HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN 1

BÁO CÁO THỰC TẬP

HỖ TRỢ CHẨN ĐOÁN BỆNH NGHỀ NGHIỆP SỬ DỤNG MÔ HÌNH HỌC SÂU ĐA PHƯƠNG THỨC

Học sinh: Nguyễn Đức Hải - B21DCDT085

Lớp: E21CNPM01

Giáo viên hướng dẫn: Nguyễn Trọng Khánh

Ngành: Công nghệ thông tin

Hà Nội – 2025

LỜI CẢM ƠN

Sau quá trình thực tập tốt nghiệp tại Học viện Công nghệ Bưu chính Viễn thông, em xin được bày tỏ lòng biết ơn sâu sắc đến tất cả những người đã tạo điều kiện, hỗ trợ và đồng hành cùng em trong suốt thời gian vừa qua.

Trước tiên, em xin chân thành cảm ơn **Ban Giám hiệu Học viện Công nghệ Bưu chính Viễn thông** đã tạo điều kiện thuận lợi và môi trường học tập, nghiên cứu tốt nhất cho sinh viên. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến **Khoa Công nghệ Thông tin** với đội ngũ lãnh đạo và giảng viên đã luôn quan tâm, tạo mọi điều kiện để em có thể hoàn thành tốt nhiệm vụ thực tập.

Em xin bày tỏ lòng biết ơn chân thành nhất đến **Thầy Nguyễn Trọng Khánh**, người thầy đã dành nhiều thời gian quý báu để hướng dẫn, chỉ bảo em trong suốt quá trình thực tập. Với kiến thức chuyên môn sâu rộng, kinh nghiệm nghiên cứu phong phú và tâm huyết trong công tác đào tạo, thầy đã giúp em định hướng đúng đắn cho đề tài nghiên cứu, giải đáp mọi thắc mắc và luôn động viên em vượt qua những khó khăn trong quá trình làm việc. Những kiến thức, kinh nghiệm và phương pháp nghiên cứu khoa học mà thầy truyền đạt sẽ là hành trang quý báu cho con đường sự nghiệp của em.

Em cũng xin gửi lời cảm ơn đến **cô Phượng** đã nhiệt tình chia sẻ kinh nghiệm, kiến thức chuyên môn và tạo điều kiện cho em tiếp cận với các công nghệ tiên tiến trong lĩnh vực trí tuệ nhân tạo và xử lý dữ liệu y tế. Đặc biệt cảm ơn các thầy cô đã hỗ trợ em trong việc tiếp cận các tài liệu nghiên cứu, cơ sở dữ liệu và môi trường cần thiết cho việc thực hiện đề tài.

Mặc dù đã có nhiều cố gắng, song do thời gian và kinh nghiệm còn hạn chế, báo cáo này chắc chắn còn nhiều thiếu sót. Em rất mong nhận được những ý kiến đóng góp, chỉ bảo của các thầy cô và các bạn để có thể hoàn thiện hơn nội dung nghiên cứu cũng như nâng cao kiến thức chuyên môn của bản thân.

NHẬN XÉT THỰC TẬP

(của giáo viên hướng dẫn)

……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………
…………………………………………

Giáo viên hướng dẫn ký, ghi rõ họ tên

Ghi chú:

-
-
-

THÔNG TIN SINH VIÊN, CÔNG TY / ĐƠN VỊ THỰC TẬP, CÁN BỘ HƯỚNG DẪN, GIẢNG VIÊN, TRƯỜNG VÀ KHOA

Thông tin sinh viên

Họ và tên sinh viên: Nguyễn Đức Hải ……

Mã số sinh viên: B21DCDT085…….

Ngành học: Công Nghệ Thông Tin…….

Sinh viên năm thứ: 4……

Địa chỉ tạm trú trong thời gian thực tập: Thôn Bầu– Kim Chung – Hà Nội…….

Số điện thoại: 0971772586 ……E-mail: nguyenhai6586@gmail.com

Thông tin công ty / đơn vị thực tập

Công ty / Đơn vị thực tập: HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG ……

Địa chỉ: Km10 Nguyễn Trãi, Hà Đông, Hà Nội

…….

Số điện thoại: 024 3756 2186…. Số fax:……

Cán bộ hướng dẫn tại đơn vị thực tập: Nguyễn Trọng Khánh……

Chức vụ (vị trí): …….

Số điện thoại: ……. E-mail: khanhnt@ptit.edu.vn…

Giảng viên phối hợp hướng dẫn: Nguyễn Trọng Khánh…….

Số điện thoại: ……. E-mail: khanhnt@ptit.edu.vn……

Thời gian thực tập: 31/06/2025……

Bắt đầu: 31/06/2025…….

Kết thúc: 24/08/2025……...

Thông tin Học viện Công nghệ Bưu chính Viễn thông

Địa chỉ: Km10 Nguyễn Trãi, Hà Đông, Hà Nội

Số điện thoại: 024 3756 2186

E-mail: ctsv@ptit.edu.vn Website: www.ptit.edu.vn

Triết lý giáo dục: Tri thức – Sáng tạo – Đạo đức – Trách nhiệm.

Tầm nhìn: Đến năm 2030, Học viện Công nghệ Bưu chính Viễn thông là trường đại học hàng đầu Việt Nam về quy mô, chất lượng đào tạo, nghiên cứu khoa học; là hình mẫu tiên phong về chuyển đổi số trong giáo dục đại học, trở thành trường đại học hàng đầu của khu vực, nằm trong nhóm trường đại học hàng đầu châu Á, nhóm 5 trường đại học hàng đầu Đông Nam Á về công nghệ số.

Sứ mạng: Sáng tạo và chuyển giao tri thức cho xã hội thông qua việc gắn kết các hoạt động đào tạo nguồn nhân lực chất lượng cao, nghiên cứu khoa học và chuyển giao công nghệ trong lĩnh vực thông tin, truyền thông và công nghệ số, góp phần xây dựng đất nước Việt Nam hùng cường.

Giá trị cốt lõi: Tiên phong – Sáng tạo; Chất lượng – Hiệu quả; Uy tín – Trách nhiệm; Tận tụy – Nghĩa tình.

Thông tin Khoa Công nghệ Thông tin 1

Địa chỉ: Tầng 9 Nhà A2, Học viện Công nghệ Bưu chính Viễn thông, Km 10 Nguyễn Trãi, Hà Đông, Hà Nội

Số điện thoại: 024 3854 5604

E-mail: phuongnd@ptit.edu.vn Website: www.it.ptit.edu.vn

Giới thiệu

Khoa Công nghệ thông tin 1 là đơn vị đào tạo trực thuộc Học viện Công nghệ Bưu chính Viễn thông. Khoa có chức năng đào tạo và nghiên cứu khoa học thuộc lĩnh vực Công nghệ thông tin (Công nghệ thông tin, Công nghệ phần mềm, Hệ thống thông tin, Khoa học máy tính, Kỹ thuật máy tính).

PHẦN I: GIỚI THIỆU VỀ ĐƠN VỊ THỰC TẬP

1. TỔNG QUAN VỀ HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG (PTIT)

1.1. Lịch sử hình thành và phát triển

Học viện Công nghệ Bưu chính Viễn thông (Posts and Telecommunications Institute of Technology - PTIT) là một trong những cơ sở đào tạo hàng đầu trong lĩnh vực công nghệ thông tin và viễn thông tại Việt Nam. Được thành lập với sứ mệnh đào tạo nguồn nhân lực chất lượng cao phục vụ cho sự phát triển của ngành bưu chính viễn thông và công nghệ thông tin trong nước.

Học viện Công nghệ Bưu chính Viễn thông, tiền thân là Trường Bưu điện – Vô tuyến điện (1953) được thành lập theo quyết định số 516/TTg của Thủ tướng Chính phủ ngày 11 tháng 7 năm 1997 trên cơ sở sắp xếp lại 4 đơn vị nghiên cứu, đào tạo thành viên thuộc Tổng Công ty Bưu chính Viễn thông Việt Nam, nay là Tập đoàn Bưu chính Viễn thông Việt Nam VNPT. Từ ngày 1/7/2014, theo Quyết định số 878/QĐ-BTTTT của Bộ trưởng Bộ Thông tin và Truyền thông, Học viện Công nghệ Bưu chính Viễn thông được điều chuyển từ Tập đoàn Bưu chính Viễn thông Việt Nam về trực thuộc Bộ Thông tin và Truyền thông (nay là Bộ Khoa học và Công nghệ).

Ngày 27/02/2025, Chính Phủ đã ban hành Quyết định số: 452/QĐ-TTg về việc Phê duyệt quy hoạch mạng lưới cơ sở giáo dục đại học và sư phạm thời kỳ 2021-2030, tầm nhìn đến năm 2050. Theo đó, Học viện Công nghệ Bưu chính Viễn thông là 1 trong 5 cơ sở giáo dục đại học được định hướng phát triển trở thành cơ sở giáo dục đại học trọng điểm Quốc gia về kỹ thuật, công nghệ có chất lượng và uy tín ngang tầm khu vực.

Trải qua nhiều năm xây dựng và phát triển, PTIT đã khẳng định được vị thế của mình trong hệ thống giáo dục đại học Việt Nam, đặc biệt trong lĩnh vực đào tạo công nghệ thông tin, viễn thông, điện tử và các ngành kỹ thuật liên quan.

1.2. Thông tin cơ bản

Tên đầy đủ: Học viện Công nghệ Bưu chính Viễn thông
Tên tiếng Anh: Posts and Telecommunications Institute of Technology (PTIT)

Loại hình trường: Công lập
Cơ quan quản lý: Bộ Thông tin và Truyền thông

Địa chỉ trụ sở quản lý: 122 Hoàng Quốc Việt, Cầu Giấy, Hà Nội
Địa chỉ cơ sở đào tạo 1: Km10, Đường Nguyễn Trãi, Hà Đông, Hà Nội

1.3. Tầm nhìn và sứ mệnh

Tầm nhìn: Trở thành một trường đại học công nghệ hàng đầu khu vực về đào tạo, nghiên cứu khoa học và chuyển giao công nghệ trong lĩnh vực công nghệ thông tin và viễn thông.

Sứ mệnh: Đào tạo nguồn nhân lực chất lượng cao, thực hiện nghiên cứu khoa học và chuyển giao công nghệ, góp phần thúc đẩy sự phát triển kinh tế - xã hội của đất nước, đặc biệt trong lĩnh vực công nghệ thông tin, viễn thông và các lĩnh vực liên quan.

2. CƠ CẤU TỔ CHỨC VÀ HOẠT ĐỘNG

2.1. Cơ cấu tổ chức

Học viện được tổ chức theo mô hình đại học công nghệ với các khoa, viện chuyên ngành:

- Khoa Công nghệ thông tin
- Khoa Điện tử Viễn thông
- Khoa Quản trị Kinh doanh
- Khoa Ngoại ngữ
- Các viện nghiên cứu chuyên ngành

2.2. Lĩnh vực đào tạo

PTIT hiện đang đào tạo các ngành học từ bậc đại học đến sau đại học, tập trung vào các lĩnh vực:

Công nghệ thông tin:

- Công nghệ phần mềm
- Khoa học máy tính
- Hệ thống thông tin

- An toàn thông tin
- Trí tuệ nhân tạo

Điện tử - Viễn thông:

- Kỹ thuật điện tử truyền thông
- Công nghệ kỹ thuật điện, điện tử
- Kỹ thuật máy tính

Kinh tế - Quản lý:

- Quản trị kinh doanh
- Kế toán
- Tài chính - Ngân hàng

3. CƠ SỞ VẬT CHẤT VÀ TRANG THIẾT BỊ

3.1. Cơ sở vật chất

Cơ sở đào tạo tại Km10, Đường Nguyễn Trãi, Hà Đông được xây dựng trên diện tích rộng lớn với kiến trúc hiện đại, bao gồm:

- Các toà nhà giảng dường với hệ thống phòng học được trang bị đầy đủ
- Thư viện hiện đại với kho tài liệu phong phú
- Các phòng thí nghiệm chuyên ngành
- Khu vực sinh hoạt và dịch vụ cho sinh viên

3.2. Trang thiết bị và công nghệ

- Hệ thống phòng máy tính với cấu hình cao, kết nối internet tốc độ cao
- Các phòng thí nghiệm chuyên ngành được trang bị thiết bị hiện đại
- Phần mềm chuyên ngành được cập nhật thường xuyên
- Hệ thống mạng nội bộ và wifi phủ sóng toàn trường

4. ĐỘI NGŨ GIẢNG VIÊN VÀ HOẠT ĐỘNG NGHIÊN CỨU

4.1. Đội ngũ giảng viên

PTIT có đội ngũ giảng viên giàu kinh nghiệm với trình độ chuyên môn cao:

- Các giáo sư, phó giáo sư đầu ngành

- Tiến sĩ và thạc sĩ tốt nghiệp từ các trường đại học danh tiếng trong và ngoài nước
- Chuyên gia có kinh nghiệm thực tiễn từ doanh nghiệp

