**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

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

ĐỒ ÁN ĐẠI HỌC

NGÀNH HỆ THỐNG THÔNG TIN

Đề tài : NGHIÊN CỨU VỀ DEEP LEARNING VÀ ỨNG DỤNG GIẢI BÀI TOÁN DỰ ĐOÁN QUAN HỆ THUỐC- BỆNH (DRUG-DISEASE)

|  |  |
| --- | --- |
| GVHD | **: 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 (Drug-disease)”. 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

[LỜI CẢM ƠN 2](#_Toc143574100)

[MỞ ĐẦU 3](#_Toc143574101)

[MỤC LỤC 5](#_Toc143574102)

[TỔNG QUAN VỀ DEEP LEARNING VÀ ỨNG DỤNG TRONG CÁC BÀI TOÁN SINH HỌC 1](#_Toc143574103)

[1.1. Tổng quan về Deep Learning 1](#_Toc143574104)

[1.1.1. Giới thiệu về Deep Learning 1](#_Toc143574105)

[1.1.2. Đặc điểm của Deep Learning 1](#_Toc143574106)

[1.1.3. Các mô hình Deep Learning tiêu biểu 3](#_Toc143574107)

[1.1.4. Một số ứng dụng của Deep Learning 7](#_Toc143574108)

[1.1.5. Một số đại lượng đánh giá mô hình 9](#_Toc143574109)

[1.2. Ứng dụng của Deep Learning trong các bài toán sinh học 16](#_Toc143574110)

[1.2.1. Một số đối tượng sinh học đáng quan tâm 16](#_Toc143574111)

[1.2.2. Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh và các đối tượng sinh học 22](#_Toc143574112)

[1.3. Kết luận 22](#_Toc143574113)

[CHƯƠNG 2: MỘT MÔ HÌNH ỨNG DỤNG DEEP LEARNING DỰ ĐOÁN QUAN HỆ THUỐC – BỆNH 22](#_Toc143574114)

[2.1. Giới thiệu chung về bài toán 23](#_Toc143574115)

[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 24](#_Toc143574116)

[2.2.1. Đầu vào của bài toán 24](#_Toc143574117)

[2.2.2. Các bước thực hiện 24](#_Toc143574118)

[2.2.3. Đầu ra của bài toán 24](#_Toc143574119)

[CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ 24](#_Toc143574120)

[3.1. Hướng dẫn chi tiết cài đặt 24](#_Toc143574121)

[3.2. Kết quả đánh giá 24](#_Toc143574122)

[3.3. So sánh với một số phương pháp khác 24](#_Toc143574123)

[TÀI LIỆU THAM KHẢO 24](#_Toc143574124)

# 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.).
* Các mô hình deep learning có thể được huấn luyện bằng cách sử dụng các thuật toán như Backpropagation và Stochastic Gradient Descent.
* 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. 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 (Graphics Processing Unit) và TPU (Tensor Processing Unit), 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.
* 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.



Hỉnh ảnh

* 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. CNNs sử dụng các lớp Convolutional để trích xuất các đặc trưng của hình ảnh và các lớp Pooling để giảm kích thước của đầu ra.
* 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.
* Transfer Learning: Transfer Learning là một kỹ thuật được sử dụng để sử dụng lại các mô hình CNNs đã được huấn luyện trước đó để giải quyết các tác vụ khác nhau. Kỹ thuật này giúp tiết kiệm thời gian và tài nguyên huấn luyện mô hình mới từ đầu.



* 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. RNNs sử dụng các lớp Recurrent để lưu trữ thông tin về các phần tử trước đó trong dãy dữ liệu.
  + Các lớp Recurrent: Các lớp Recurrent trong RNNs được sử dụng để lưu trữ thông tin về các phần tử trước đó trong dãy dữ liệu. Các lớp Recurrent thường sử dụng các hàm kích hoạt như Tanh hoặc ReLU để tính toán giá trị của các hidden state.
* Các kiến trúc RNNs: Có nhiều kiến trúc RNNs khác nhau được sử dụng trong Deep Learning, bao gồm Simple RNNs, LSTM và GRU. Mỗi kiến trúc có những ưu điểm và hạn chế riêng.
  + 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 Machine Learning, 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, cần lưu ý rằng 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 hoặc có mất cân bằng. Trong các trường hợp này, accuracy có thể bị nhiễu bởi tỷ lệ dự đoán đúng của lớp thiểu số.

