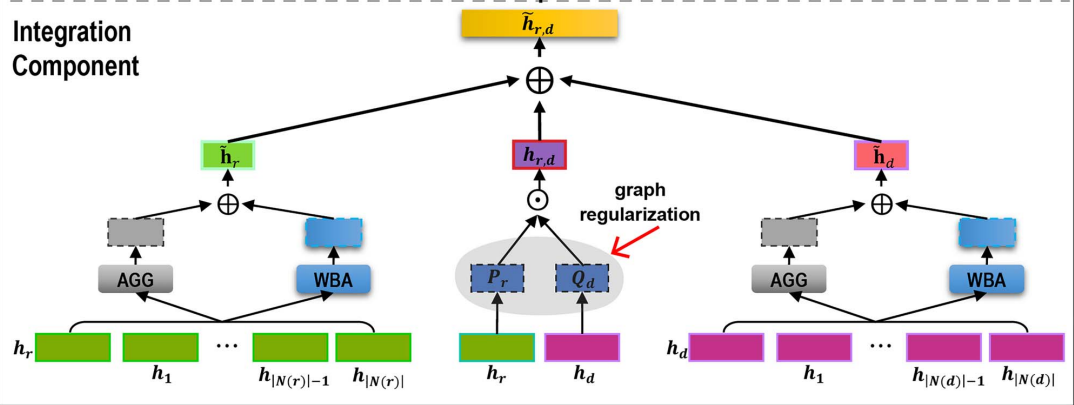
(12)

*Tích hợp thông tin về liên kết đã biết với thông tin về hàng xóm và tương tác có trọng số giữa các hàng xóm*

* + Sau khi ta đã có biểu diễn cho liên kết biết đến giữa loại thuốc và bệnh, cũng như biểu diễn cho hàng xóm và tương tác hàng xóm của cả loại thuốc và bệnh, chúng ta tiếp tục kết hợp những vector này thành một biểu diễn thống nhất như sau:

= hr,d ⊕ ⊕ (13)

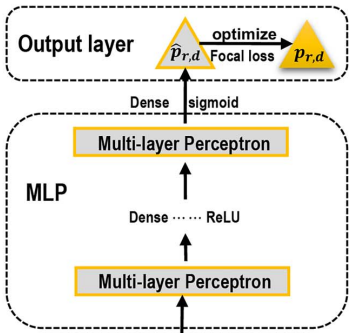
Trong đó ⊕ biểu thị phép nối vector



Hình ‑ Dữ liệu đầu vào của mạng nơ-ron thuộc mô hình DRWBNCF [7]

*Mạng nơ-ron để dự đoán (MLP-based prediction)*

* + Ta có thể thấy mối quan hệ giữa thuốc và bệnh có thể rất phức tạp. Để giải quyết vấn đề này, ta xây dựng mạng nơ – ron nhân tạo để dự đoán mối quan hệ thuốc – bệnh.
  + Mạng nơ – ron được xây dựng từng lớp một. Với mỗi lớp, kết quả đầu vào của lớp trước đó được kết hợp với trọng số tương ứng và độ lệch (bias) để tính toán kết quả đầu ra của lớp hiện tại. Sau khi xử lý qua tất cả các lớp, kết quả cuối cùng được đưa qua hàm kích hoạt sigmoid để đảm bảo rằng kết quả dự đoán nằm trong khoảng từ 0 đến 1, có ý nghĩa như xác suất dự đoán rằng mối quan hệ giữa loại thuốc và bệnh là 1 (tức là có liên kết) dựa trên thông tin đã xử lý



Hình ‑ Sơ đồ mạng nơ-ron của DRWBNCF [7]

Mạng nơ – ron nhân tạo được trong DRWBNCF xây dựng như sau:

,

,

,

Trong đó:

* + - f (x) (·) là hàm kích hoạt phi tuyến cho lớp thứ x. Ta ReLU làm hàm kích hoạt để tránh tình trạng quá bão hòa.
    - Wx và bx biểu thị ma trận trọng số và vector bias tương ứng.
    - Ta thiết lập z0 = (được định nghĩa ở công thức (13)).