4.2. Hoạt động nghiên cứu

- Thực hiện các đề tài nghiên cứu khoa học cấp nhà nước, bộ, ngành
- Hợp tác nghiên cứu với các trường đại học và viện nghiên cứu trong nước và quốc tế
- Chuyển giao công nghệ cho doanh nghiệp
- Tổ chức các hội thảo khoa học quốc tế

5. HOẠT ĐỘNG ĐÀO TẠO VÀ LIÊN KẾT DOANH NGHIỆP

5.1. Chương trình đào tạo

- Chương trình đào tạo theo tiêu chuẩn quốc tế, cập nhật theo xu hướng công nghệ
- Liên kết với các trường đại học nước ngoài trong đào tạo
- Chú trọng thực hành, thực tập tại doanh nghiệp
- Phát triển kỹ năng mềm và năng lực nghề nghiệp cho sinh viên

5.2. Liên kết với doanh nghiệp

- Hợp tác với các tập đoàn công nghệ lớn trong nước và quốc tế
- Tổ chức các chương trình thực tập, việc làm cho sinh viên
- Phát triển các dự án hợp tác nghiên cứu và chuyển giao công nghệ
- Mời chuyên gia doanh nghiệp tham gia giảng dạy

6. THÀNH TỰU VÀ ĐỊNH HƯỚNG PHÁT TRIỂN

6.1. Những thành tựu đạt được

- Đào tạo được hàng nghìn kỹ sư, cử nhân chất lượng cao
- Sinh viên đạt nhiều giải thưởng trong các cuộc thi khoa học kỹ thuật
- Tỷ lệ sinh viên có việc làm sau tốt nghiệp cao
- Được công nhận chất lượng đào tạo bởi các tổ chức quốc tế

6.2. Định hướng phát triển

- Nâng cao chất lượng đào tạo theo hướng hiện đại, đáp ứng nhu cầu thị trường
- Mở rộng hợp tác quốc tế trong đào tạo và nghiên cứu
- Phát triển nghiên cứu ứng dụng và chuyển giao công nghệ
- Xây dựng môi trường học tập và làm việc chuyên nghiệp, sáng tạo
- Trờ thành hình mẫu về chuyển đổi số đại học của Việt Nam

PHẦN II: NỘI DUNG THỰC TẬP

1. Đặt vấn đề: Thực trạng và Nhu cầu

Bệnh bụi phổi silic là một trong những bệnh nghề nghiệp nguy hiểm và phổ biến nhất trên toàn cầu, đặc biệt tại các quốc gia đang phát triển với các ngành công nghiệp khai khoáng, xây dựng và sản xuất vật liệu. Bệnh gây ra do việc hít phải bụi silic tinh thể trong thời gian dài, dẫn đến tình trạng xơ hóa phổi không thể phục hồi, suy giảm chức năng hô hấp và có thể gây tử vong. Tại Việt Nam, công tác khám sàng lọc sức khỏe cho người lao động được thực hiện định kỳ, tuy nhiên, quy trình này đang đối mặt với nhiều thách thức lớn.

Thực trạng hiện tại cho thấy quy trình chẩn đoán sớm phụ thuộc nặng nề vào đội ngũ y bác sĩ chuyên khoa có kinh nghiệm, trong khi số lượng chuyên gia còn hạn chế so với số lượng lớn người lao động cần được thăm khám. Việc phân tích hàng nghìn bộ hồ sơ bệnh án và phim chụp X-quang trong thời gian ngắn dẫn đến nguy cơ quá tải, sai sót và làm chậm quá trình phát hiện bệnh. Dữ liệu y tế thu thập được thường không đồng nhất, thiếu hụt thông tin ở nhiều trường hợp, gây khó khăn cho việc áp dụng các phương pháp thống kê và học máy truyền thống.

Từ thực trạng trên, **nhu cầu cấp thiết** đặt ra là cần có một hệ thống thông minh có khả năng tự động hóa một phần quy trình sàng lọc, hỗ trợ các y bác sĩ trong việc nhận diện sớm các trường hợp có nguy cơ cao. Một hệ thống như vậy không chỉ giúp giảm tải áp lực cho đội ngũ y tế mà còn có tiềm năng nâng cao độ chính xác, rút ngắn thời gian chẩn đoán và giảm chi phí, từ đó cải thiện hiệu quả của công tác chăm sóc sức khỏe nghề nghiệp.

2. Nhiệm vụ được giao

Nhận thức được tiềm năng của AI trong việc giải quyết các thách thức trên, trong khuôn khổ kỳ thực tập, em đã được giao nhiệm vụ nghiên cứu và phát triển một mô hình học sâu tiên tiến để hỗ trợ chẩn đoán bệnh bụi phổi silic dựa theo dataset Sillicosis và bài báo https://thesai.org/Downloads/Volume15No7/Paper_128-Graph_Convolutional_Network_for_Occupational_Disease_Prediction.pdf. Các công việc cụ thể được phân công bao gồm:

Nghiên cứu lý thuyết và các công trình liên quan:

- **Tìm hiểu sâu về Mạng Nơ-ron Đồ thị (GNN):** Tập trung nghiên cứu các bài báo khoa học, đặc biệt là các công trình ứng dụng Mạng Tích chập Đồ thị (Graph Convolutional Network - GCN) trong chẩn đoán y khoa. Mục tiêu là nắm vững nguyên lý hoạt động, ưu điểm của GNN trong việc xử lý dữ liệu y tế không đồng nhất.

Cải tiến và Tích hợp Dữ liệu Đa phương thức:

- **Phân tích và xử lý bộ dữ liệu mới:** Tiếp nhận và phân tích bộ dữ liệu thực tế về bệnh bụi phổi silic, bao gồm cả dữ liệu bệnh án dạng bảng (tuần tự) và một tập hợp con dữ liệu hình ảnh X-quang phổi.
- **Cải tiến mô hình GCN:** Dựa trên các kiến thức đã nghiên cứu, đề xuất phương án cải tiến mô hình GCN ban đầu để có thể tích hợp và xử lý đồng thời cả hai nguồn dữ liệu trên.

Xây dựng, Huấn luyện và Đánh giá Mô hình:

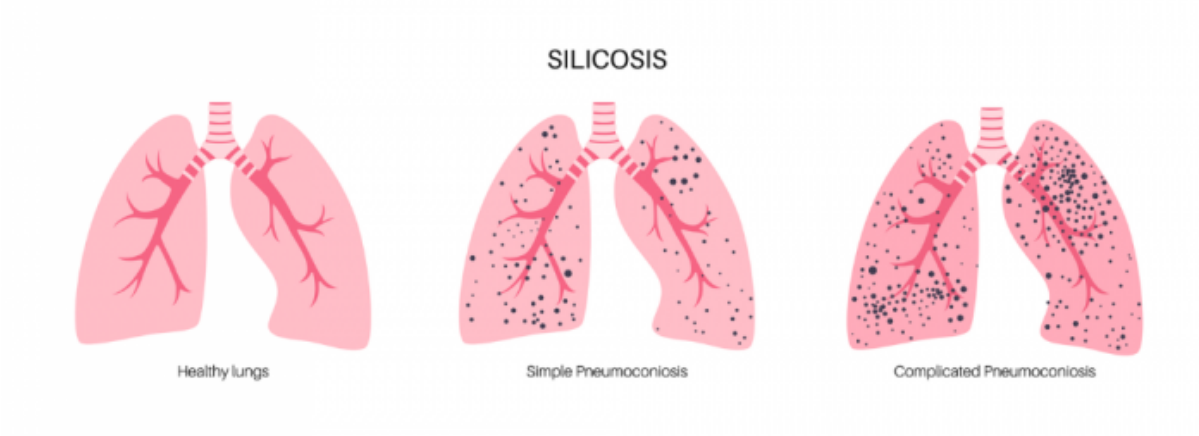
- **Triển khai thực nghiệm:** Áp dụng các mô hình đã nghiên cứu và cải tiến vào bộ dữ liệu thực tế.
- **Thiết kế các kịch bản thử nghiệm:**
  - **Kịch bản 1 (Baseline):** Đánh giá khả năng tương thích của mô hình GCN cũ với bộ dữ liệu mới, chỉ sử dụng dữ liệu dạng bảng.
  - **Kịch bản 2 (Tích hợp ảnh):** Phát triển một kiến trúc đa đầu vào, giữ nguyên nhánh GCN cho dữ liệu bảng và xây dựng một nhánh CNN mới để trích xuất đặc trưng từ dữ liệu ảnh. Thử nghiệm với các kiến trúc CNN phổ biến như ResNet50, EfficientNet, DenseNet để tìm ra bộ trích xuất đặc trưng hiệu quả nhất.
- **Đánh giá kết quả:** Huấn luyện các mô hình trên các kịch bản khác nhau, sau đó tiến hành đánh giá hiệu năng một cách toàn diện bằng các chỉ số phù hợp để đưa ra kết luận về tính hiệu quả của phương pháp đề xuất.
- Tạo một web app để nhận ảnh và đánh giá có bệnh hay không bệnh, và tỉ lệ chính xác

PHẦN III: BÁO CÁO ĐỀ TÀI THỰC TẬP

1. CƠ SỞ LÝ THUYẾT (THEORETICAL BASIS)

1.1 Giới thiệu

Bệnh nghề nghiệp, một hệ quả trực tiếp của quá trình công nghiệp hóa và hiện đại hóa, là nhóm các bệnh lý mà nguyên nhân phát sinh có mối liên hệ trực tiếp đến các yếu tố nguy cơ tại nơi làm việc. Mối liên hệ giữa nghề nghiệp và sức khỏe đã được nhận thức từ rất sớm trong lịch sử nhân loại. Một trong những ghi nhận chi tiết đầu tiên đến từ Bernardino Ramazzini vào thế kỷ 18, người được xem là cha đẻ của y học lao động, khi ông mô tả một cách hệ thống các bệnh lý đặc thù của hàng chục ngành nghề khác nhau, từ bệnh "ho của thợ đá" (bệnh bụi phổi silic) đến các vấn đề sức khỏe của những người thợ mạ vàng do tiếp xúc với thủy ngân. Những nghiên cứu này đã đặt nền móng cho sự hiểu biết rằng môi trường lao động chứa đựng những rủi ro tiềm ẩn, có khả năng gây ra những tổn thương sức khỏe lâu dài và nghiêm trọng cho người lao động.



Sự phát triển của học máy (machine learning) đã mở ra nhiều hướng tiếp cận mới cho bài toán chẩn đoán y khoa, với các thuật toán kinh điển như Support Vector Machines (SVM), Random Forests (RF), và các Mạng Nơ-ron Nhân tạo (ANN). Mặc dù các phương pháp này đã đạt được những thành công nhất định, chúng thường đối mặt với một thách thức cố hữu của dữ liệu y tế: **tính không đồng nhất và thiếu hụt thông tin**. Dữ liệu bệnh án trong thực tế hiếm khi đầy đủ và có cấu trúc đồng nhất giữa các bệnh nhân, khiến việc áp dụng các mô hình yêu cầu đầu vào dạng bảng cố định trở nên khó khăn.

Để vượt qua rào cản này, Mạng Nơ-ron Đồ thị (Graph Neural Networks - GNNs) đã nổi lên như một giải pháp đầy hứa hẹn. Thay vì buộc dữ liệu phải tuân theo một khuôn mẫu cứng nhắc, GNN cho phép biểu diễn mỗi bệnh nhân như một đồ thị độc lập, trong đó các đặc trưng y khoa là các nút (node) và các mối liên hệ giữa chúng là các cạnh (edge). Cách tiếp cận này mang lại sự linh hoạt vượt trội, cho phép mô hình xử lý hiệu quả các bệnh án bị thiếu hụt thông tin mà không cần loại bỏ các đặc trưng quan trọng. Mô hình có thể học từ cả cấu trúc và thuộc tính của dữ liệu, nắm bắt các mối tương quan phức tạp mà các phương pháp truyền thống có thể bỏ qua. Tìm hiểu cải thiện từ Graph Convolutional Network for Occupational Disease Prediction with Multiple Dimensional Data bởi Khanh Nguyen-Trong, Tuan Vu-Van, Phuong Luong Thi Bich và "Foacto/OccupationalDiseaseWebAPI"

Bên cạnh dữ liệu bệnh án, ảnh X-quang lồng ngực cung cấp một nguồn thông tin trực quan vô giá, chứa đựng các dấu hiệu tổn thương mà đôi khi không thể hiện rõ qua các chỉ số lâm sàng. Việc phân tích các ảnh này đòi hỏi kiến thức chuyên môn sâu và kinh nghiệm dày dặn. Các Mạng Nơ-ron Tích chập (Convolutional Neural Networks - CNNs) đã chứng tỏ năng lực vượt trội trong việc tự động trích xuất các đặc trưng hình ảnh từ dữ liệu y tế. Bằng cách kết hợp GNN để phân tích dữ liệu bệnh án có cấu trúc và CNN để phân tích dữ liệu ảnh phi cấu trúc, chúng ta có thể xây dựng một hệ thống chẩn đoán đa phương thức, mô phỏng gần hơn với quy trình chẩn đoán tổng hợp của một bác sĩ chuyên khoa.

