VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Đồ án tổng hợp - Hướng trí tuệ nhân tạo

Course Project

"EHR Transfer Learning"

Instructor(s): Nguyễn Tuấn Khôi
Ngô Hoàng Anh

Students: Bùi Đăng Khoa - 2252344
Trần Thành Trọng - 2353238

HO CHI MINH CITY, September 2025

Lời cảm ơn

Chúng em xin gửi lời cảm ơn chân thành đến các thầy Nguyễn Tuấn Khôi, thầy Ngô Hoàng Anh. Các thầy đã trực tiếp hướng dẫn, cung cấp tài liệu và ủng hộ chúng em trong suốt quá trình thực hiện đồ án này. Sự chỉ bảo, hướng dẫn và những kiến thức chuyên môn của các thầy đã giúp chúng em hoàn thiện bài báo cáo với hiểu biết sâu sắc hơn về môn học. Chúng em rất trân trọng những gì Thầy đã đóng góp và hy vọng sẽ tiếp tục nhận được sự hướng dẫn quý báu từ Thầy trong các dự án tương lai. Kính chúc các Thầy luôn mạnh khỏe, hạnh phúc và gặt hái nhiều thành công hơn nữa trong sự nghiệp giảng dạy và nghiên cứu.

Member list & Workload

No.	Fullname	Student ID	Problems	% done
1	Bùi Đăng Khoa	2252344	- Random Over-Sampling (ROS)/XGBClassifier	100%
2	Trần Thành Trọng	2353238	- SMOTE và XGBClassifier/XGBRegressor	100%

Mục lục

1 Tóm tắt

Báo cáo này xây dựng hệ thống chẩn đoán bệnh hô hấp mạn tính (CRD) bằng cách kết hợp đặc trưng hình ảnh X-quang ngực (CXR) trích xuất từ mô hình DenseNet theo hướng học chuyển giao và các đặc trưng lâm sàng từ hồ sơ sức khỏe điện tử (EHR). Để xử lý mất cân bằng lớp, hai chiến lược tăng cường dữ liệu ở mức đặc trưng được khảo sát gồm SMOTE và Random Over-Sampling (ROS), kết hợp với các mô hình dựa trên XGBoost. Kết quả cho thấy SMOTE giúp tăng đáng kể Recall khi chỉ dùng đặc trưng ảnh, trong khi ROS kết hợp XGBClassifier cho hiệu quả rõ rệt hơn khi tích hợp thêm EHR, cải thiện Precision/Recall ở nhiều bệnh hiếm.

2 Giới thiệu

2.1 Bối cảnh và các nghiên cứu liên quan

Bệnh hô hấp mạn tính (Chronic Respiratory Diseases – CRDs) là một trong những nguyên nhân hàng đầu gây bệnh tật và tử vong trên toàn cầu [1], đặc biệt tại các khu vực có mức độ ô nhiễm không khí cao hoặc phơi nhiễm lâu dài với khói thuốc lá [2]. Việc phát hiện sớm CRD đóng vai trò quan trọng trong việc ngăn ngừa tiến triển bệnh, cải thiện chất lượng sống của bệnh nhân và giảm gánh nặng chi phí y tế [3]. Chụp X-quang ngực (Chest X-ray – CXR) là phương pháp chẩn đoán không xâm lấn, phổ biến, cho phép bác sĩ đánh giá các bất thường cấu trúc trong phổi [4]. Tuy nhiên, việc đọc và phân tích CXR đòi hỏi trình độ chuyên môn cao, trong khi nhiều dấu hiệu bệnh lý mang tính tinh vi, dễ gây nhầm lẫn, dẫn đến sai sót trong chẩn đoán [5].

Trong những năm gần đây, học máy (ML), đặc biệt là học sâu (DL), đã được ứng dụng rộng rãi trong chẩn đoán bệnh từ ảnh CXR. Các mạng nơ-ron tích chập (CNN) cho thấy khả năng mạnh mẽ trong việc trích xuất đặc trưng hình ảnh phức tạp và đạt độ chính xác cao trong nhiều bài toán y tế [6]. Tuy nhiên, các mô hình DL vẫn tồn tại những hạn chế quan trọng. Thứ nhất, quá trình ra quyết định của mô hình thiếu tính minh bạch, gây khó khăn cho việc triển khai trong môi trường lâm sàng. Các kỹ thuật trực quan hóa như heatmap chỉ cung cấp thông tin định tính và chưa phản ánh đầy đủ cơ sở của dự đoán [7]. Thứ hai, dữ liệu CRD thường bị mất cân bằng nghiêm trọng, khi số lượng mẫu dương tính của nhiều bệnh hiếm rất hạn chế, làm suy giảm đáng kể hiệu suất của mô hình đối với lớp thiểu số [8].