#### 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.
* Confusion matrix 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).



Hình ảnh

* + 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.

#### F1-Score

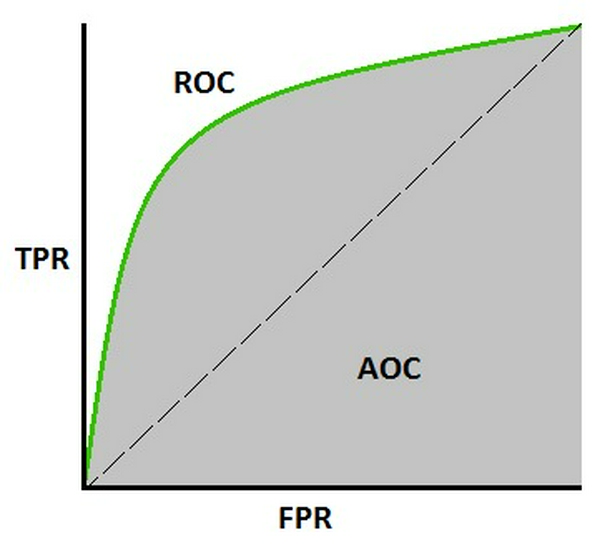
* Một mô hình tốt khi cả Precision và Recall đều cao, thể hiện cho mô hình ít phân loại nhầm giữa các lớp cũng như tỉ lệ bỏ sót các đối tượng thuộc lớp cần quan tâm là thấp. Tuy nhiên, hai giá trị Precision và Recall thường không cân bằng với nhau (giá trị này tăng thì giá trị kia thường có xu hướng giảm). Để đánh giá cùng lúc cả Precision và Recall, ta sử dụng độ đo F-Score.



* Tham số β quyết định mức độ coi trọng giữa Precision và Recall
  + β > 1: Recall được coi trọng hơn Precision
  + β < 1: Precision được coi trọng hơn Recall
  + β = 1: Precision và Recall được coi trọng ngang nhau
* Việc quyết định nên ưu tiên Precision hay Recall phụ thuộc vào từng bài toán. Với những bài toán mà Precision và Recall được cân nhắc ngang nhau, ta chọn β = 1, khi đó ta đang sử dụng F1-Score. F1-Score là kỳ vọng harmonic (harmonic mean) của Precision và Recall.
* F1-score lớn khi cả 2 giá trị Precision và Recall đều lớn. Ngược lại, chỉ cần 1 giá trị nhỏ sẽ làm cho F1-Score nhỏ.

#### 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.
* Ý nghĩa của AUROC có thể diễn giải như sau: Là xác suất rằng một mẫu dương tính được lấy ngẫu nhiên sẽ được xếp hạng cao hơn một mẫu âm tính được lấy ngẫu nhiên. Biểu diễn theo công thức, ta có AUC = P(score(x+) > score(x-)). Chỉ số AUC càng cao thì mô hình càng chính xác trong việc phân loại các lớp.
* Đường cong ROC biểu diễn các cặp chỉ số (TPR, FPR) tại mỗi ngưỡng với TPR là trục tục và FPR là trục hoành.



Hình ảnh

* 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, được tính theo công thức:

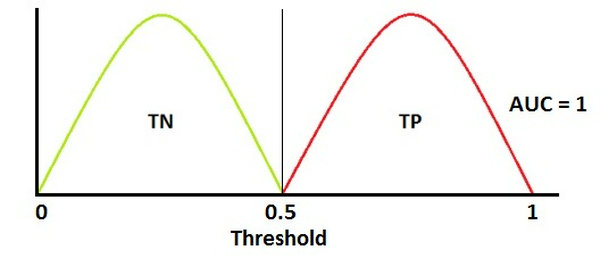
TPR càng cao thì các mẫu dương tính càng được phân loại chính xác.

* 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, được tính theo công thức:

FPR càng cao thì Specificity càng giảm và số lượng các mẫu âm tính bị gắn nhãn sai càng lớn.