Trong nghiên cứu này, Em đề xuất một kiến trúc học sâu đa đầu vào, kết hợp Mạng Tích chập Đồ thị (GCN) và Mạng Nơ-ron Tích chập, nhằm xây dựng một mô hình hỗ trợ chẩn đoán bệnh nghề nghiệp. Mô hình của eem được thiết kế để học đồng thời từ hai nguồn dữ liệu dị thể, tận dụng sự linh hoạt của GCN để xử lý dữ liệu bệnh án không đồng nhất và sức mạnh của CNN để khai thác thông tin từ ảnh X-quang. Em kỳ vọng rằng cách tiếp cận tổng hợp này sẽ cải thiện đáng kể độ chính xác và độ tin cậy trong việc sàng lọc và phát hiện sớm nguy cơ bệnh tật.

1.2 Mục tiêu

Báo cáo này trình bày cơ sở lý thuyết, phương pháp luận và quy trình triển khai một hệ thống học sâu nhằm hỗ trợ chẩn đoán bệnh bụi phổi silic. Hệ thống

tận dụng sức mạnh của hai nguồn dữ liệu dị thể: dữ liệu bệnh án dạng bảng (structured data) và dữ liệu hình ảnh X-quang phổi (unstructured data) để đưa ra dự đoán với độ tin cậy cao hơn.

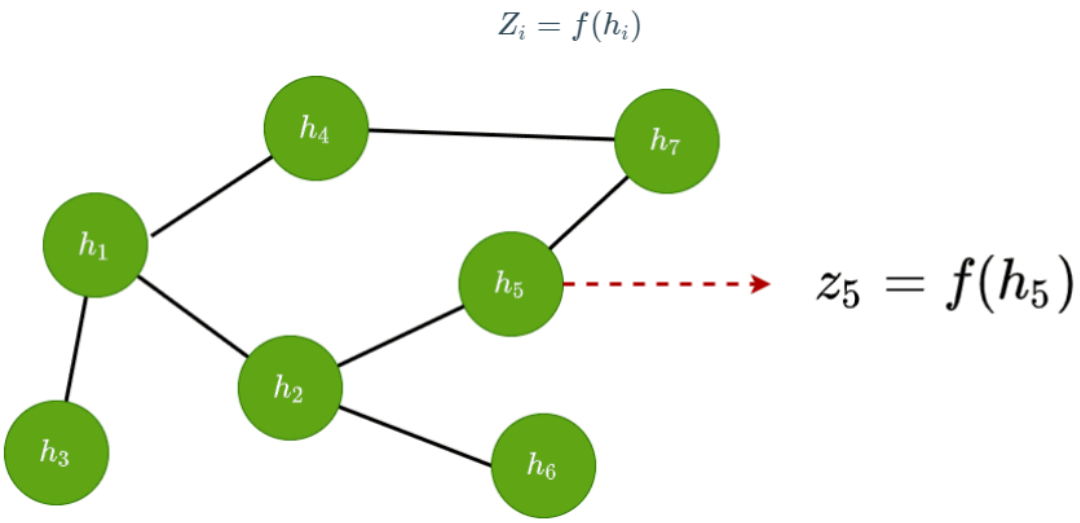
Nền tảng của dự án được xây dựng dựa trên sự kết hợp của ba lĩnh vực tiên tiến trong Trí tuệ Nhân tạo: Học sâu trên Đồ thị, Thị giác Máy tính ứng dụng Học chuyển tiếp và Học Đa phương thức.

1.3. Học sâu trên Đồ thị (Graph Deep Learning)

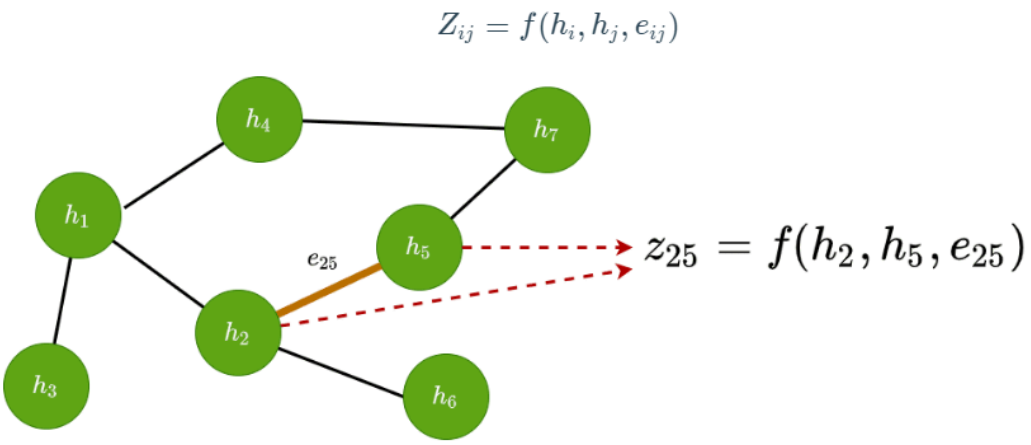
Trong lĩnh vực y khoa, thông tin bệnh án không phải là các thực thể độc lập mà tồn tại trong một mạng lưới quan hệ phức tạp. Ví dụ, triệu chứng "ho" có mối liên hệ mật thiết với "tiền sử hút thuốc" và "kết quả đo chức năng phổi". Các mô hình học máy truyền thống xử lý dữ liệu dạng bảng thường bỏ qua các mối quan hệ nội tại này.

Mạng Nơ-ron Đồ thị (Graph Neural Networks - GNNs) ra đời để giải quyết vấn đề này. Thay vì xem mỗi bệnh nhân là một hàng dữ liệu, GNN biểu diễn mỗi bệnh nhân như một đồ thị, trong đó:

- **Node (Nút):** Đại diện cho một đặc trưng cụ thể (ví dụ: tuổi, số năm làm việc, chỉ số FEV1).



- **Edge (Cạnh):** Đại diện cho mối quan hệ y khoa hoặc logic giữa các đặc trưng (ví dụ: cạnh nối giữa node "hút thuốc" và node "ho").



Cơ chế hoạt động cốt lõi của GNN là **lan truyền thông điệp (message passing)**, cho phép mỗi node tổng hợp thông tin từ các node lân cận. Qua nhiều tầng GNN, một node có thể học được biểu diễn (embedding) không chỉ từ giá trị của chính nó mà còn từ ngữ cảnh của toàn bộ mạng lưới đặc trưng, giúp mô hình nắm bắt được các quy luật phức tạp và tinh vi hơn.

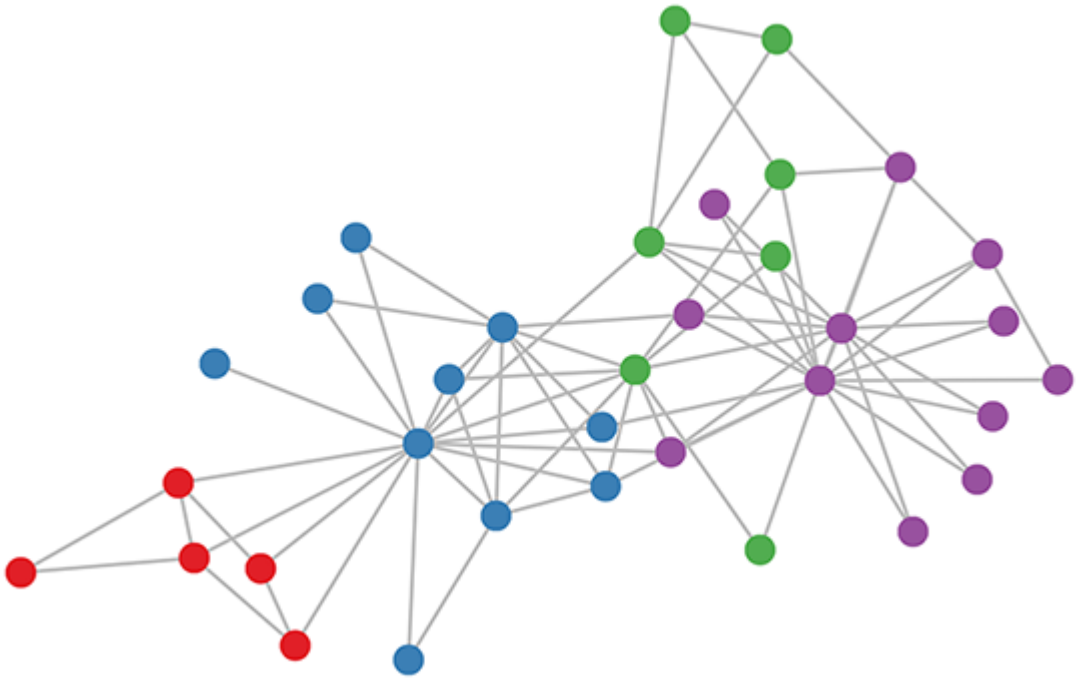
1.4. Mạng Nơ-ron Đồ thị (Graph Convolutional Network – GCN)

GCN (Graph Convolutional Network) là một kiến trúc cụ thể trong GNN, được giới thiệu nổi bật bởi Kipf & Welling (2016).Mạng Nơ-ron Đồ thị (GCN) là một lớp các kiến trúc học sâu được thiết kế đặc biệt để làm việc với dữ liệu có cấu trúc đồ thị. Khác với các loại dữ liệu truyền thống như hình ảnh (cấu trúc lưới) hay văn bản (cấu trúc tuần tự), dữ liệu đồ thị biểu diễn các thực thể và các mối quan-hệ phức tạp giữa chúng, vốn là một dạng cấu trúc tự nhiên và phổ biến trong nhiều lĩnh vực, bao gồm cả y học.

Về mặt toán học, một đồ thị G được định nghĩa là một cặp G = (V, E), trong đó V là một tập hợp các đỉnh (vertices) hay còn gọi là nút (nodes), và E là một

tập hợp các cạnh (edges) nối các cặp nút lại với nhau. Trong bối cảnh học máy và đặc biệt là với đề tài này, khái niệm đồ thị được áp dụng để mô hình hóa dữ liệu bệnh án của mỗi bệnh nhân:

- **Nút (Node):** Mỗi nút đại diện cho một đặc trưng y khoa hoặc một thông tin cụ thể của bệnh nhân. Ví dụ, trong một bệnh án, các nút có thể là tuổi, số năm làm việc trong môi trường độc hại, chỉ số FEV1, tiền sử hút thuốc, v.v.
- **Cạnh (Edge):** Mỗi cạnh biểu diễn một mối quan-hệ logic, y khoa hoặc thống kê giữa hai đặc trưng. Ví dụ, một cạnh có thể được tạo ra để nối nút tiền sử hút thuốc và nút triệu chứng ho, phản ánh mối liên hệ nhân quả đã biết giữa hai yếu tố này.

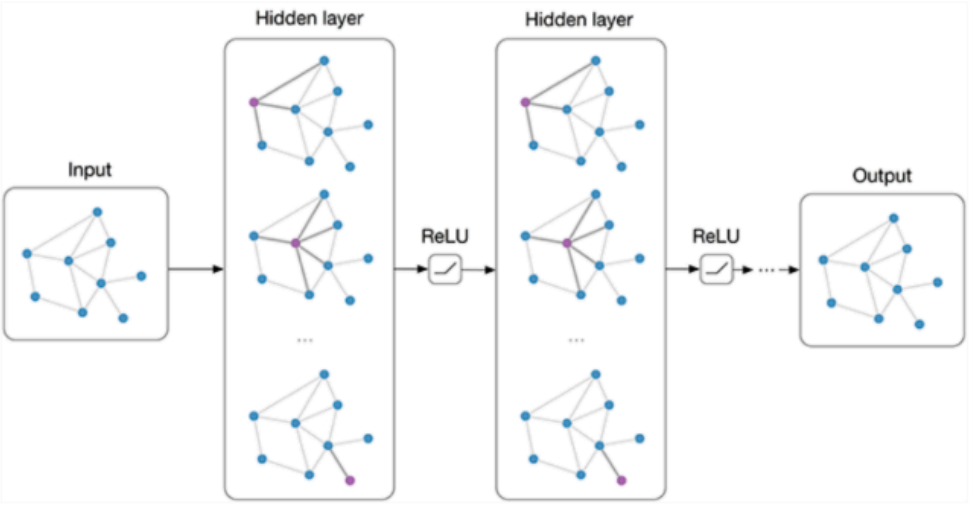


Để máy tính có thể xử lý, một đồ thị thường được biểu diễn bằng hai ma trận chính:

- **Ma trận Đặc trưng Nút (Node Feature Matrix - X):** Một ma trận có kích thước (N × D), trong đó N là số lượng nút và D là số chiều của vector đặc trưng cho mỗi nút. Trong dự án này, mỗi nút đại diện cho một chỉ số y khoa duy nhất, do đó D=1.
- **Ma trận Kề (Adjacency Matrix - A):** Một ma trận vuông có kích thước (N × N), trong đó phần tử A(i, j) = 1 nếu có một cạnh nối giữa nút i và

nút j, và bằng 0 nếu không có. Ma trận này mã hóa toàn bộ **cấu trúc** và các mối quan-hệ của đồ thị.