Hồ sơ sức khỏe điện tử (Electronic Health Records – EHR) chứa các thông tin lâm sàng quan trọng như nhân khẩu học, chỉ số sinh tồn, bệnh nền và tiền sử điều trị, có giá trị bổ trợ cho chẩn đoán CRD. Các mô hình ML truyền thống trên EHR có ưu điểm về khả năng giải thích và đánh giá tầm quan trọng của đặc trưng [9]. Tuy nhiên, EHR thường bị thiếu dữ liệu và phân bố thưa thớt, khiến việc tích hợp trực tiếp với DL gặp nhiều khó khăn. Đồng thời, vấn đề mất cân bằng lớp vẫn là thách thức lớn khi chỉ sử dụng EHR.

Để giải quyết các hạn chế trên, dự án này đề xuất một khuôn khổ dự đoán CRD đa chiến lược, kết hợp dữ liệu hình ảnh CXR và dữ liệu EHR. Học chuyển giao được sử dụng để trích xuất đặc trưng hình ảnh từ các mô hình CNN tiền huấn luyện, trong khi các mô hình ML truyền thống khai thác thông tin lâm sàng có cấu trúc. Nhằm giảm thiểu mất cân bằng lớp và cải thiện khả năng khái quát, các kỹ thuật tăng cường dữ liệu ở mức đặc trưng, bao gồm SMOTE [10] và Gaussian Noise, được áp dụng. Các chiến lược này giúp tăng cường biểu diễn của lớp thiểu số và hạn chế ảnh hưởng của dữ liệu thiếu.

Đóng góp chính của dự án bao gồm:

1. Đề xuất một hệ thống chẩn đoán CRD đa mô thức, kết hợp đặc trưng CXR và EHR.
2. Áp dụng chiến lược tăng cường dữ liệu đa dạng nhằm xử lý mất cân bằng lớp.
3. Đánh giá toàn diện hiệu suất mô hình và phân tích hành vi ra quyết định để nâng cao tính minh bạch.

3 Phương pháp

3.1 Lựa chọn thiết kế

3.1.1 Nguồn dữ liệu

Dự án trong báo cáo này sử dụng hai nguồn dữ liệu: Medical Information Mart for Intensive Care (MIMIC) và CheXpert.

MIMIC-IV [11, 12]: chứa 299.712 hồ sơ chi tiết về các lần thăm khám của bệnh nhân tại Beth Israel Deaconess Medical Center (BIDMC). Cơ sở dữ liệu này được sử dụng để trích xuất EHR của bệnh nhân, cũng như làm ground truth cho nhiệm vụ này: mã International Classification of Diseases (ICD), ghi lại các bệnh được phát hiện khi bệnh nhân xuất viện.

MIMIC-CXR [13]: bao gồm 227.835 nghiên cứu hình ảnh đã được ẩn danh của 64.588 bệnh nhân tại BIDMC và MIMIC-IV. Cơ sở dữ liệu này được sử dụng để truy xuất hình ảnh CXR của bệnh nhân.

CheXpert [14]: là một bộ dữ liệu CXR khác được sử dụng trong dự án. Nó chứa 224.316 hình ảnh CXR của 65.240 bệnh nhân tại Stanford Hospital, hoàn toàn độc lập với BIDMC. Trong dự án này, bộ dữ liệu được dùng để pre-train mô hình DL đã chọn, nhằm thực hiện Transfer Learning trên dữ liệu CXR từ MIMIC. Việc này đảm bảo tính tổng quát của mô hình đối với dữ liệu thực tế, khi các trường hợp mới chưa được biết trước bởi mô hình đã huấn luyện.

Data access control: nhằm tuân thủ các khía cạnh đạo đức liên quan đến dữ liệu y tế, các tác giả đã hoàn thành tất cả các khóa đào tạo cần thiết để truy cập cơ sở dữ liệu. Để đảm bảo an toàn, dữ liệu sử dụng trong dự án này sẽ không được cung cấp công khai. Thay vào đó, các script lập trình được cung cấp, trong đó yêu cầu đăng nhập để xác nhận trạng thái đào tạo trước khi truy cập và tạo bộ dữ liệu.