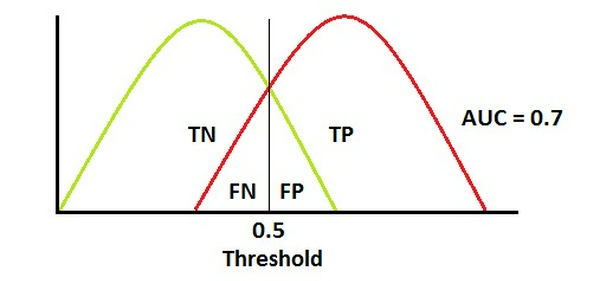
* Để hợp chúng lại thành 1 chỉ số duy nhất, ta sử dụng đường cong ROC để hiển thị từng cặp (TPR, FPR) cho các ngưỡng khác nhau với mỗi điểm trên đường cong biểu diễn 1 cặp (TPR, FPR) cho 1 ngưỡng, sau đó tính chỉ số AUC cho đường cong này. Chỉ số AUC chính là con số thể hiện hiệu suất phân loại của mô hình.
* 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).
  + AUC = 1



Hình ảnh

Đây là trường hợp tốt nhất. Mô hình phân loại hoàn toàn chính xác khi 2 đường cong không chồng lên nhau.

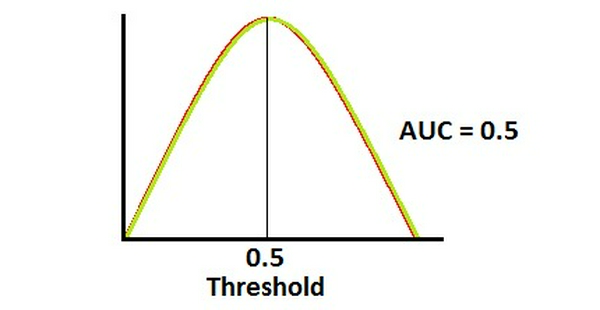
* + AUC = 0.7



Hình ảnh

Khi 2 đường cong chồng lên nhau, việc phân loại sẽ xảy ra 2 dạng lỗi đó là FP(Type 1 Error) và FN(Type 2 Error). Ta có thể thay đổi giá trị của 2 chỉ số lỗi này bằng cách thay đổi ngưỡng. Có thể thấy đường cong ROC đã hạ xuống một chút, tuy nhiên nó vẫn nằm ở góc trên bên trái của đồ thị, tức là hiệu suất phân loại vẫn ổn định.

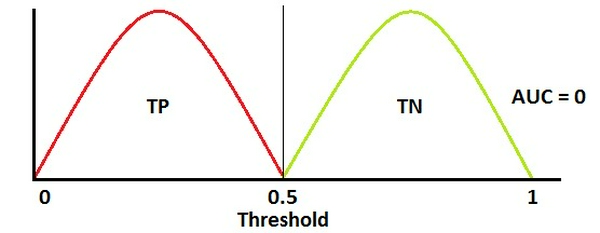
* + AUC = 0.5



Hình ảnh

Đây là trường hợp tệ nhất. Mô hình hoàn toàn không có khả năng phân loại giữa 2 lớp.

* + AUC = 0



Hình ảnh

Khi AUC xâp xỉ 0, mô hình phân loại ngược hoàn toàn 2 lớp với việc phân loại âm tính thành dương tính - dương tính thành âm tính. Để sửa điều này ta chỉ cần đảo ngược đầu ra của mô hình.