*Tối ưu hoá*

* + Ta xem xét các cặp thuốc - bệnh là các mẫu dương (positive samples) là Y+ và các cặp khác (tức là các cặp thuốc - bệnh chưa được quan sát hoặc không biết) là các mẫu âm (negative samples) Y-
  + Các cặp thuốc - bệnh đã được xác nhận là có liên kết có tính đáng tin cậy cao và quan trọng để cải thiện hiệu suất dự đoán. Tuy nhiên, số lượng các mẫu dương thường ít hơn nhiều so với số lượng các mẫu âm. Tình trạng mất cân đối này gây ra hai vấn đề:
    - Quá trình huấn luyện không hiệu quả vì các mẫu âm dễ dàng phân loại chiếm đa số tổng lượng lỗi mà không đóng góp thông tin nào hữu ích
    - Các mẫu âm quá nhiều khiến quá trình huấn luyện khó khăn và dẫn đến hiệu suất kém của các mô hình.
  + Để giải quyết tính trạng mất cân đối trên, ta sẽ sử dụng đến hệ số cân bằng α cho lớp Y+ và 1 – α cho lớp Y-, với α
  + Ta áp dụng hàm mất mát lấy cân bằng α (α-balanced focal loss) và các thành phần chính quy hóa đồ thị vào hàm mục tiêu của mô hình DRWBNCF:

(15)

* + Hàm mục tiêu của DRWBNCF cân bằng một số mục tiêu:
    - Phần đầu tiên (tức hàm mất mát lấy cân bằng α) tập trung huấn luyện vào tập dày đặc của các mẫu dương, từ đó giảm trọng lượng mất mát gán cho số lượng lớn các mẫu âm.
    - Phần thứ hai là để chính quy hóa đồ thị thuốc, giảm thiểu khoảng cách giữa các vector đặc trưng tiềm ẩn của hai thuốc lân cận.
    - Phần thứ ba là để chính quy hóa đồ thị bệnh.

# THỰC NGHIỆM VÀ KẾT QUẢ

## Cài đặt và thực nghiệm

### Môi trường cài đặt

Để phục vụ cho việc huấn luyện mô hình DRWBNCF, tôi đã chọn Python làm ngôn ngữ cũng như môi trường để cài đặt. Phiên bản Python chính thức mà tôi sử dụng là 3.6.12, một phiên bản ổn định và tương thích với nhiều thư viện phổ biến. Cùng với đó, tôi sử dụng Pycharm làm IDE (Integrated Development Environment) làm môi trường cho việc huấn luyện.

Để giúp tôi trong việc phát triển và huấn luyện mô hình máy học, tôi đã tích hợp PyTorch phiên bản 1.6.0 vào môi trường của mình. PyTorch là một thư viện mạnh mẽ cho học máy và deep learning, cho phép chúng tôi xây dựng và huấn luyện các mô hình phức tạp một cách dễ dàng và hiệu quả.

Tôi cũng sử dụng PyTorch Lightning phiên bản 1.0.8 để giảm bớt sự phức tạp của quá trình huấn luyện và đảm bảo tính tái sử dụng của mã nguồn. PyTorch Lightning cung cấp một cách tiếp cận mô hình học máy dựa trên PyTorch với cấu trúc rõ ràng và tối ưu hóa, giúp tôi tập trung vào việc phát triển mô hình mà không cần lo lắng về việc quản lý quy trình huấn luyện.

Cuối cùng, tôi sử dụng thư viện scikit-learn để thực hiện các tác vụ liên quan đến học máy và xử lý dữ liệu. Scikit-learn là một thư viện mạnh mẽ và đáng tin cậy cho học máy và phân tích dữ liệu, cung cấp các công cụ và thuật toán phổ biến cho nhiều tác vụ khác nhau như phân loại, hồi quy, gom cụm, và xử lý dữ liệu.

### Tham số của mô hình DRWBNCF

Các tham số đóng vai trò quan trọng trong việc đưa ra kết quả, ảnh hưởng đến hiệu suất của mô hình. Một số tham số quan trọng trong mô hình:

* λ (được nêu ở công thức 11, 12) ảnh hưởng đến quá trình tiền xử lý đầu vào dữ liệu của mạng nơ ron
* k (được nêu ở công thức 11, 12) là số hàng xóm gần nhất của thuốc (bệnh), ảnh hưởng đến quá trình tiền xử lý đầu vào dữ liệu của mạng nơ-ron
* α (được nêu ở công thức 15) ảnh hưởng đến quá trình tối ưu hoá mô hình
* Dropout rate: là giá trị phần trăm các đơn vị (node) bị bỏ qua trong mạng neural network trong quá trình huấn luyện mô hình học máy. Điều này sẽ giúp mô hình chống hiện tượng over-fitting. [11]
* Learning rate: tốc độ học tập của mô hình, ảnh hưởng trực tiếp đến việc giảm thiểu sự mất mát giữa thực tế và dự đoán từ các mẫu dữ liệu đầu vào. [11]
* Epoch: là số lần toàn bộ tập dữ liệu đầu vào được vào mạng nơ-ron để huấn luyện. [11]
* Hidden units: số lớp ẩn của mạng nơ-ron, lớp này nằm ở giữa lớp đầu ra và đầu vào. [11]

### Hướng dẫn chạy chương trình

* Tôi cài đặt PyCharm Edu 2022 làm IDE chính để chạy chương trình cũng như huấn luyện model.