3.1.2 Hình thành dữ liệu

Các CRD cần dự đoán: do có quá nhiều CRD, việc bao quát tất cả là không khả thi, vì vậy chúng ta sẽ chọn 8 bệnh tiêu biểu, được kỳ vọng có sự đa dạng về đặc điểm và phân bố dữ liệu [15, 16, 17, 18, 19]. Bảng tóm tắt các bệnh được trình bày trong Figure 1. Bệnh nhân được xác định đã mắc CRD dựa trên mô tả mã ICD được cấp khi xuất viện, lấy trực tiếp từ MIMIC-IV.

Các đặc trưng EHR: dựa trên 8 CRD đã chọn, 12 đặc trưng EHR được lựa chọn để trích xuất từ MIMIC-IV. Đối với các đặc trưng này, chúng ta loại bỏ tất cả thông tin về thời gian nhằm giảm rò rỉ thông tin bệnh nhân [15, 16, 17, 18, 19].

Các đặc trưng hình ảnh DL: DenseNet là kiến trúc mạng nơ-ron sâu có khả năng duy trì thông tin ở nhiều mức chi tiết khác nhau [20]. Mô hình này được chọn để trích xuất các chi tiết quan trọng nhất từ hình ảnh CXR. Dự án này sẽ triển khai mô hình CheXpert đã được pre-train sẵn, có sẵn trong gói torchxrayvision [21]. Lớp dense thứ hai tính từ cuối sẽ được chọn để trích xuất đặc trưng, tạo ra 18 đặc trưng tiềm ẩn từ mỗi hình ảnh CXR.

Bệnh	Keywords ICD	Số ca bệnh	% Tỷ lệ của từng bệnh
arthritis	arthritis	1191	4.10%
bronchitis	bronchitis	1485	5.1%
fracture	fracture, broken	2078	7.1%
lung_cancer	tumor, cancer	7787	26.8%
lung_infection	infection, infectious	952	3.2%
pneumonia	pneumonia	7980	27.4%
scoliosis	scoliosis, curvature	150	0.5%
tuberculosis	tuberculosis	771	2.7%

Figure 1: Thống kê số ca và tỷ lệ mắc các bệnh CRD trong tập dữ liệu

3.2 Thu thập và tổ chức dữ liệu

3.2.1 Xác định nhóm dự án

Các bệnh nhân trong MIMIC-IV sẽ bị loại trừ nếu họ không có CXR. Những bệnh nhân còn lại sẽ được kiểm tra mã ICD để xác định xem có khớp với một trong 8 CRD đã chọn hay không, từ đó chia bệnh nhân thành hai nhóm. Tất cả bệnh nhân có mã ICD khớp sẽ được đưa vào nhóm dự án, trong khi những bệnh nhân còn lại, có số lượng lớn hơn đáng kể, sẽ được chọn ngẫu nhiên với số lượng bằng nhóm đầu tiên. Quá trình này được tóm tắt trong Figure 2C.

Tiếp theo, 11 đặc trưng EHR sẽ được trích xuất từ cơ sở dữ liệu MIMIC-IV, như minh họa trong Figure 2A. Mỗi bệnh nhân trong nhóm dự án được chọn sẽ có CXR gần nhất trước khi xuất viện được lựa chọn. Việc này giúp đơn giản hóa mô hình và tránh trùng lặp dữ liệu, vì một bệnh nhân có thể có nhiều kết quả CXR trong MIMIC [13]. Trong tập dữ liệu cuối cùng của nhóm dự án, mỗi bệnh nhân dự kiến có 18 đặc trưng hình ảnh và 12 đặc trưng EHR, tổng cộng 30 đặc trưng.