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

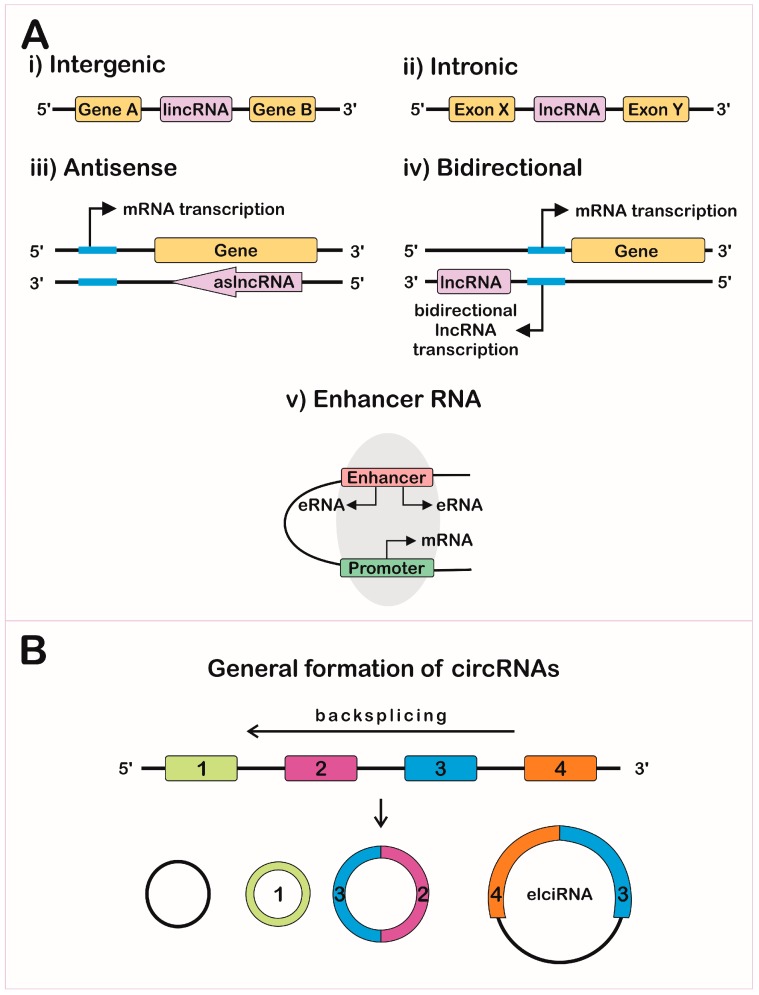
### Một số đối tượng sinh học đáng quan tâm

#### 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.
* RNA có cấu trúc tương tự như DNA, với một chuỗi các nucleotide liên kết với nhau bởi các liên kết đường phènyl (phosphodiester bonds). Các nucleotide trong RNA bao gồm adenine (A), cytosine (C), guanine (G) và uracil (U), khác với thymine (T) trong DNA.



* IncRNA (long non-coding RNA) là một dạng của RNA, được định nghĩa là các RNA không mã hóa protein khác với các RNA như tRNA, rRNA và snRNA, và độc lập với các RNA nhỏ có máy móc xử lý phân tử cụ thể như micro- hoặc piwi-RNA.
* Các nghiên cứu về lncRNA trên các loài khác nhau đã phát hiện ra một quần thể đa dạng các phân tử RNA khác nhau về kích thước và chức năng. Các lncRNA có thể được phân loại dựa trên kích thước (nhỏ hơn hoặc lớn hơn 200 cặp bazơ - thường được sử dụng là định nghĩa của lncRNA dài so với RNA ngắn), vị trí (ví dụ: các loại RNA được tạo ra từ 3 'UTR hoặc 5' UTR), tương tác phân tử (ví dụ: phụ thuộc vào Drosha hoặc Dicer), và chức năng phân tử.



Hình ảnh

* 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. Các nghiên cứu gần đây cũng cho thấy rằng lncRNA 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.
* Trong bối cảnh này, 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. Các nghiên cứu gần đây cũng đã chỉ ra rằng lncRNA có thể được sử dụng như một chỉ số tiên đoán cho các bệnh lý khác nhau, và có thể được sử dụng để phát triển các phương pháp chẩn đoán và điều trị mới.
* Tuy nhiên, việc hiểu rõ hơn về cấu trúc và chức năng của các lncRNA vẫn còn rất nhiều thách thức và cần thêm nhiều nghiên cứu để giải quyết. Một trong những thách thức lớn nhất đối với việc nghiên cứu lncRNA là khả năng dự đoán chức năng của chúng dựa trên cấu trúc của phân tử. Do đó, các nghiên cứu tiếp theo cần tập trung vào việc xác định các tương tác phân tử của lncRNA và các protein hoặc RNA khác, và phân tích các tác động của chúng đến các quá trình sinh học. Ngoài ra, việc phát hiện và xác định các lncRNA mới cũng là một lĩnh vực nghiên cứu tiềm năng, đặc biệt là trong việc tìm kiếm các chỉ số tiên đoán cho các bệnh lý khác nhau.
* Tóm lại, lncRNA là một lĩnh vực nghiên cứu đầy triển vọng và cần thêm nhiều nghiên cứu để hiểu rõ hơn về chúng và ứng dụng của chúng trong các lĩnh vực khác nhau của sinh học và y học.