Việc biểu diễn dữ liệu dưới dạng đồ thị cho phép các mô hình học máy không chỉ học từ các giá trị đặc trưng riêng lẻ (ma trận X) mà còn học từ cấu trúc quan-hệ phức tạp giữa chúng (ma trận A), một lợi thế vượt trội so với các phương pháp truyền thống.



Nguyên lý cốt lõi của Mạng Tích chập Đồ thị (GCN) dựa trên khái niệm **lan truyền thông điệp (message passing)** hay **tổng hợp lân cận (neighborhood aggregation)**. Ý tưởng cơ bản là biểu diễn của một nút không chỉ được quyết định bởi chính nó, mà còn bởi các nút lân cận kết nối trực tiếp với nó. Quá trình này diễn ra theo từng lớp, tương tự như các lớp trong mạng nơ-ron truyền thống.

Tại mỗi lớp GCN, việc cập nhật biểu diễn cho một nút (node) diễn ra qua hai bước chính:

1. **Bước Tổng hợp (Aggregation):** Đối với một nút mục tiêu, mô hình sẽ thu thập các vector đặc trưng từ tất cả các nút hàng xóm của nó (các nút được kết nối trực tiếp). Quá trình này giống như việc một nút "lắng nghe" thông tin từ môi trường xung quanh nó.
2. **Bước Cập nhật (Update):** Thông tin đã được tổng hợp từ các nút hàng xóm, cùng với thông tin của chính nút mục tiêu từ lớp trước đó, sẽ được đưa qua một phép biến đổi. Phép biến đổi này thường bao gồm một phép nhân với một ma trận trọng số có thể học được (W) và sau đó đi qua một hàm kích hoạt phi tuyến (ví dụ: ReLU). Ma trận trọng số W này được chia sẻ trên tất cả các nút, tương tự như cách một bộ lọc (kernel) trong

mạng CNN được áp dụng trên toàn bộ ảnh. Đây chính là quá trình "học" của GCN, nơi mô hình học cách kết hợp thông tin từ chính nó và các nút lân cận một cách hiệu quả nhất.

Khi xếp chồng nhiều lớp GCN lên nhau, phạm vi "lắng nghe" của mỗi nút được mở rộng. Sau lớp đầu tiên, một nút học được từ các hàng xóm cách nó 1 bước (1-hop). Sau lớp thứ hai, nó có thể tổng hợp thông tin từ các hàng xóm của hàng xóm, tức là các nút cách nó 2 bước (2-hop). Quá trình này cho phép mô hình nắm bắt các mối quan-hệ ở phạm vi rộng hơn và các quy luật phức tạp hơn trên toàn bộ đồ thị.

Ưu điểm:

- **Xử lý Dữ liệu Phi cấu trúc (Non-Euclidean):** Đây là ưu điểm lớn nhất. GCN được thiết kế tự nhiên để làm việc với dữ liệu đồ thị, vốn không có cấu trúc dạng lưới như hình ảnh. Điều này làm cho nó trở nên lý tưởng cho các bài toán như phân tích mạng xã hội, tương tác thuốc, và đặc biệt là dữ liệu bệnh án không đồng nhất.
- **Khai thác Thông tin Quan hệ:** GCN có khả năng học đồng thời từ cả đặc trưng của các nút và cấu trúc kết nối giữa chúng, cho phép nó phát hiện các mẫu tinh vi dựa trên mối quan-hệ mà các mô hình truyền thống thường bỏ qua.
- **Tính linh hoạt với Dữ liệu Thiếu hụt:** Đối với một bệnh nhân bị thiếu một vài chỉ số y khoa, đồ thị tương ứng của họ sẽ chỉ đơn giản là có ít nút hơn. GCN có thể xử lý các đồ thị có kích thước và cấu trúc khác nhau một cách tự nhiên, không yêu cầu tất cả các mẫu đầu vào phải có cùng một bộ đặc trưng cố định.
- **Chia sẻ Tham số (Parameter Sharing):** Tương tự như CNN, GCN sử dụng cùng một bộ trọng số (bộ lọc) để xử lý thông tin trên tất cả các nút, giúp mô hình trở nên hiệu quả về mặt tính toán và có khả năng tổng quát hóa tốt hơn.

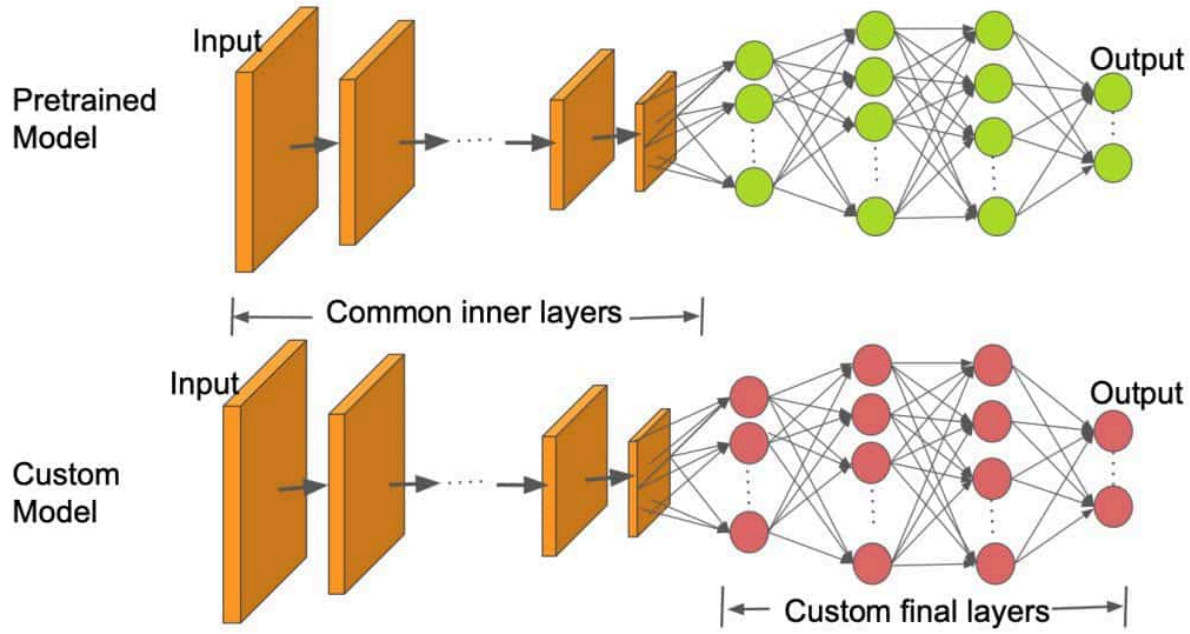
Nhược điểm:

- **Hiện tượng Làm mịn quá mức (Oversmoothing):** Khi xếp chồng quá nhiều lớp GCN, các biểu diễn của các nút trong đồ thị có xu hướng trở nên rất giống nhau. Điều này xảy ra do thông tin được trung bình hóa qua một vùng lân cận ngày càng lớn, làm mất đi các đặc trưng riêng biệt và hữu ích của từng nút.

- **Khả năng mở rộng:** Mặc dù hiệu quả, việc huấn luyện GCN trên các đồ thị đơn lẻ có quy mô cực lớn (hàng tỷ nút và cạnh) vẫn là một thách thức về mặt tính toán và bộ nhớ. Tuy nhiên, vấn đề này ít ảnh hưởng đến dự án hiện tại, vốn xử lý nhiều đồ thị nhỏ thay vì một đồ thị khổng lồ.
- **Giả định về tính Tương đồng (Homophily):** Hiệu năng của GCN tốt nhất khi các nút được kết nối với nhau có xu hướng giống nhau hoặc thuộc cùng một lớp. Trong các bài toán mà các kết nối biểu thị sự khác biệt, hiệu quả của GCN có thể bị giảm sút.

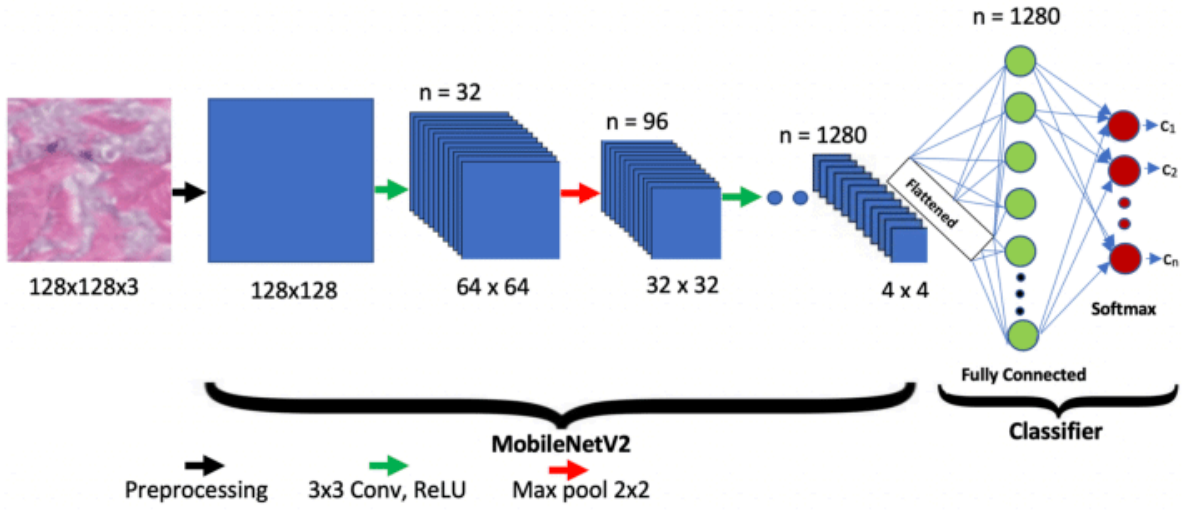
1.5. Thị giác Máy tính và Học chuyển tiếp (Computer Vision & Transfer Learning)

Ảnh X-quang là một nguồn thông tin trực quan quan trọng nhưng phi cấu trúc. Để "đọc" và "hiểu" được các ảnh này, Em sử dụng Mạng Tích chập (Convolutional Neural Networks - CNNs), một kiến trúc học sâu tiêu chuẩn trong lĩnh vực thị giác máy tính, có khả năng tự động học các đặc trưng hình ảnh từ cấp thấp (cạnh, góc) đến cấp cao (hình dạng, kết cấu tổn thương).



Tuy nhiên, việc huấn luyện một mạng CNN từ đầu đòi hỏi một bộ dữ liệu ảnh y tế cực lớn. Để khắc phục hạn chế này, dự án áp dụng kỹ thuật **Học chuyển**

tiếp (Transfer Learning). Cụ thể, Em sử dụng một mô hình CNN (ví dụ: MobileNetV2) đã được huấn luyện trước trên một bộ dữ liệu khổng lồ (ImageNet). Mô hình này đã học được cách nhận diện hàng ngàn loại đặc trưng hình ảnh khác nhau. Em tận dụng kiến thức nền tảng này, loại bỏ lớp phân loại cuối cùng của nó và sử dụng nó như một **bộ trích xuất đặc trưng (feature extractor)** mạnh mẽ, chuyển đổi mỗi ảnh X-quang thành một vector số học cô đọng, giàu thông tin.

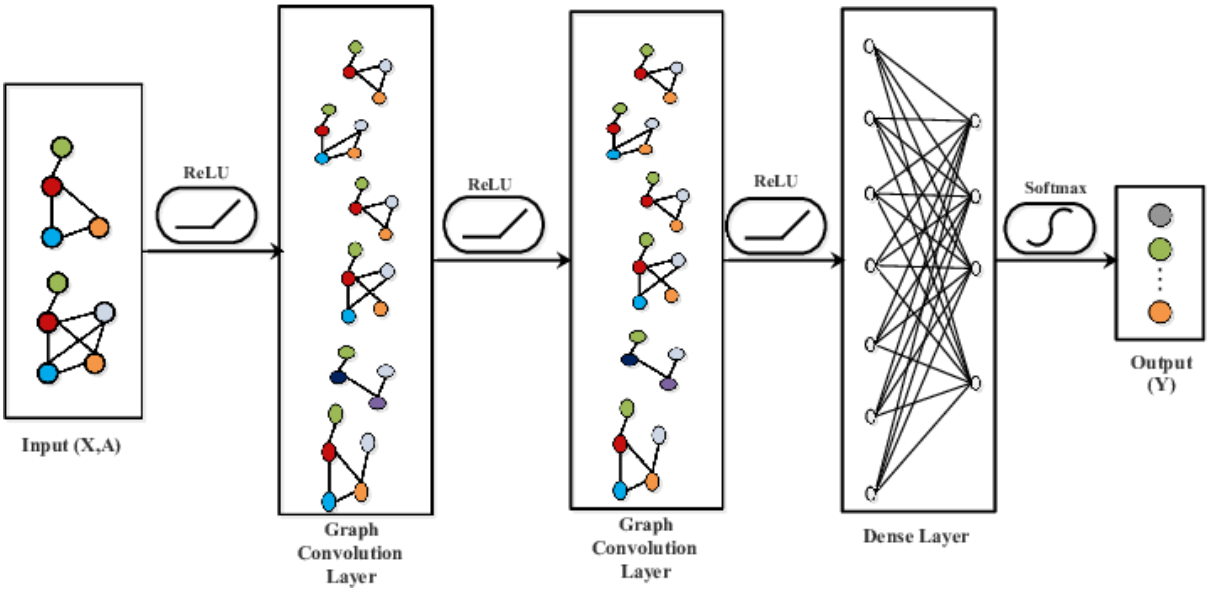


1.6. Học Đa phương thức (Multimodal Learning)

Thực tế chẩn đoán y khoa đòi hỏi sự tổng hợp thông tin từ nhiều nguồn. Một hệ thống AI chỉ dựa trên dữ liệu bảng hoặc chỉ dựa trên hình ảnh đều sẽ thiếu sót. Học Đa phương thức là phương pháp kết hợp thông tin từ các loại dữ liệu (phương thức) khác nhau để đưa ra quyết định chính xác hơn.