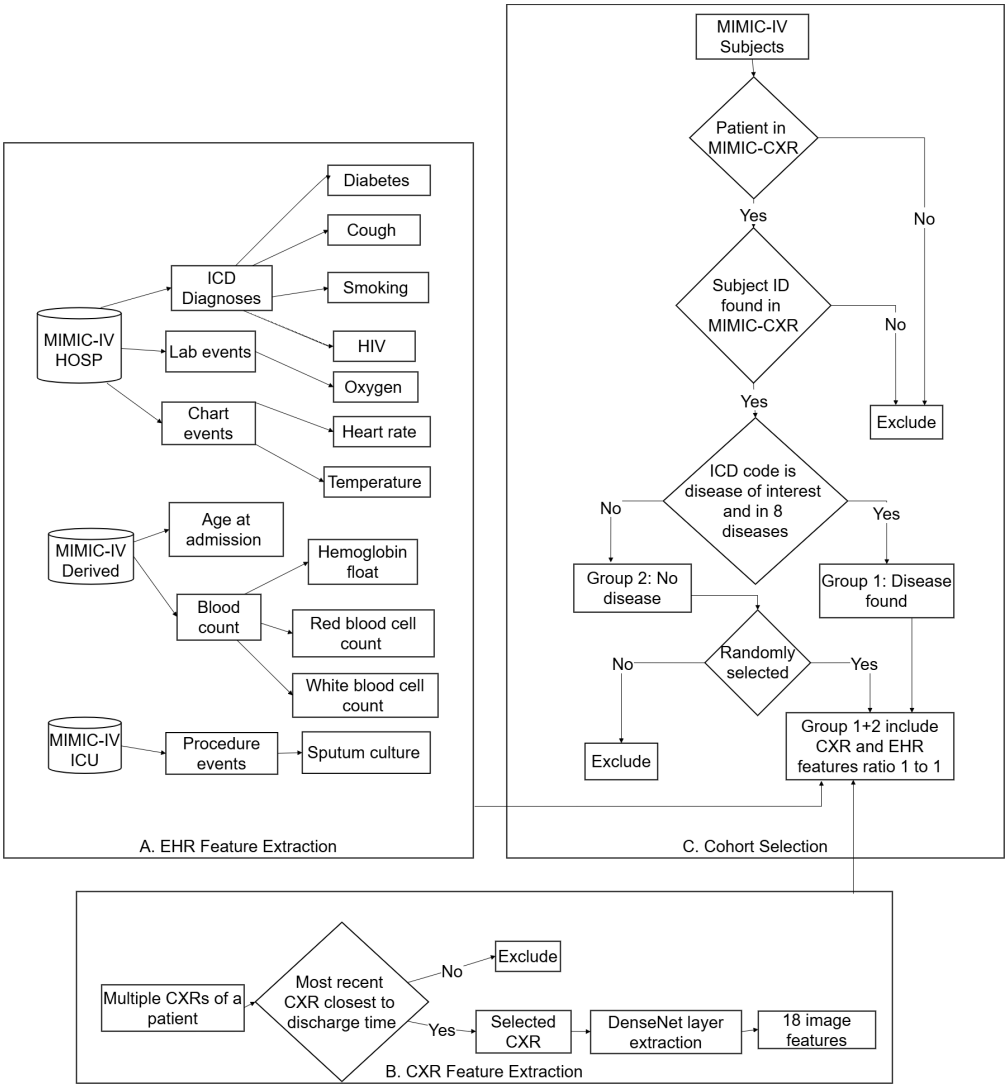


Figure 2: Sơ lược quy trình trích xuất nhóm đối tượng nghiên cứu

3.2.2 Làm sạch dữ liệu

Trong khi các đặc trưng hình ảnh không cần xử lý thêm do giao thức dữ liệu và việc làm sạch trong MIMIC-CXR đã được thực hiện đầy đủ [13], thì dữ liệu EHR lại chứa một số phép đo bất thường, chẳng hạn như nhiệt độ cơ thể là 150°C. Do đó, để đảm bảo tính toàn vẹn của dữ liệu đối với các đặc trưng liên tục, chúng ta đã loại bỏ bất kỳ phép đo nào là outlier trong phân bố đặc trưng, tức là những giá trị cách trung bình hơn 3 độ lệch chuẩn.

3.3 Quy trình xử lý

Quy trình xử lý của dự án được chia thành hai giai đoạn chính: Tăng cường dữ liệu (Data Augmentation) và Đánh giá (Evaluation). Mỗi mô hình được huấn luyện và đánh giá trên tập dữ liệu đã được tăng cường nhằm phân tích mức độ cải thiện hiệu suất.

Các chỉ số đánh giá: Để đánh giá mô hình, chúng ta chọn độ chính xác (accuracy), độ chính xác dự đoán (precision) và khả năng phát hiện (recall) làm các chỉ số hiệu suất. Những chỉ số này hữu ích để đánh giá khả năng cân bằng giữa dự đoán chính xác và giảm bỏ sót các mẫu thuộc lớp thiểu số, đặc biệt khi các trường hợp âm tính giả (False Negatives) được coi là ít mong muốn do hậu quả nghiêm trọng hơn.