#### Micro RNAs

* MiRNAs được phát hiện lần đầu vào những năm 1990, và nhanh chóng thu hút sự quan tâm của cộng đồng nghiên cứu sinh học phân tử. Chúng là một loại RNA nhỏ, có độ dài khoảng 21-25 nucleotide. Cấu trúc của miRNA bao gồm một dãy nucleotide đáng kể, gọi là "seed region," có vai trò quan trọng trong việc nhận dạng và tương tác với các mục tiêu gen.



Hình ảnh

* MiRNAs tham gia vào quá trình điều chỉnh biểu hiện gen ở mức post-transcriptional. Chúng hoạt động bằng cách tương tác với các messenger RNA (mRNA), ngăn cản quá trình dịch mã gen hoặc thúc đẩy phân hủy mRNA. MiRNAs được sản xuất thông qua một chuỗi các bước tiền lưu và xử lý tại cả nhân (nucleus) và cytoplasm.
* 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.
* Tóm lại, microRNAs (miRNAs) là một loại phân tử RNA nhỏ có vai trò quan trọng trong điều chỉnh sự biểu hiện gen và quá trình sinh học ở cả thực vật và động vật. Sự hiểu biết về miRNAs đã và đang góp phần quan trọng vào sự tiến bộ của nghiên cứu sinh học phân tử và y học.

#### Diseases

* Disease (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, ...
* Các cơ chế gây bênh bao gồm:
  + Bệnh cấp tính (Acute Diseases): Là các bệnh phát triển nhanh chóng và có thể kéo dài trong một thời gian ngắn. Ví dụ: cảm lạnh, viêm họng cấp, ...
  + Bệnh mãn tính (Chronic Diseases): Là các bệnh kéo dài trong thời gian dài và thường liên quan đến sự suy yếu chức năng dần dần. Ví dụ: bệnh tim mạch mãn tính, tiểu đường type 2, ...
  + Bệnh di truyền (Hereditary Diseases): Là các bệnh được truyền từ thế hệ này sang thế hệ khác thông qua gene. Ví dụ: bệnh bạch cầu, hội chứng Marfan, ...
* Việc phát hiện ra bệnh sẽ được dựa trên triệu chứng và biểu hiện
  + Triệu chứng là các dấu hiệu hoặc tình trạng mà người bệnh cảm nhận, chẳng hạn như đau, mệt mỏi, hoặc khó thở.
  + Biểu hiện là các tín hiệu mà người y tế có thể nhận thấy trong quá trình khám và kiểm tra, ví dụ như kết quả xét nghiệm máu, tăng huyết áp.
* Dựa vào các triệu chứng và biểu hiện, ta đưa ra các chẩn đoán và lộ trình điều trị
  + Chẩn đoán bệnh thường liên quan đến việc thu thập thông tin về triệu chứng, tiến hành các xét nghiệm hình ảnh và xét nghiệm thử nghiệm, và phân tích dữ liệu để xác định nguyên nhân và loại bệnh.
  + Điều trị bệnh có thể bao gồm sử dụng thuốc, phẫu thuật, liệu pháp vật lý, thay đổi lối sống, và các biện pháp hỗ trợ khác như tâm lý học và chăm sóc y tế.
* 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.
* Tóm lại, disease là một khái niệm quan trọng trong lĩnh vực y học, thể hiện sự bất thường trong cơ thể dẫn đến sự suy yếu chức năng và gây ra các triệu chứng không mong muốn. Nó có nhiều nguyên nhân và biểu hiện khác nhau, cần được chẩn đoán và điều trị một cách khoa học và hiệu quả.