Dự án triển khai một **kiến trúc đa đầu vào (multi-input)**, một dạng của kỹ thuật **kết hợp muộn (late fusion)**. Trong kiến trúc này:

- Một nhánh chuyên biệt(mạng GCN) xử lý dữ liệu đồ thị từ bệnh án.
- Một nhánh chuyên biệt khác (mạng MLP) xử lý vector đặc trưng từ ảnh X-quang.
- Các biểu diễn (embedding) đầu ra từ hai nhánh này sau đó được ghép lại (concatenate) và đưa vào một khối phân loại cuối cùng để đưa ra dự đoán tổng hợp.



Cách tiếp cận này cho phép mỗi nhánh của mô hình học sâu các đặc trưng riêng của từng loại dữ liệu trước khi kết hợp chúng lại để có một cái nhìn toàn diện.

2. PHƯƠNG PHÁP LUẬN VÀ LỰA CHỌN KỸ THUẬT

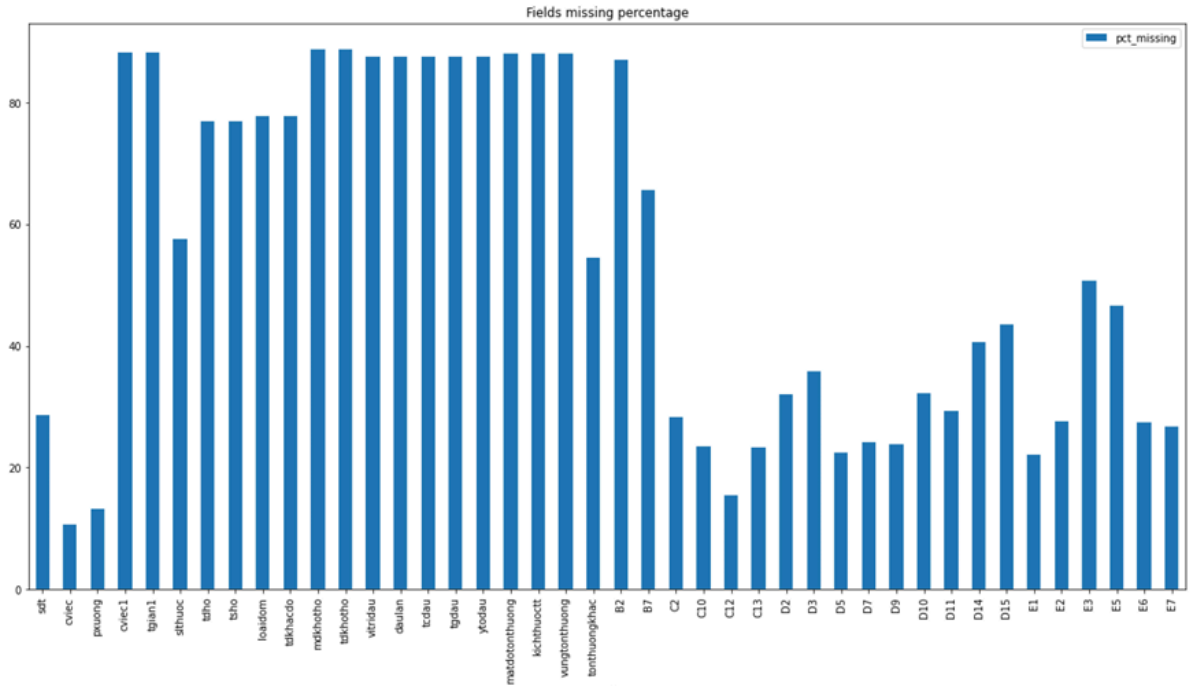
Dựa trên cơ sở lý thuyết, Em đã xây dựng một phương pháp luận có hệ thống để giải quyết bài toán.

2.1. Lựa chọn Mô hình và Nền tảng

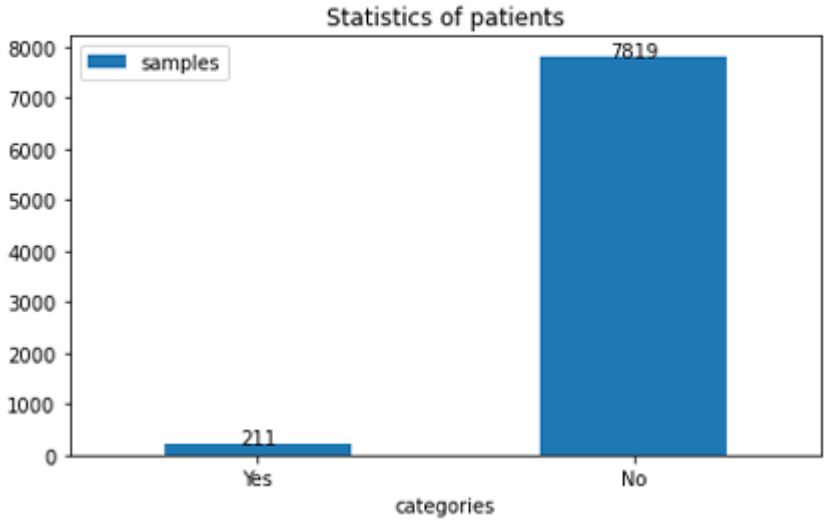
- **Mô hình cho Dữ liệu Bảng:** Mạng Tích chập Đồ thị (Graph Convolutional Network - GCN) được lựa chọn do khả năng mô hình hóa hiệu quả các mối quan hệ giữa các đặc trưng y khoa.
- **Mô hình cho Dữ liệu Ảnh:** MobileNetV2 được chọn làm bộ trích xuất đặc trưng do sự cân bằng tốt giữa hiệu năng và tốc độ, phù hợp cho việc triển khai ứng dụng.
- **Nền tảng:** PyTorch và thư viện PyTorch Geometric (PyG) được lựa chọn thay thế cho kiến trúc Stellargraph/Keras cũ. Lý do chính là PyG cung cấp một giao diện linh hoạt, hiệu suất cao, được tối ưu hóa cho các nghiên cứu về đồ thị và được cộng đồng hỗ trợ mạnh mẽ.

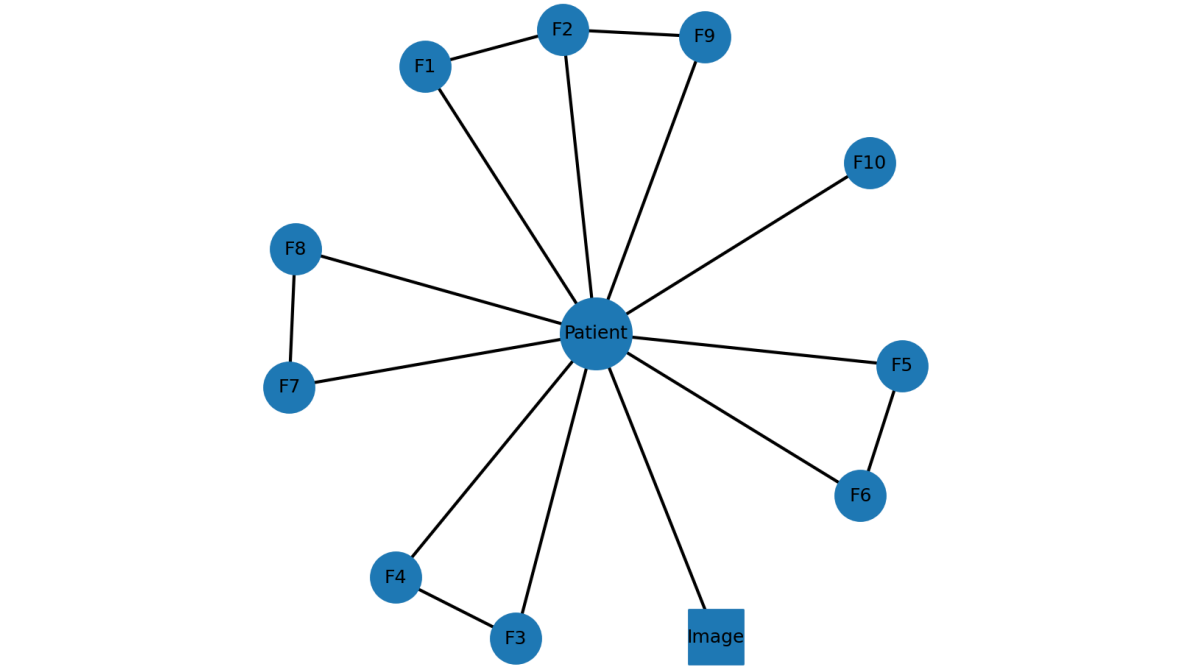
2.2. Biểu diễn và Cấu trúc hóa Dữ liệu

- **Biểu diễn Dữ liệu Bảng:** Mỗi bệnh nhân được chuyển đổi thành một đồ thị riêng biệt. Mỗi cột trong file Excel (sau khi loại bỏ các thông tin định danh) được coi là một node. Giá trị của cột sau khi chuẩn hóa là thuộc tính của node đó.



- **Cấu trúc Đồ thị:** Để nắm bắt tối đa thông tin, một cấu trúc đồ thị hỗn hợp được áp dụng:
    1. **Nền tảng kết nối đầy đủ:** Ban đầu, tất cả các node trong đồ thị của một bệnh nhân được kết nối với nhau, cho phép mô hình tự do khám phá mọi mối tương quan tiềm ẩn.
    2. **Bổ sung cạnh chuyên biệt:** Dựa trên tri thức y khoa từ mô hình cũ, các cạnh bổ sung được tạo ra để nhấn mạnh các mối quan hệ đã biết (ví dụ: nhóm triệu chứng về ho, nhóm tiền sử nghề nghiệp,...).





- **Biểu diễn Dữ liệu Ảnh:** Mỗi ảnh X-quang được chuyển đổi thành một vector đặc trưng 1280 chiều duy nhất. Đối với các bệnh nhân không có ảnh, một vector không (zero-vector) được sử dụng để biểu thị sự thiếu vắng thông tin này, đảm bảo tính nhất quán của dữ liệu đầu vào.