3.3.1 Giai đoạn 1: Tăng cường dữ liệu

3.3.1.1 Phương pháp 1: SMOTE và XGBClassifier/XGBRegressor

Trong tập dữ liệu, số lượng nhãn âm tính và dương tính chênh lệch lớn, ảnh hưởng đến khả năng dự đoán. SMOTE [10] được áp dụng để tạo mẫu nhân tạo cho lớp thiểu số, giúp cân bằng dữ liệu và cải thiện khả năng học của mô hình. Sau đó, XGBoost [22] được huấn luyện trên dữ liệu tăng cường, đồng thời xử lý tốt mất cân bằng lớp và các giá trị thiếu.

SMOTE cho đặc trưng ảnh: SMOTE được áp dụng trên 18 đặc trưng ảnh từ các hình chụp X-quang. Mục đích là đảm bảo rằng số lượng mẫu dương tính và âm tính trong tập huấn luyện gần bằng nhau, giúp mô hình học được các đặc trưng của bệnh một cách chính xác hơn. Sau khi tăng cường, các mẫu hình ảnh mới được kết hợp với dữ liệu EHR còn thiếu để tạo ra tập dữ liệu đầy đủ, sẵn sàng cho bước huấn luyện mô hình.

XGBClassifier và XGBRegressor cho đặc trưng EHR: Dữ liệu EHR bao gồm cả các đặc trưng phân loại (categorical: diabetes, HIV, smoke, cough) và các đặc trưng số (numeric: age, oxygen, heart_rate, temperature, hemoglobin, rbc, wbc). Để dự đoán các đặc trưng này:

- XGBClassifier được dùng cho các đặc trưng phân loại. Nó giúp mô hình xác định giá trị nhãn (0 hoặc 1) dựa trên các đặc trưng của bệnh nhân (Figure 3.B).
- XGBRegressor được dùng cho các đặc trưng số (Figure 3.D). Trước đó, mô hình sẽ dự đoán mask (Figure 3.C) bằng XGBClassifier để cho biết giá trị đặc trưng có tồn tại hay không, sau đó XGBRegressor dự đoán giá trị thực tế của các đặc trưng số nếu mask = 1 (Figure 3.E).

Sự kết hợp giữa SMOTE, XGBClassifier và XGBRegressor giúp đảm bảo rằng tập dữ liệu đầu vào cho mô hình XGBoost cuối cùng là cân bằng, đầy đủ và phản ánh cả thông tin hình ảnh lẫn thông tin EHR, từ đó nâng cao hiệu suất dự đoán CRD.

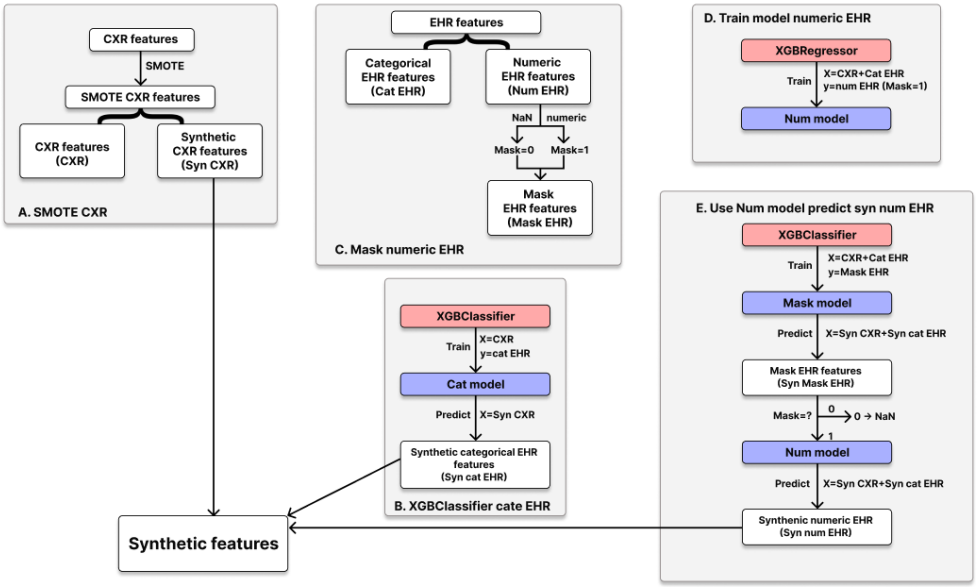


Figure 3: Quy trình tăng cường dữ liệu cho 2 đặc trưng CXR và EHR

3.3.1.2 Phương pháp 2: Random Over-Sampling (ROS) + XGBClassifier

Bên cạnh SMOTE, chúng tôi triển khai *Random Over-Sampling* (ROS)[23] như một phương pháp cân bằng lớp đơn giản cho bài toán mất cân bằng nhãn trong dữ liệu CRD. Khác với SMOTE (tạo mẫu tổng hợp mới), ROS cân bằng dữ liệu bằng cách **lặp lại ngẫu nhiên các mẫu thuộc lớp thiểu số** trong tập huấn luyện cho đến khi số lượng hai lớp trở nên cân bằng hơn. Phương pháp này giúp giảm thiên lệch của mô hình về lớp âm tính và tăng khả năng học được đặc trưng của lớp dương tính (các ca bệnh hiếm).