A screenshot of a computer

Description automatically generated

Hình ‑ PyCharm Edu 2022.2.2

* Sau khi đã cài đặt PyCharm, tôi tiếp tục cài đặt thêm Anaconda. Anaconda là một bản phân phối Python và môi trường phát triển, cung cấp công cụ và thư viện đã được tối ưu hoá. Anaconda được sử dụng rộng rãi trong lĩnh vực khoa học dữ liệu, học máy và phân tích dữ liệu hiện nay.

A screenshot of a computer

Description automatically generated

Hình ‑ Anaconda Navigator – quản lý môi trường ảo

* Mở project đã được cung cấp mã nguồn trên Github của tác giả bằng Pycharm, tôi cài đặt thêm một số thư viện như Pytorch 1.6.0, Pytorch-Lightning 1.0.8, Scikit-learn phục vụ cho việc huấn luyện mô hình.

A screenshot of a computer

Description automatically generated

Hình ‑ Mã nguồn của mô hình DRWBNCF

* Trong mã nguồn này, thư mục dataset là nơi để tập dữ liệu đầu vào gồm 3 bộ dữ liệu Fdataset, Cdataset và LRSSL. Thư mục src là nơi lưu trữ các mã nguồn liên quan đến việc tiền xử lý dữ liệu đầu vào cũng như xây dựng mạng nơ-ron của mô hình. Cuối cùng là các file main.py, demo.py, demo.sh là các file python được sử dụng để cài đặt một số thông số cũng như để chạy mô hình. Tôi sẽ chạy file demo.sh để chạy mô hình.

A screenshot of a computer

Description automatically generated

Hình ‑ Sử dụng file demo.sh để chạy mô hình

### Thực nghiệm

Tôi đã xét giá trị của các tham số liên quan của phần tiền xử lý và phần mạng nơ-ron để thực hiện huấn luyện mô hình.

A computer code with many colored text

Description automatically generated with medium confidence

Hình ‑ Mã nguồn các tham số

A computer screen shot of a program

Description automatically generated

Hình ‑ Mã nguồn các tham số

## Kết quả thu được và so sánh với một số phương pháp khác

### Kết quả thu được

Để đánh giá hiệu suất của mô hình DRWBNCF, ta sử dụng phương pháp chia dữ liệu kiểm tra và huấn luyện theo kiểu 10 - fold cross-validation (10 - CV) để dự đoán liên kết loại thuốc - bệnh. Cụ thể, các liên kết loại thuốc - bệnh đã xác nhận và các liên kết chưa quan sát được chia ngẫu nhiên thành 10 phần bằng nhau. Mỗi phần gồm các liên kết đã xác nhận và các liên kết chưa quan sát được xem như các mẫu kiểm tra lần lượt, trong khi 9 phần còn lại được sử dụng làm mẫu huấn luyện để huấn luyện mô hình được đề xuất. Các liên kết đã xác nhận được coi là các mẫu dương và các liên kết chưa quan sát được coi là các mẫu âm. [7]

Sau khi tất cả các liên kết đã được kiểm tra, chúng ta tính tỷ lệ True-Positive (TPR), tỷ lệ False-Positive (FPR) và Precision như sau:

TPR (hoặc Recall) = (16)

Trong đó TP là số lượng mẫu dương được xác định đúng, và FN biểu thị số lượng mẫu âm được xác định sai.

FPR = (17)

Precision = (18)

Trong đó FP là số lượng mẫu âm được xác định sai, và TN biểu thị số lượng mẫu âm được xác định đúng.

Bằng cách thay đổi ngưỡng xếp hạng, ta có thể tính toán TPR, FPR và Precision để xây dựng đường cong đặc trưng (ROC) và đường cong Precision-recall (PR). Đối với đường cong ROC, FPR được đặt trên trục x và TPR được đặt trên trục y. Đối với đường cong PR, Recall và Precision được đặt trên các trục x và y tương ứng. Diện tích dưới đường cong ROC (AUROC) và diện tích dưới đường cong precision–recall (AUPR) được sử dụng để đánh giá hiệu suất tổng thể của DRWBNCF. Để có kết quả thuyết phục, ta lặp lại 10 lần phương pháp 10-fold CV và báo cáo giá trị trung bình là kết quả cuối cùng.