2.3. Xử lý Vấn đề Mất cân bằng Dữ liệu và vấn đề số lượng dữ liệu còn thiếu

Phân tích sơ bộ cho thấy bộ dữ liệu có sự mất cân bằng nghiêm trọng giữa số lượng ca "Bệnh" và "Không Bệnh". Nếu không xử lý, mô hình sẽ có xu hướng dự đoán thiên về lớp chiếm đa số. Để giải quyết vấn đề này, kỹ thuật **trọng số lớp (class weighting)** được áp dụng. Cụ thể, trong quá trình tính toán hàm mất mát (BCEWithLogitsLoss), một trọng số (pos_weight) cao hơn được gán cho

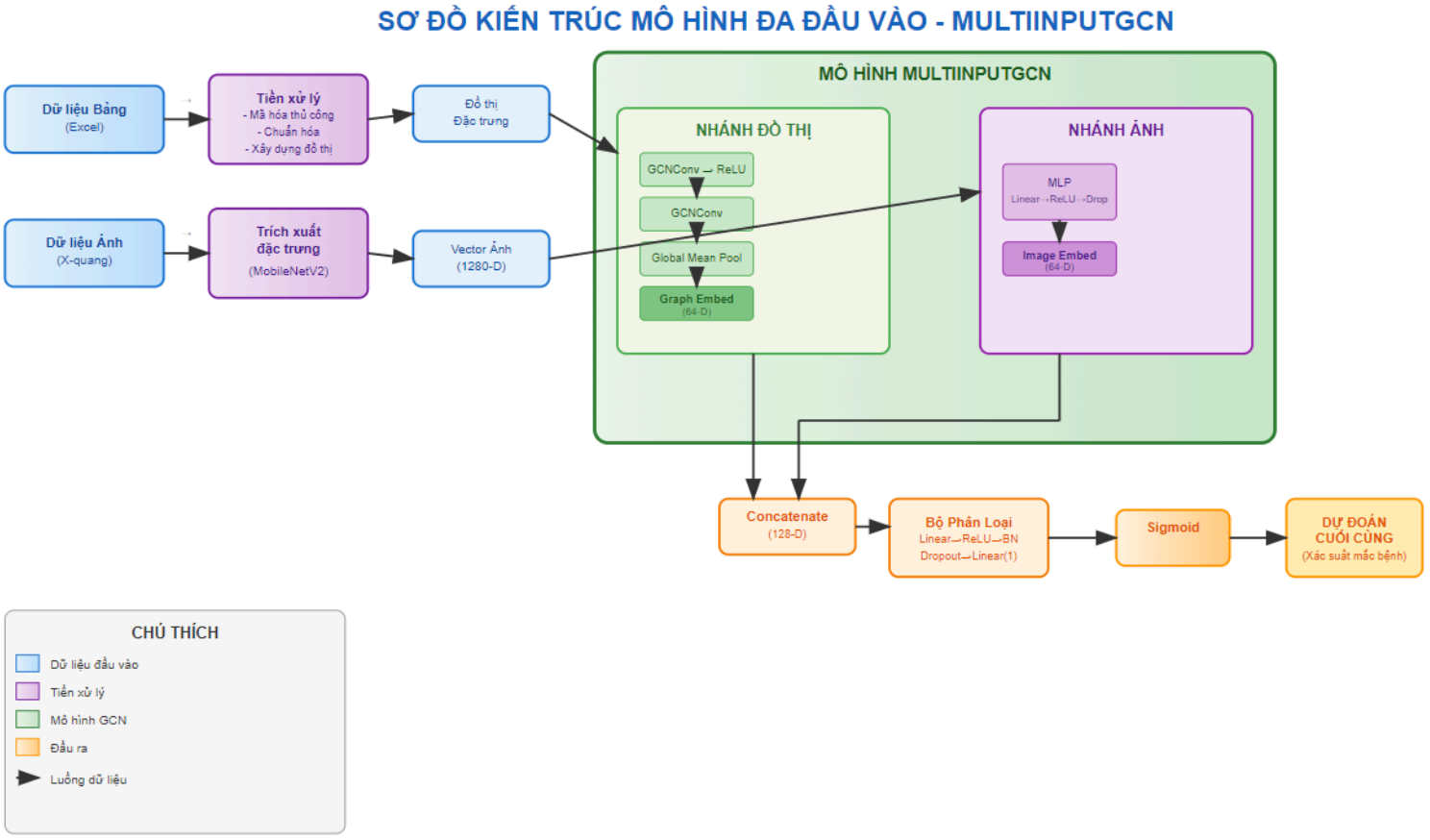
lớp thiểu số ("Bệnh"), khiến cho việc dự đoán sai một ca bệnh thật sẽ bị "phạt" nặng hơn, buộc mô hình phải nỗ lực học cách nhận diện lớp này.

Với tỷ lệ mất cân bằng 1:37, số lượng mẫu thuộc lớp "Bệnh" là cực kỳ hạn chế. Điều này khiến mô hình không có đủ các trường hợp đa dạng để học được các quy luật nhận diện một cách vững chắc.

3. QUÁ TRÌNH TRIỂN KHAI THỰC NGHIỆM

Quá trình triển khai được chia thành các giai đoạn tuần tự và logic, từ xử lý dữ liệu thô đến huấn luyện và đánh giá mô hình.

SƠ ĐỒ KIẾN TRÚC:



3.1. Giai đoạn Tiền xử lý và Chuẩn bị Dữ liệu

Đây là giai đoạn nền tảng và quan trọng nhất.

1. **Đọc và Làm sạch Dữ liệu:** Dữ liệu được đọc từ file Excel. Các hàng có nhãn không hợp lệ hoặc thiếu nhãn (bnn) sẽ bị loại bỏ. Cột nhãn được ánh xạ từ "Co"/"Khong" sang định dạng nhị phân (1/0).
2. **Tách Dữ liệu:** Dữ liệu được tách thành hai phần chính: DataFrame chứa các đặc trưng huấn luyện và Series chứa tên file ảnh.
3. **Xử lý Dữ liệu Bảng:** Toàn bộ logic mã hóa thủ công từ mô hình cũ được tái áp dụng để đảm bảo tính nhất quán. Các quy tắc replace được thực thi để chuyển đổi các giá trị dạng chữ và hạng mục sang dạng số. Sau đó, dữ

liệu được chuẩn hóa bằng cách sử dụng giá trị trung bình và lớn nhất tính toán từ tập huấn luyện.

4. **Xử lý Dữ liệu Ảnh:** Một vòng lặp duyệt qua danh sách tên file ảnh. Với mỗi tên file, đường dẫn đầy đủ được tạo và đưa vào hàm trích xuất đặc trưng. Hàm này sử dụng MobileNetV2 để chuyển đổi ảnh thành vector và xử lý các trường hợp không có ảnh bằng cách trả về một vector không.
5. **Phân chia Dữ liệu:** Toàn bộ bộ dữ liệu (bao gồm dữ liệu bảng, vector ảnh và nhãn) được phân chia thành hai tập: Huấn luyện (80%) và Kiểm tra (20%) bằng cách sử dụng train_test_split. Quá trình này đảm bảo tỷ lệ phân bổ giữa các lớp được giữ nguyên ở cả hai tập (stratify).

3.2. Giai đoạn Xây dựng Cấu trúc Dữ liệu cho PyTorch Geometric

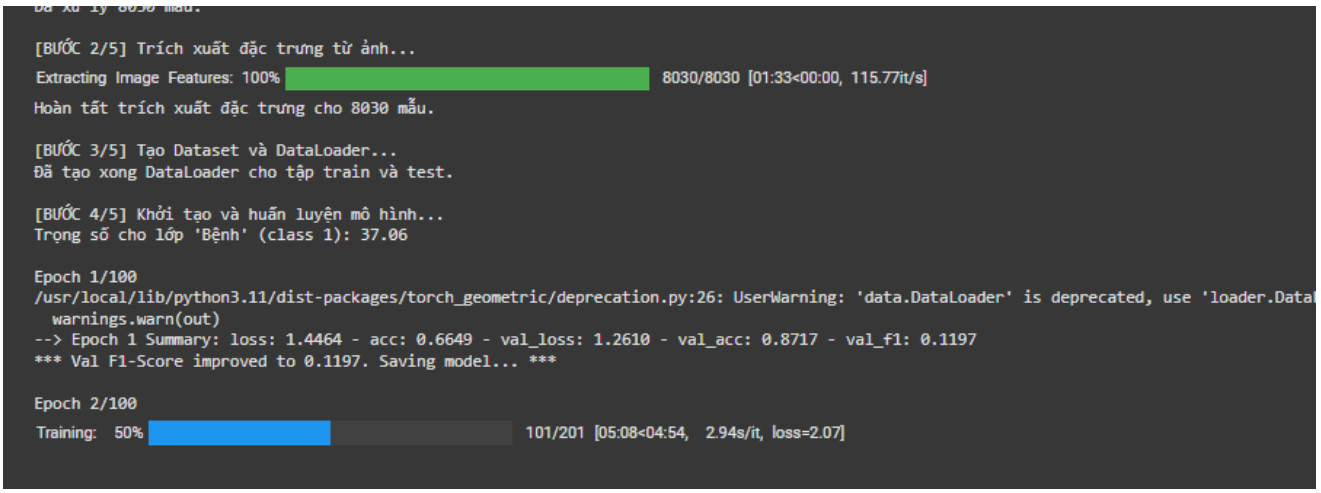
Một lớp Dataset tùy chỉnh được tạo ra để PyTorch Geometric có thể làm việc. Lớp này thực hiện việc chuyển đổi từng hàng dữ liệu bảng trong các tập huấn luyện/kiểm tra thành một đối tượng Data của PyG. Mỗi đối tượng Data chứa x (ma trận đặc trưng của các node) và edge_index (cấu trúc kết nối của các cạnh), sẵn sàng để được đưa vào mô hình.

3.3. Giai đoạn Thiết kế và Huấn luyện Mô hình

1. **Khởi tạo:** Mô hình MultiInputGCN, hàm tối ưu Adam, và hàm mất mát BCEWithLogitsLoss (với trọng số lớp đã tính) được khởi tạo.
2. **Vòng lặp Huấn luyện:** Quá trình huấn luyện diễn ra qua nhiều epoch. Trong mỗi epoch, dữ liệu từ tập huấn luyện được chia thành các batch nhỏ. Với mỗi batch, mô hình thực hiện lượt đi xuôi (forward pass), tính toán giá trị mất mát, thực hiện lan truyền ngược (backpropagation) để tính toán gradient, và cuối cùng cập nhật trọng số thông qua hàm tối ưu.

3.4. Giai đoạn Đánh giá và Lưu trữ Mô hình

Sau mỗi epoch huấn luyện, mô hình được chuyển sang chế độ đánh giá và chạy trên toàn bộ tập kiểm tra. Các chỉ số hiệu năng như val_loss, val_acc, và đặc biệt là val_f1 (do dữ liệu mất cân bằng) được tính toán. Nếu chỉ số val_f1 của epoch hiện tại vượt qua giá trị tốt nhất đã ghi nhận, trạng thái của mô hình sẽ được lưu lại. Quá trình này được lặp lại cho đến khi val_f1 không còn cải thiện trong một số lượng epoch nhất định (cơ chế Early Stopping).


[BƯỚC 2/5] Trích xuất đặc trưng từ ảnh...
Extracting Image Features: 100% 8030/8030 [01:33<00:00, 115.77it/s]
Hoàn tất trích xuất đặc trưng cho 8030 mẫu.
[BƯỚC 3/5] Tạo Dataset và DataLoader...
Đã tạo xong DataLoader cho tập train và test.
[BƯỚC 4/5] Khởi tạo và huấn luyện mô hình...
Trọng số cho lớp 'Bệnh' (class 1): 37.06
Epoch 1/100
/usr/local/lib/python3.11/dist-packages/torch_geometric/deprecation.py:26: UserWarning: 'data.DataLoader' is deprecated, use 'loader.Data
warnings.warn(out)
--> Epoch 1 Summary: loss: 1.4464 - acc: 0.6649 - val_loss: 1.2610 - val_acc: 0.8717 - val_f1: 0.1197
*** Val F1-Score improved to 0.1197. Saving model... ***
Epoch 2/100
Training: 50% 101/201 [05:08<04:54, 2.94s/it, loss=2.07]

4. QUY TRÌNH HUẤN LUYỆN MÔ HÌNH

4.1. Chuẩn bị dữ liệu đầu vào

- **Dữ liệu bảng**: Đọc từ file Excel, áp dụng logic mã hóa thủ công (chuyển đổi text thành số) theo phong cách legacy
- **Dữ liệu hình ảnh**: Trích xuất đặc trưng bằng MobileNet v2 pre-trained, tạo ra vector 1280 chiều cho mỗi ảnh
- **Tiền xử lý**: Chuẩn hóa dữ liệu bảng bằng công thức (value - mean) / (max - mean)

4.2. Xây dựng Graph Neural Network

- **Tạo đồ thị**: Mỗi mẫu dữ liệu được chuyển thành đồ thị, mỗi đặc trưng là một node.
- **Kết nối nodes**:
    - Kết nối tất cả nodes với nhau (fully connected)
    - Thêm các kết nối đặc biệt cho các nhóm đặc trưng có quan hệ (ví dụ: công việc-thời gian, hút thuốc-số lượng thuốc).

4.3. Kiến trúc mô hình kết hợp

- **Nhánh GCN**: Xử lý dữ liệu đồ thị bằng 2 lớp Graph Convolutional Network.
- **Nhánh Image**: Xử lý đặc trưng ảnh qua Multi-Layer Perceptron.
- **Fusion**: Nối (concatenate) 2 embedding vectors từ 2 nhánh
- **Classification**: Đưa qua classifier cuối cùng để dự đoán nhị phân(có bệnh/không bệnh).

4.4. Phân tích Quy trình huấn luyện

Quy trình huấn luyện được thiết kế như một vòng lặp có giám sát nhằm tối ưu hóa các tham số của mô hình MultiInputGCN để nó có thể học được mối liên hệ giữa dữ liệu đầu vào (bệnh án và ảnh X-quang) và kết quả chẩn đoán (Bệnh/Không Bệnh). Quá trình này mô phỏng lại cách một chuyên gia con

người học hỏi từ kinh nghiệm, bao gồm các bước học, kiểm tra, rút kinh nghiệm và ghi nhớ.