Bước 1: Với mỗi bệnh trong 8 CRD, thực hiện chia dữ liệu thành tập huấn luyện và tập kiểm thử (*train–test split*) trước khi tăng cường dữ liệu. Việc này đảm bảo **tránh rò rỉ dữ liệu** (data leakage), vì nếu oversampling trước khi chia thì các mẫu bị lặp có thể xuất hiện ở cả train và test.

Bước 2: Chỉ áp dụng ROS trên tập train để tạo tập huấn luyện cân bằng hơn giữa lớp 0 (không bệnh) và lớp 1 (có bệnh). Tập test được giữ nguyên nhằm phản ánh đúng phân bố dữ liệu thực tế khi đánh giá.

Bước 3: Huấn luyện mô hình phân loại trên dữ liệu đã được cân bằng bởi ROS bằng thuật toán `XGBClassifier` [24]. Thí nghiệm được thực hiện theo hai cấu hình đặc trưng: (i) chỉ sử dụng đặc trưng ảnh (*image features only*), và (ii) kết hợp đặc trưng ảnh và EHR (*image features with EHR features*).

ROS được xem như một baseline hữu ích vì: (i) không tạo mẫu tổng hợp mới (giảm nguy cơ sinh dữ liệu không hợp lý), và (ii) khi EHR có missing values, việc lặp mẫu giúp giữ nguyên cấu trúc thiếu dữ liệu của mẫu gốc, phù hợp với đặc thù dữ liệu lâm sàng[25].

3.3.2 Giai đoạn 2: Đánh giá

3.3.2.1 Mô hình 1: XGBoost

Trong giai đoạn đánh giá, mô hình XGBoost [22] được sử dụng để phân tích hiệu suất dự đoán các bệnh CRD trên tập dữ liệu đã qua tăng cường và không tăng cường dữ liệu. Mô hình được huấn luyện và đánh giá trong hai kịch bản: chỉ sử dụng đặc trưng hình ảnh và kết hợp đặc trưng hình ảnh với dữ liệu EHR, nhằm đánh giá tác động của từng nguồn dữ liệu đến hiệu suất dự đoán.

3.3.2.2 Mô hình 2: XGBClassifier (ROS)

Ngoài mô hình XGBoost ở Mô hình 1, chúng tôi xây dựng **Mô hình 2** dựa trên `XGBClassifier`. như một phương án đối sánh, đặc biệt để đánh giá tác động của chiến lược cân bằng lớp bằng ROS. `XGBClassifier` là mô hình boosting trên cây quyết định, phù hợp cho dữ liệu bảng và có khả năng mô hình hóa quan hệ phi tuyến giữa các đặc trưng, thường cho hiệu quả tốt trong bài toán phân loại trong lĩnh vực y sinh học [24].

Trong giai đoạn đánh giá, với mỗi bệnh trong 8 CRD, mô hình được huấn luyện và kiểm thử theo hai cấu hình đầu vào:

- **Image features only:** chỉ sử dụng 18 đặc trưng trích xuất từ ảnh CXR.
- **Image features with EHR features:** kết hợp 18 đặc trưng CXR và 12 đặc trưng EHR .

Với mỗi cấu hình đầu vào, mô hình được chạy ở hai thiết lập dữ liệu: (i) **baseline** (không ROS), và (ii) **ROS** (chỉ áp dụng trên tập train). Các chỉ số đánh giá sử dụng gồm Accuracy, Precision và Recall nhằm phản ánh đồng thời độ chính xác tổng thể và khả năng phát hiện lớp thiểu số.