#### Drugs

* Drugs (thuốc) trong ngữ cảnh y học và khoa học là 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 drugs 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

#### Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh – micro RNAs

#### Ứng dụng của Deep Learning trong giải bài toán dự đoán quan hệ giữa bệnh – thuốc

## Kết luận

# 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 đề xuất hấp dẫn trong lĩnh vực phát triển thuốc, tái định vị thuốc, 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. May mắn thay, 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.
* 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.
  + DRWBNCF 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

### Đầu vào của bài toán

* Ta 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 này tương ứng với bộ dữ liệu thử nghiệm chất lượng vàng về tương tác thuốc - bệnh được báo cáo trong công trình của Gottlieb et al. . Nó 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.
* Ta tính toán độ tương tự giữa các cặp thuốc bằng cách dựa trên cấu trúc hóa học của chuỗi đầu vào đơn giản (SMILES), và đại diện bằng chỉ số Tanimoto của các dấu vân tay hóa học của cặp thuốc thông qua Chemical Development Kit. Độ tương tự giữa các cặp bệnh được đo lường dựa trên độ tương tự ngữ nghĩa của các hiện tượng bệnh thông qua phân tích khai thác thông tin mô tả y học của cặp bệnh.

### 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)

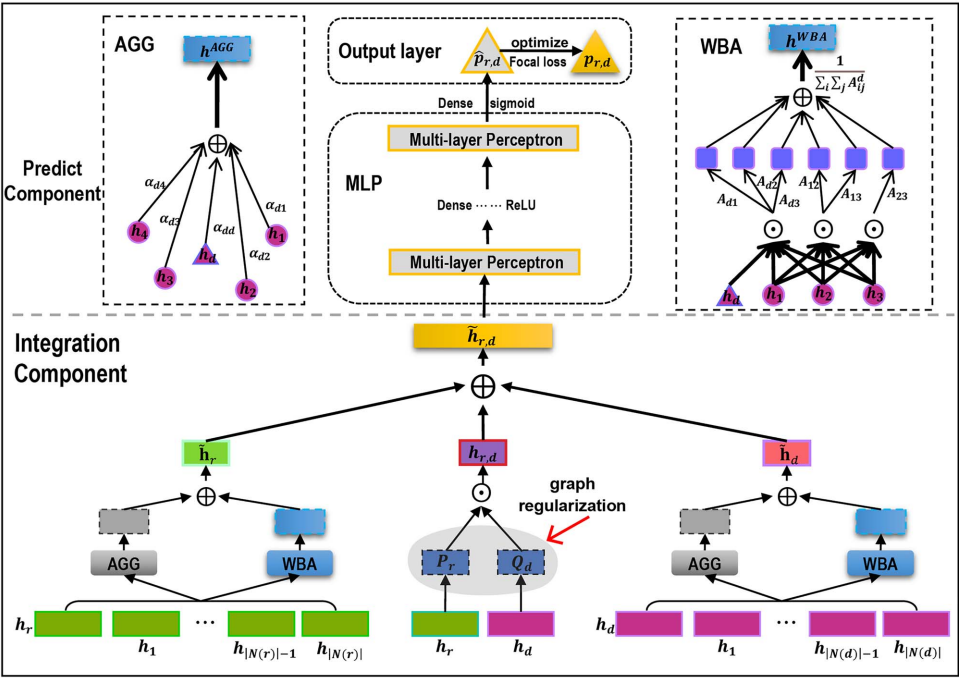
Trong đó: Nk (di)là tập hợp gồm k thuốc có mức độ tương tự cao nhất với bệnh d­i. Như vậy

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

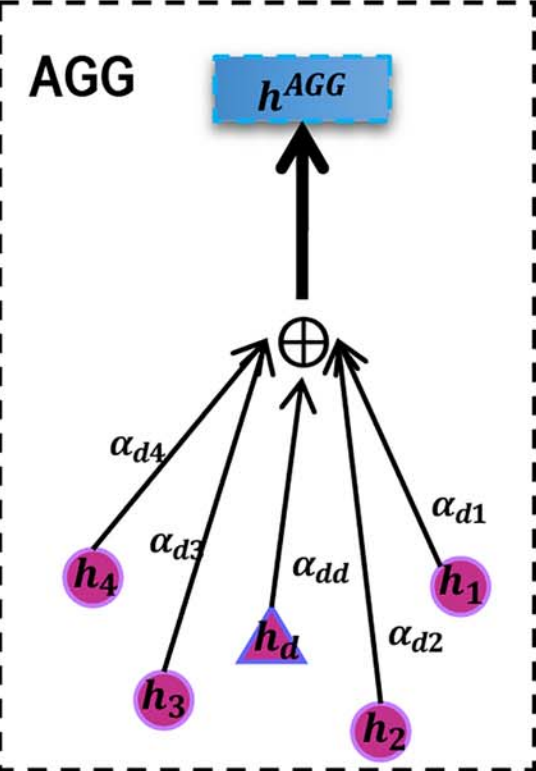
* + Số lượng bệnh (m) = 3
  + Ma trận tương đồng bệnh – bệnh (Sd)