Sau khi tiến hành thực nghiệm trên Fdataset, ta thu được các biểu đồ AUROC và AUPR sau mỗi lần huấn luyện và kiểm tra chéo như sau:

#### AUROC

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ nhất

A graph with blue and orange lines

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ hai

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ ba

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ tư

A graph with blue and orange lines

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ năm

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ sáu

A graph with blue and orange lines

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ bảy

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ tám

A graph with blue and orange lines

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ chín

A graph with blue and orange lines

Description automatically generated

Hình ‑ Biểu đồ AUROC cho lần huấn luyện thứ mười

#### AUPR

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ nhất

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ hai

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ ba

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ tư

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ năm

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ sáu

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ bảy

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ tám

A graph with a line

Description automatically generated

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ chín

A graph of a line

Description automatically generated with medium confidence

Hình ‑ Biểu đồ AUPR cho lần huấn luyện thứ mười

### So sánh với một số phương pháp khác

Ta có các phương pháp khác đã được công bố trước đây cho vấn đề về xác định mối quan hệ thuốc – bệnh như sau:

* BNNR (bounded nuclear norm regularization) là một phương pháp trong việc hoàn thành ma trận drug-disease (ma trận mối quan hệ giữa các loại thuốc và bệnh tật) dưới giả thiết về chiều thấp (low-rank assumption).[7]
* LAGCN (Layer Attention Graph Convolutional Network) trong việc dự đoán mối quan hệ giữa các loại thuốc và bệnh tật. LAGCN sử dụng một mạng đồ thị (graph) để học các biểu diễn của các loại thuốc và bệnh tật từ dữ liệu đa dạng, bao gồm các mối quan hệ drug-disease đã biết, độ tương tự giữa các loại thuốc và độ tương tự giữa các bệnh tật.[7]

Với các kết quả được công bố từ trước của các phương pháp trên, kết hợp các kết quả thực nghiệm, ta có bảng so sánh kết quả như sau

A screenshot of a graph

Description automatically generated

Bảng 3. Kết quả sau khi huấn luyện mô hình DRWBNCF trên 3 tập dữ liệu so sánh với kết quả của các phương pháp khác đã được công bố trước đó [7]

* Chỉ số AUROC của DRWBNCF lớn hơn hầu hết các phương pháp khác, chứng tỏ mô hình này hiệu quả hơn các mô hình đã được công bố trước đó
* Ta thấy BNNR cũng có chỉ số AUROC tương tự như DRWBNCF, ta sẽ xét đến chỉ số AUPR cho trường hợp này

A screenshot of a graph

Description automatically generated

Bảng 3. Kết quả sau khi huấn luyện mô hình DRWBNCF trên 3 tập dữ liệu so sánh với kết quả của các phương pháp khác đã được công bố trước đó [7]

* DRWBNCF có AUPR trung bình tốt nhất là 0.4688, cao hơn 5.7% so với BNNR. Điều này chứng tỏ rằng khả năng dự đoán các cặp thuốc - bệnh chưa được biết đến từ tập dữ liệu ban đầu tốt hơn.

### Tổng kết chương

Như vậy, sau khi sử dụng phương pháp huấn luyện theo kiểu 10 - fold cross-validation (10 - CV) để dự đoán liên kết loại thuốc - bệnh trên bộ dữ liệu đã có, ta đã thu được kết quả đầu ra của bài toán. Kết quả đó đã được xây dựng thành các đồ thị AUROC và AUPR, từ đó đánh giá hiệu suất dự đoán của phương pháp. Ngoài ra, bằng cách so sánh với các dữ liệu của các phương pháp giải quyết bài toán dự đoán mối liên kết thuốc – bệnh khác như BNNR, LAGCN đã cho ta thấy DRWBNCF vượt trội hơn so với các phương pháp khác về AUPR cũng như AUC trên tất cả ba bộ dữ liệu. Tóm lại, DRWBNCF là một phương pháp tiềm năng cho việc tái định vị thuốc và có thể được áp dụng cho các lĩnh vực khác của phát hiện và phát triển thuốc.