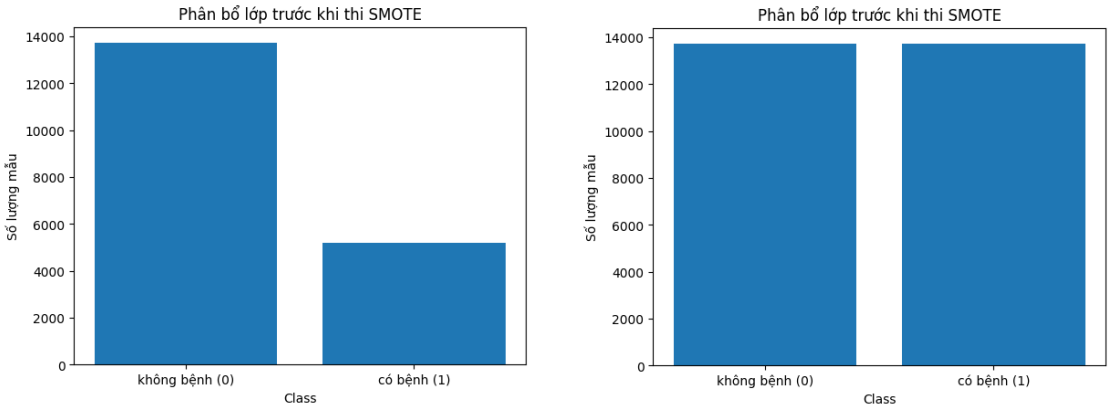
4 Kết quả

4.1 Giai đoạn 1: Tăng cường dữ liệu

4.1.1 Phương pháp 1: SMOTE và XGBClassifier/XGBRegressor

Với 8 bệnh được nghiên cứu, mỗi bệnh sử dụng một phép chia dữ liệu train–test độc lập, tuy nhiên quy trình tăng cường dữ liệu bằng SMOTE và dự đoán đặc trưng EHR là hoàn toàn giống nhau. Do đó, trong phần kết quả, chúng ta chỉ trình bày một ví dụ đại diện nhằm minh họa hiệu

quả của giai đoạn tăng cường dữ liệu, đồng thời cung cấp thống kê tổng quát cho toàn bộ các bệnh để đảm bảo tính nhất quán của phương pháp.



(a) Trước khi áp dụng SMOTE (b) Sau khi áp dụng SMOTE

Figure 4: Phân bố dữ liệu CXR và EHR trước và sau khi tăng cường dữ liệu

Sau khi áp dụng SMOTE trên các đặc trưng hình ảnh CXR, số lượng mẫu thuộc lớp dương tính đã được tăng đáng kể. Cụ thể, số mẫu dương tính được tạo thêm là 8,533 mẫu, giúp phân bố dữ liệu giữa hai lớp trở nên cân bằng hơn (Figure 4). Sự cân bằng này đóng vai trò quan trọng trong việc giảm thiên lệch của mô hình về lớp âm tính và tạo điều kiện thuận lợi cho quá trình học các đặc trưng liên quan đến bệnh.

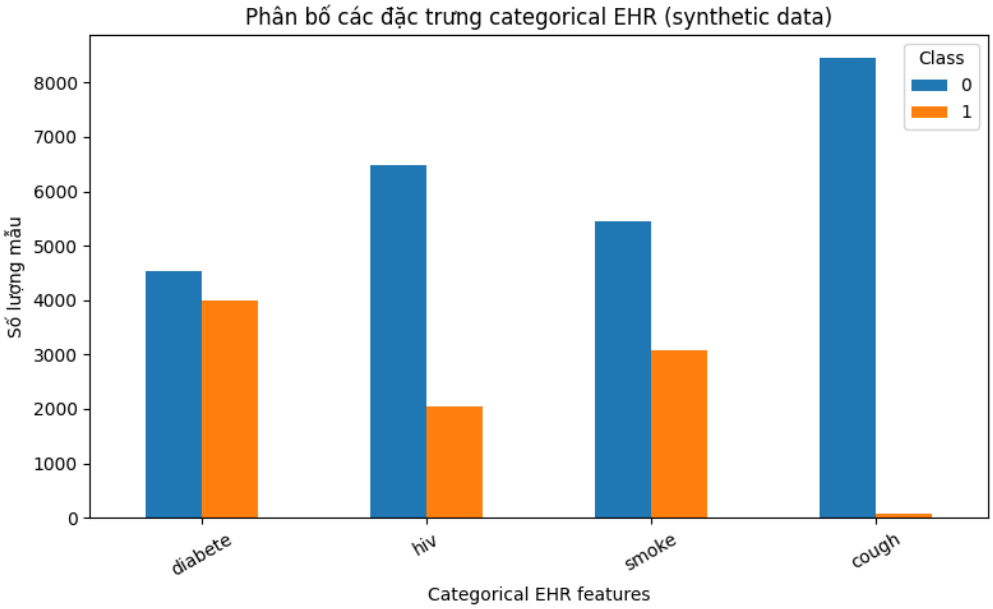


Figure 5: Phân bố các đặc trưng EHR dạng phân loại tại các mẫu CXR được tạo bởi SMOTE