- **Khởi tạo:** Mô hình được khởi tạo với các trọng số ngẫu nhiên. Nó chưa có bất kỳ kiến thức nào về bài toán. Một hàm tối ưu (Adam optimizer) được chỉ định để thực hiện việc cập nhật trọng số, và một hàm mất mát (BCEWithLogitsLoss có trọng số lớp) được chọn để đo lường mức độ sai sót trong các dự đoán của mô hình.
- **Học theo từng Lô (Batch-wise Learning):** Thay vì học trên toàn bộ 8030 mẫu dữ liệu cùng một lúc, dữ liệu huấn luyện được chia thành các lô nhỏ (ví dụ: 32 mẫu/lô). Điều này giúp quá trình học ổn định và hiệu quả hơn về mặt bộ nhớ. Với mỗi lô, mô hình thực hiện:
    - **Lượt đi xuôi (Forward Pass):** Dữ liệu được đưa qua hai nhánh của mô hình (GCN cho dữ liệu bảng và MLP cho dữ liệu ảnh). Các kết quả được kết hợp để tạo ra một dự đoán thô (logit) cho mỗi mẫu.
    - **Tính toán Sai số (Loss Calculation):** Dự đoán thô được so sánh với kết quả chẩn đoán thực tế. Hàm mất mát tính toán một giá trị số học đại diện cho mức độ sai sót. Do có sự mất cân bằng lớp, các sai sót trên lớp "Bệnh" (thiểu số) sẽ bị "phạt" nặng hơn, buộc mô hình phải chú ý đến chúng.
    - **Lan truyền ngược và Cập nhật (Backpropagation & Optimization):** Dựa trên giá trị sai số, thuật toán lan truyền ngược tính toán trách nhiệm của từng trọng số trong mô hình đối với sai sót đó. Hàm tối ưu Adam sau đó sẽ điều chỉnh nhẹ các trọng số này theo hướng làm giảm sai số trong lần dự đoán tiếp theo.

- **Đánh giá sau mỗi Epoch (Epoch-wise Evaluation):** Một epoch hoàn thành khi mô hình đã học qua tất cả các lô trong tập huấn luyện. Ngay sau đó, mô hình được chuyển sang chế độ đánh giá và được kiểm tra trên một tập dữ liệu riêng biệt mà nó chưa từng thấy (tập kiểm tra). Các chỉ số hiệu năng chính, bao gồm độ chính xác (val_acc) và đặc biệt là F1-Score (val_f1), được tính toán. F1-Score được chọn làm thước đo chính do khả năng phản ánh hiệu năng tốt hơn trên các bộ dữ liệu mất cân bằng

- **Lưu trữ và Dừng sớm (Checkpointing & Early Stopping):** Trạng thái (các trọng số) của mô hình được lưu lại bất cứ khi nào chỉ số val_f1 trên tập kiểm tra đạt một kỷ lục mới. Nếu sau một số lượng Epoch nhất định (ví dụ: 15) mà val_f1 không thể cải thiện thêm, quá trình huấn luyện sẽ tự động dừng lại. Cơ chế này giúp tiết kiệm thời gian tính toán và ngăn chặn hiện tượng học vẹt (overfitting), đồng thời đảm bảo phiên bản mô hình cuối cùng là phiên bản có hiệu năng tốt nhất trên dữ liệu thực tế.

5. NHẬN ĐỊNH KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN

Dựa trên các kết quả thu được từ quá trình huấn luyện và đánh giá, Em đưa ra các nhận định sau:

5.1. Phân tích Kết quả Huấn luyện

- **Mô hình có khả năng học:** Quá trình huấn luyện cho thấy các chỉ số mất mát (loss) và độ chính xác (acc) trên tập huấn luyện có sự biến động, và chỉ số val_f1 trên tập kiểm tra có xu hướng tăng nhẹ trong các epoch đầu tiên (từ 0.1452 lên 0.2020). Điều này chứng tỏ kiến trúc mô hình và quy trình huấn luyện là hợp lệ, mô hình có khả năng học và điều chỉnh các tham số của nó.
- **Hiệu quả của việc xử lý mất cân bằng lớp:** Việc áp dụng trọng số lớp (pos_weight = 37.06) đã thành công trong việc ngăn chặn mô hình dự đoán hoàn toàn theo lớp đa số. Báo cáo phân loại cho thấy mô hình đã cố gắng nhận diện lớp "Bệnh", với chỉ số recall là 0.24 (tìm thấy 24% số ca bệnh thật), thay vì bỏ qua hoàn toàn.
- **Hiệu năng tổng thể còn hạn chế:** Mặc dù có khả năng học, hiệu năng cuối cùng của mô hình vẫn còn khiêm tốn. Chỉ số F1-Score tốt nhất đạt được trên tập kiểm tra là **0.2020**. Báo cáo phân loại chi tiết hơn cho thấy mặc dù recall của lớp "Bệnh" đạt 0.24, precision lại rất thấp (0.18), nghĩa là mô hình tạo ra nhiều cảnh báo dương tính giả.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$FPR = \frac{FP}{FP + TN}$$

$$ROC\text{-}AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

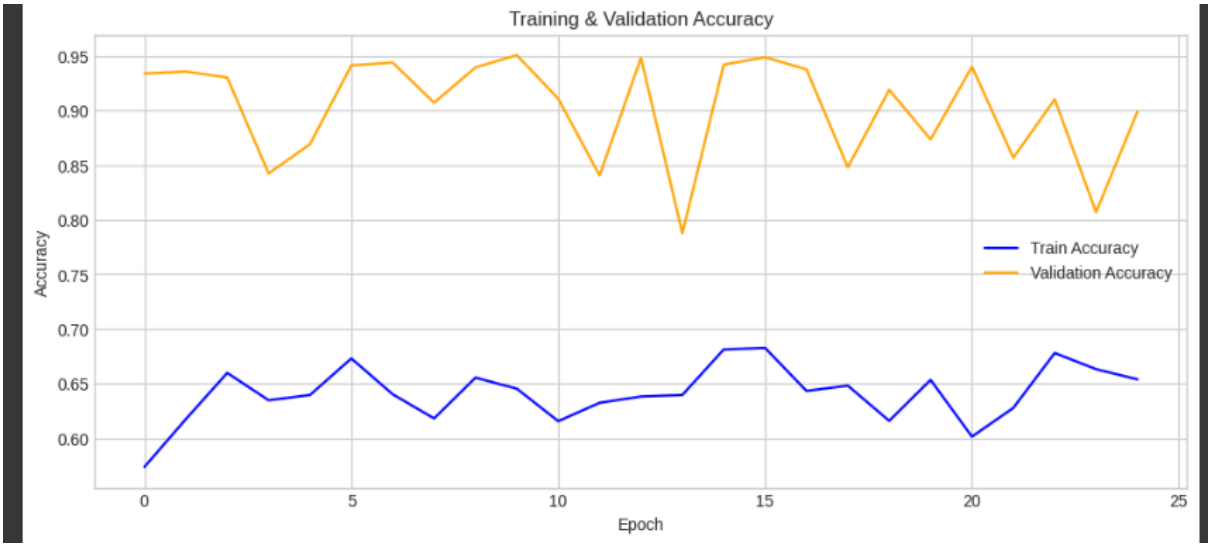
$$PR\text{-}AUC = \int_0^1 Precision(Recall)\, d(Recall)$$

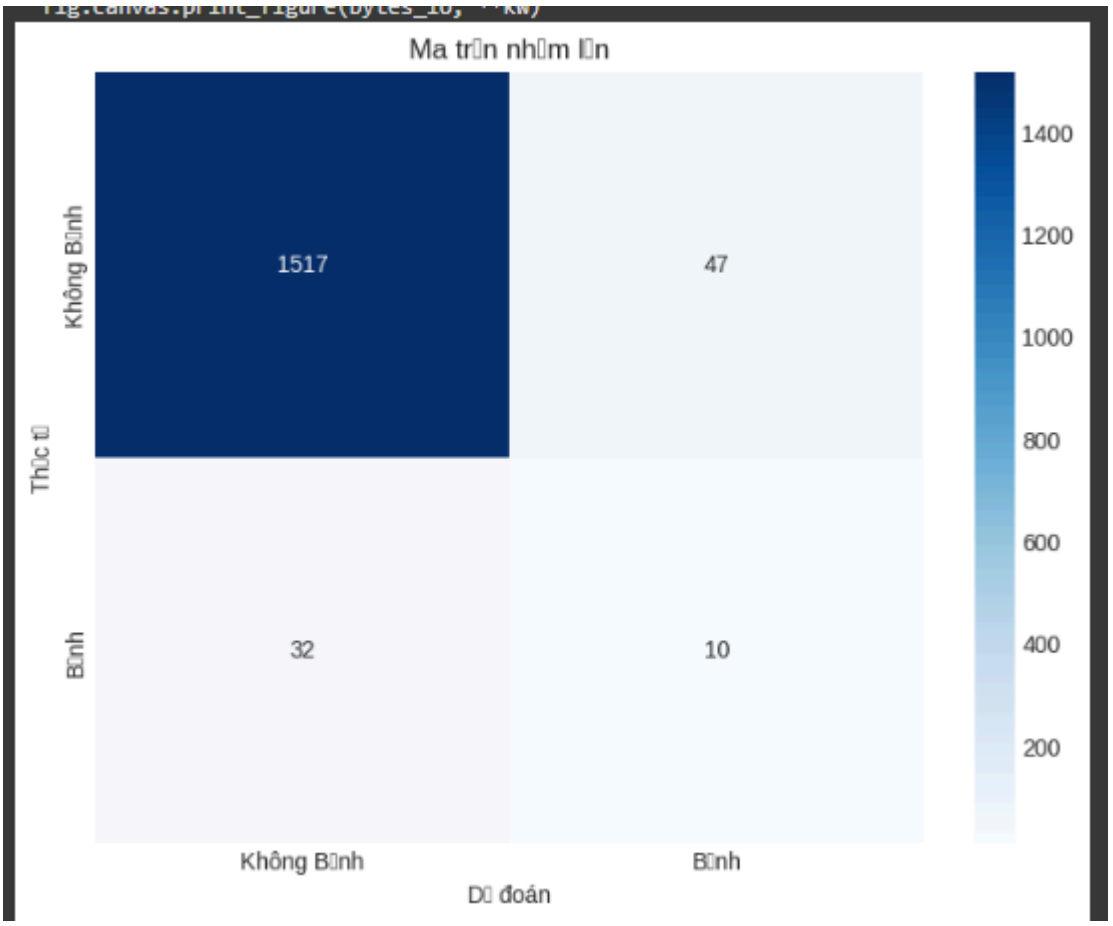
```
Classification Report trên tập Test:
                precision    recall  f1-score   support

Không Bệnh (0)       0.98      0.97      0.97      1564
      Bệnh (1)       0.18      0.24      0.20        42

      accuracy                           0.95      1606
     macro avg       0.58      0.60      0.59      1606
  weighted avg       0.96      0.95      0.95      1606
```


Training & Validation Loss
Train Loss
Validation Loss
Loss
Epoch
1.400
1.375
1.350
1.325
1.300
1.275
1.250
1.225
0
5
10
15
20
25

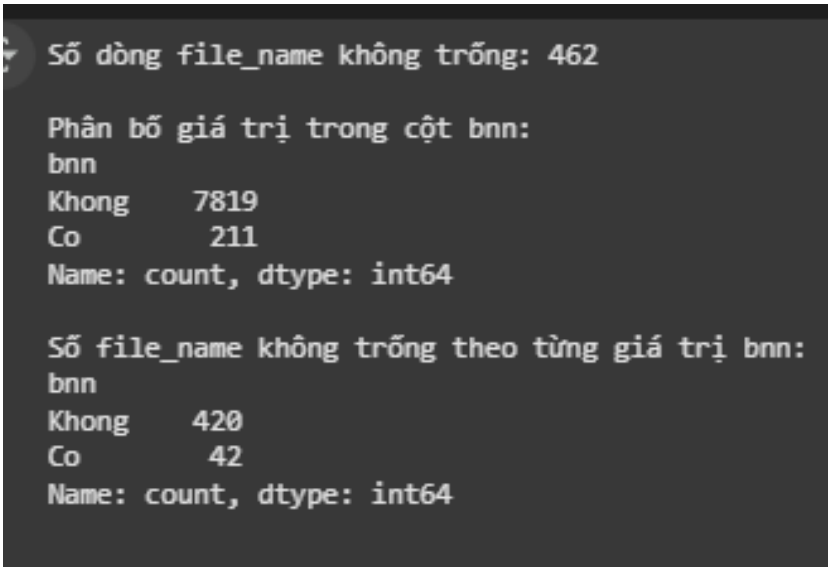

Training & Validation Accuracy
Train Accuracy
Validation Accuracy
Accuracy
Epoch
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
5
10
15
20
25



5.2. Nhận định về Nguyên nhân và Hạn chế

Kết quả trên cho thấy thách thức chính không nằm ở kiến trúc mô hình hay quy trình huấn luyện, mà có khả năng xuất phát từ chính bản chất của dữ liệu đầu vào.

- **Hạn chế về Quy mô và Chất lượng Dữ liệu:** Đây được xem là nguyên nhân cốt lõi.
    - **Số lượng mẫu thiểu số quá ít:** Với tỷ lệ mất cân bằng 1:37, số lượng mẫu thuộc lớp "Bệnh" là cực kỳ hạn chế. Điều này khiến mô hình không có đủ các trường hợp đa dạng để học được các quy luật nhận diện một cách vững chắc.
    - **Tín hiệu yếu trong Dữ liệu:** Có khả năng các đặc trưng hiện có trong dữ liệu bảng và dữ liệu ảnh không chứa đủ "tín hiệu" để phân biệt rõ ràng hai lớp. Các trường hợp bệnh có thể có các biểu hiện rất tinh vi và không đồng nhất, gây khó khăn cho mô hình trong việc tìm ra một quy luật chung.