# KẾT LUẬN

Trong đồ án này, tôi đã tiến hành nghiên cứu sâu vào lĩnh vực Deep Learning và ứng dụng nó để giải quyết bài toán dự đoán quan hệ thuốc-bệnh. Quan hệ này có ảnh hưởng lớn đến lĩnh vực y tế và có tiềm năng cải thiện quản lý bệnh và điều trị.

Tôi đã bắt đầu bằng việc nghiên cứu các khía cạnh cơ bản của Deep Learning, bao gồm khái niệm cơ bản, đặc điểm và ứng dụng của Deep Learning. Tiếp đó, tôi cũng nghiên cứu về mạng nơ-ron của Deep Learning và một số mạng nơ-ron tiêu biểu, để hiểu cách chúng hoạt động và cách chúng có thể áp dụng vào bài toán của tôi. Sau đó, tôi tìm hiểu về các đại lượng đánh giá cho mô hình nhằm hiểu được một mô hình hoạt động hiệu quả được dựa trên tiêu chí nào.

Để giải quyết bài toán dự đoán mối liên kết thuốc – bệnh, tôi đã tìm hiểu một số khái niệm sinh học như RNA, thuốc, bệnh, và một số tương tác giữa chúng. Tiếp theo, tôi đã tìm hiểu các thuật toán Deep Learning đã được ứng dụng như thế nào trong việc dự đoán mối liên kết thuốc – bệnh. Cuối cùng, tôi lựa chọn mô hình DRWBNCF để tìm hiểu sâu hơn, thu thập dữ liệu y tế và tiến hành tiền xử lý để chuẩn bị cho quá trình đào tạo và kiểm tra mô hình.

Trong quá trình thử nghiệm và đào tạo, tôi đã tìm hiểu và đánh giá mô hình DRWBNCF dưới nhiều dữ liệu đầu vào khác nhau và với các thiết lập tham số khác nhau. Kết quả thử nghiệm đã cho thấy rằng mô hình Deep Learning có khả năng dự đoán quan hệ thuốc-bệnh với độ chính xác đáng kể, giúp nâng cao hiệu suất và hiệu quả trong việc quản lý thông tin y tế và điều trị bệnh.

Tuy nhiên, nghiên cứu này cũng đã tiếp cận một số thách thức, như việc thu thập và tiền xử lý dữ liệu y tế phức tạp, cũng như việc điều chỉnh các siêu tham số của mô hình để đạt được hiệu suất tốt nhất. Các nghiên cứu và phát triển tương lai có thể tiếp tục nghiên cứu các phương pháp và kiến thức mới để cải thiện khả năng dự đoán và áp dụng rộng rãi trong lĩnh vực y tế.

# TÀI LIỆU THAM KHẢO

1. Michahial Kaaya Tech, S., Kumar, H., Michahial, S., Kumar, H. P. & Ahmed, F. Feature Extraction and Classification of EEG Signal Using Neural Network Based Techniques Breast cancer detection View project Feature Extraction and Classification of EEG Signal Using Neural Network Based Techniques. *Certified International Journal of Engineering and Innovative Technology (IJEIT* **9001**, (2008).

2. Detection and Tracking of Pallets using a Laser Rangefinder and Machine Learning Techniques. (2017) doi:10.13140/RG.2.2.30795.69926.

3. Shajihan, N. Classification of stages of Diabetic Retinopathy using Deep Learning Classification of stages of Diabetic Retinopathy using Deep Learning View project Classification of stages of Diabetic Retinopathy using Deep Learning. doi:10.13140/RG.2.2.10503.62883.

4. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J Cheminform* **12**, 35 (2020).

5. Silva, A. B. O. V. & Spinosa, E. J. Graph Convolutional Auto-Encoders for Predicting Novel lncRNA-Disease Associations. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 2264–2271 (2022).

6. Zhang, L., Chen, X. & Yin, J. Prediction of potential miRNA–disease associations through a novel unsupervised deep learning framework with variational autoencoder. *Cells* **8**, (2019).

7. Meng, Y. *et al.* A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform* **23**, (2022).

8. Cấu trúc và Chức năng của các ARN. https://thpt-lequydon-quangtri.edu.vn/cau-truc-va-chuc-nang-arn.

9. micro RNA | What is microRNA (miRNA)? | How miRNAs work? | How miRNAs are detected experimentally? https://www.youtube.com/watch?v=hY3TYEo24PY.

10. What is overfitting? https://aws.amazon.com/vi/what-is/overfitting/.

11. JOHN D. KELLEHER. *DEEP LEARNING*. (2019).