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.



* + 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)

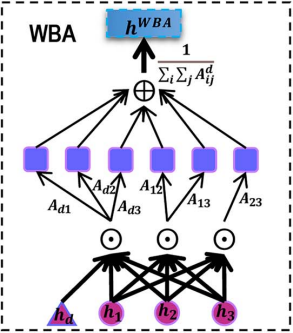
Ở đây, pr và qd là các vector có kích thước 1 \* K3 cho thuốc và bệnh tương ứng.

* + 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)



* + 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

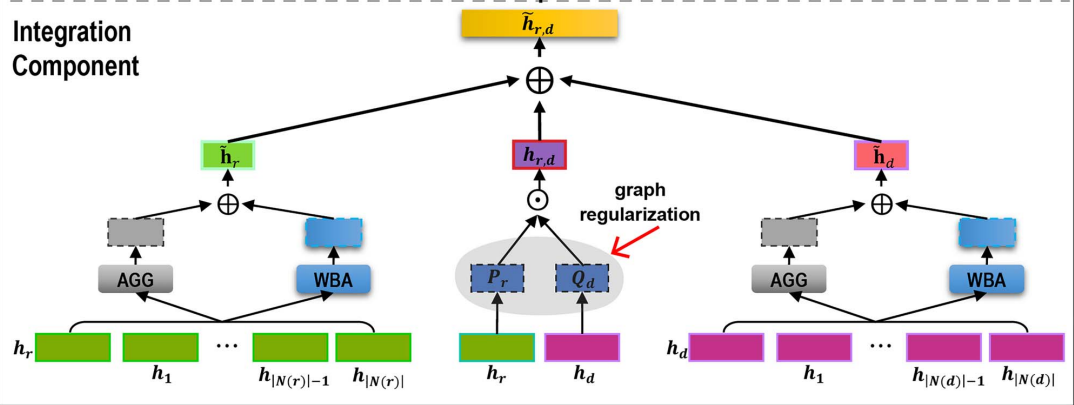
(11)

(12)

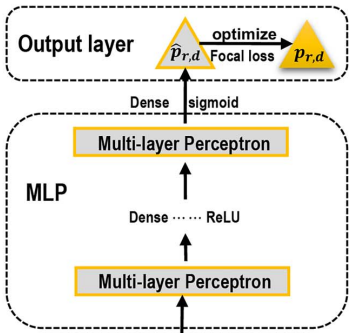
* *Tích hợp thông tin về liên kết đã biết với thông tin về hàng xóm và tương tác có trọng số giữa các hàng xóm*
  + Sau khi ta đã có biểu diễn cho liên kết biết đến giữa loại thuốc và bệnh, cũng như biểu diễn cho hàng xóm và tương tác hàng xóm của cả loại thuốc và bệnh, chúng ta tiếp tục kết hợp những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

= hr,d ⊕ ⊕ (13)

Trong đó ⊕ biểu thị phép nối vector



* *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



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
* Đánh giá mô hình
  + 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. Đáng chú ý, 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.
  + 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, chúng ta lặp lại 10 lần phương pháp 10-fold CV và báo cáo giá trị trung bình là kết quả cuối cùng.

### Đầu ra của bài toán

* 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 ROC và PR, 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).

# THỰC NGHIỆM VÀ KẾT QUẢ

## Hướng dẫn chi tiết cài đặt

## Kết quả đánh giá

## So sánh với một số phương pháp khác

# TÀI LIỆU THAM KHẢO