- **Hạn chế của Học chuyển tiếp:** Mô hình MobileNetV2 được huấn luyện trước trên bộ dữ liệu ImageNet (ảnh sinh hoạt thông thường). Mặc dù mạnh mẽ, các đặc trưng mà nó học được có thể không phải là các đặc trưng tối ưu để phân tích các chi tiết y khoa vi tế trên ảnh X-quang.

5.3. Đề xuất Hướng phát triển trong Tương lai

Dựa trên các nhận định trên, các hướng phát triển sau được đề xuất để cải thiện hiệu năng của hệ thống:

- **Làm giàu và Tăng cường Dữ liệu:** Đây là hướng đi quan trọng nhất.
    - **Thu thập thêm dữ liệu:** Ưu tiên hàng đầu là thu thập thêm các mẫu bệnh án, đặc biệt là các ca được chẩn đoán là "Bệnh".
    - **Áp dụng Kỹ thuật Augmentation và Oversampling:** Sử dụng các kỹ thuật như SMOTE (Synthetic Minority Over-sampling Technique) trên dữ liệu bảng và các kỹ thuật tăng cường ảnh (xoay, lật, thay đổi độ sáng) trên dữ liệu ảnh để tạo ra các mẫu huấn luyện mới một cách nhân tạo, giúp mô hình học được các quy luật đa dạng hơn.

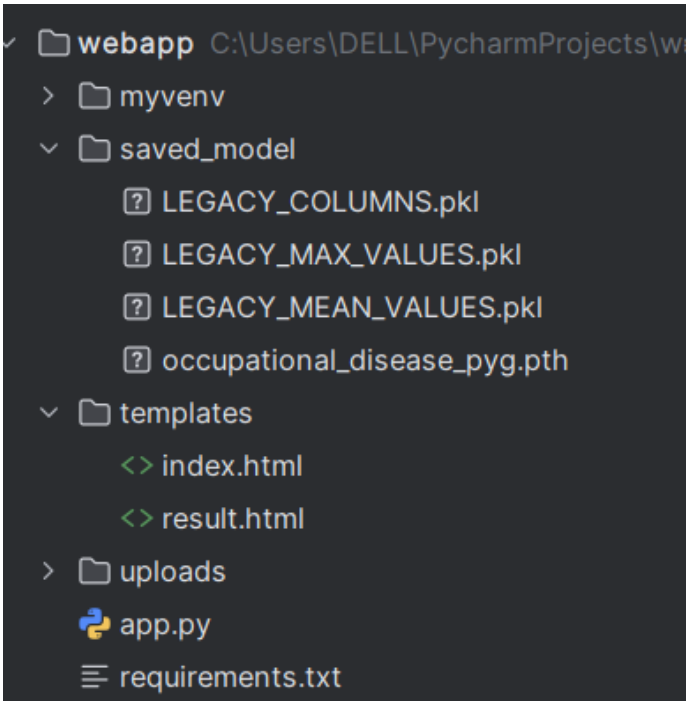
- **Sử dụng các Mô hình được Huấn luyện chuyên biệt:**
    - **Transfer Learning với Mô hình Y tế:** Thay thế MobileNetV2 bằng các mô hình CNN đã được huấn luyện trước trên các bộ dữ

liệu X-quang lớn như CheXpert hoặc MIMIC-CXR. Các mô hình này đã học được cách "nhìn" các đặc trưng y khoa và có khả năng cung cấp các vector đặc trưng chất lượng hơn.

- **Tinh chỉnh và Kỹ thuật Đặc trưng:**
    - **Làm việc với Chuyên gia:** Tham vấn ý kiến của các chuyên gia y tế để xác định và có thể tạo ra các đặc trưng (feature engineering) mới từ dữ liệu bảng có khả năng phân loại cao hơn.
    - **Tinh chỉnh Siêu tham số:** Thử nghiệm với các kiến trúc GCN khác nhau (ví dụ: GAT), các hàm tối ưu và tốc độ học khác nhau để tìm ra cấu hình phù hợp nhất cho bộ dữ liệu.

6. XÂY DỰNG WEB APP

Dựa trên mô hình occupational_disease_pyg.pth và các tệp tiền xử lý đã tạo, Em sẽ xây dựng một trang web đơn giản bằng Flask:



Hệ thống Hỗ trợ Chẩn đoán Bệnh Bụi phổi Silic

Vui lòng nhập đầy đủ thông tin của bệnh nhân và tải lên ảnh X-quang phổi để nhận dự đoán.

1. Tải lên Ảnh X-quang Phổi (*.jpg, *.png)

Choose File | No file chosen

---

2. Nhập Thông tin Bệnh án

tinh

Nhập giá trị cho tinh

gioitinh

Nhập giá trị cho gioitinh

namsinh

Nhập giá trị cho namsinh

cviec

Nhập giá trị cho cviec

pxuong

Nhập giá trị cho pxuong

tuoinghe

Nhập giá trị cho tuoinghe

nampx

Nhập giá trị cho nampx

cv5nam

Nhập giá trị cho cv5nam

cviec1

Nhập giá trị cho cviec1

tgian1

Nhập giá trị cho tgian1

cviec2

Nhập giá trị cho cviec2

tgian2

Nhập giá trị cho tgian2

cao

Nhập giá trị cho cao

can

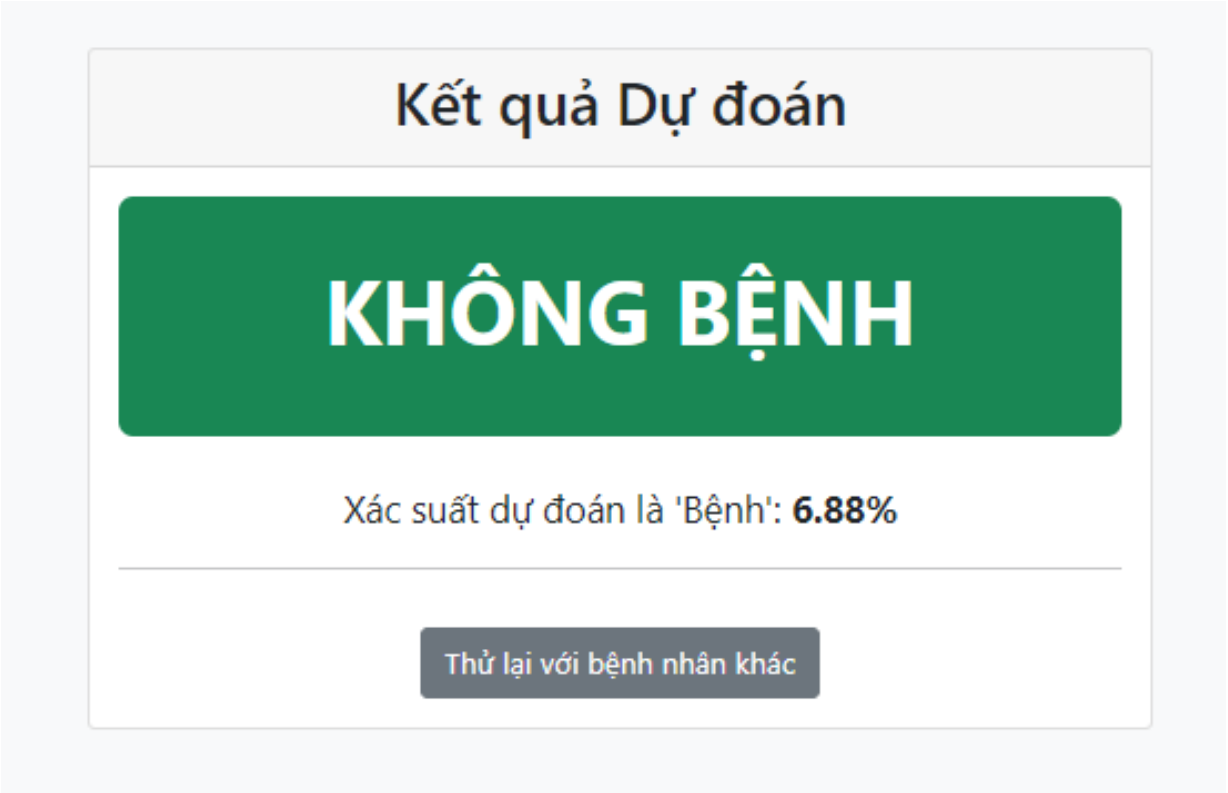
Nhập giá trị cho can

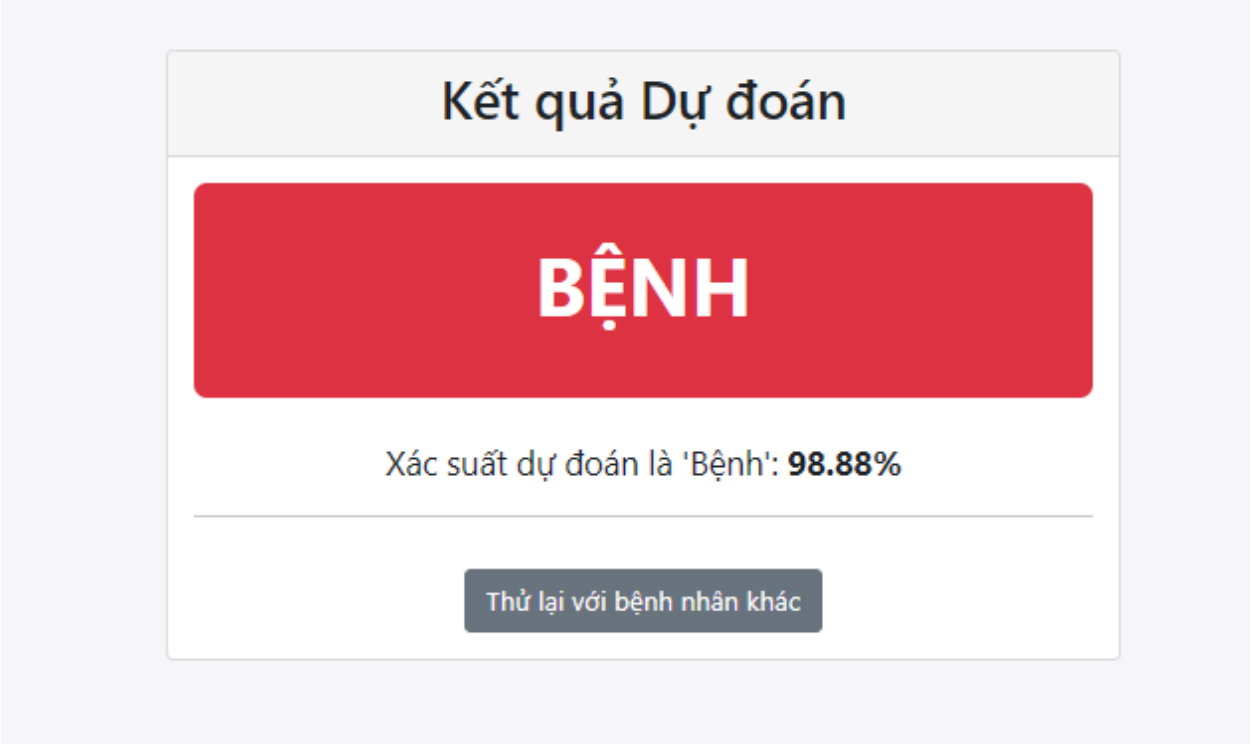
hatd

Nhập giá trị cho hatd

hatt

Nhập giá trị cho hatt





LINK PROJECT HOÀN THIỆN: https://github.com/CosmicAlpaca/webapp.git

7. TÀI LIỆU THAM KHẢO

1. https://github.com/Foacto/OccupationalDiseaseWebAPI
2. Khanh Nguyen-Trong1, Tuan Vu-Van2, Phuong Luong Thi Bich K Nguyen-Trong Paper_128-Graph_Convolutional_Network_for_Occupational_Disease_Prediction_2- Graph Convolutional Network for Occupational Disease Prediction with Multiple Dimensional Data
3. https://pytorch-geometric.readthedocs.io/
4. https://docs.pytorch.org/docs/
5. S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, and D. Rueckert, "Spectral graph convolutions for population-based dis ease prediction," in Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. Springer, 2017, pp. 177–185.
6. Thomas N. Kipf, Max Welling ICLR 2017 Semi-Supervised Classification with Graph Convolutional Networks.
7. C. Data61, "Stellargraph machine learning library," https://github.com/stellargraph/stellargraph, 2018.
8. Ruoyu Li, Sheng Wang, Feiyun Zhu, Junzhou Huang AAAI 2018 Adaptive Graph Convolutional Neural Networks
9. Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst NIPS 2016 Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering
10. F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6861–6871. [Online]. Available: https://proceedings.mlr.press/v97/wu19e.html

LINK PROJECT HOÀN THIỆN: https://github.com/CosmicAlpaca/webapp.git