Figure 5 minh họa phân bố của các đặc trưng EHR dạng phân loại (categorical) tại các mẫu CXR tổng hợp được tạo bởi SMOTE. Có thể quan sát rằng các giá trị nhị phân (0/1) của các đặc trưng như diabetes, HIV, smoke và cough không còn bị thiếu và được phân bổ một cách nhất quán trên toàn bộ tập dữ liệu tổng hợp. Điều này cho thấy mô hình XGBClassifier đã học được mối quan hệ giữa các đặc trưng hình ảnh và các biến lâm sàng phân loại, từ đó gán nhãn hợp lý cho các mẫu tổng hợp mà không phá vỡ phân bố dữ liệu ban đầu.

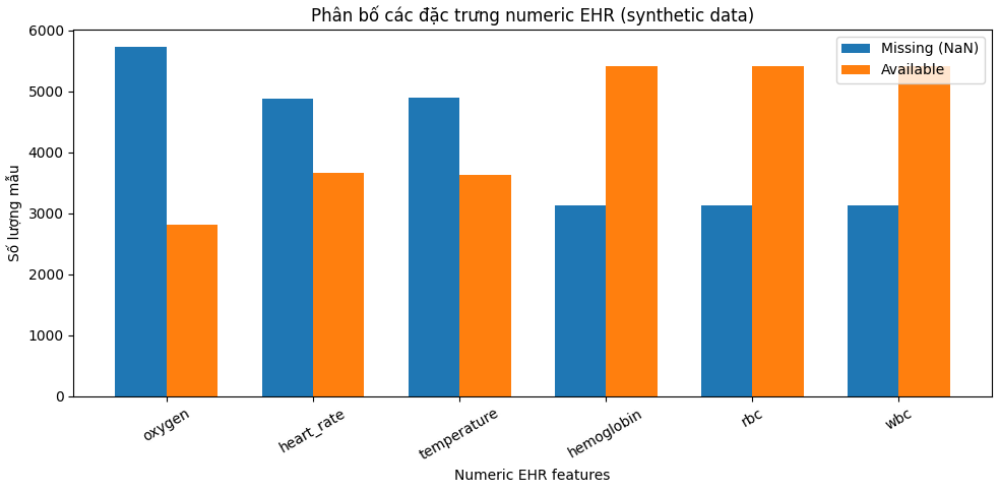


Figure 6: Phân bố các đặc trưng EHR dạng số tại các mẫu CXR được tạo bởi SMOTE

Đối với các đặc trưng EHR dạng số, Figure 6 thể hiện phân bố trạng thái tồn tại dữ liệu (available) và thiếu dữ liệu (missing/NaN) tại các mẫu CXR được tạo bởi SMOTE. Kết quả cho thấy mô hình XGBClassifier dự đoán mask đã giúp xác định hợp lý sự tồn tại của các đặc trưng số, trong khi XGBRegressor chỉ thực hiện dự đoán giá trị tại các vị trí mà mask cho phép. Cách tiếp cận hai bước này giúp duy trì cấu trúc thiếu dữ liệu vốn có của EHR, đồng thời tránh việc sinh dữ liệu giả không hợp lý về mặt lâm sàng.

Nhìn chung, các kết quả trên cho thấy quy trình kết hợp giữa SMOTE cho đặc trưng hình ảnh và XGBClassifier/XGBRegressor cho đặc trưng EHR đã tạo ra một tập dữ liệu tăng cường vừa cân bằng về nhãn bệnh, vừa đầy đủ về thông tin đa nguồn. Tập dữ liệu này đóng vai trò đầu vào nhất quán và đáng tin cậy cho giai đoạn đánh giá hiệu suất mô hình ở phần tiếp theo.

4.1.2 Phương pháp 2: Random Over-Sampling (ROS)

Tương tự SMOTE, ROS được sử dụng nhằm giải quyết vấn đề mất cân bằng lớp trong các bài toán CRD. Tuy nhiên, thay vì tạo mẫu tổng hợp, ROS cân bằng dữ liệu bằng cách lặp lại ngẫu nhiên các mẫu thuộc lớp dương tính trong tập huấn luyện. Vì vậy, để đánh giá hiệu quả của giai đoạn tăng cường dữ liệu, chúng tôi tập trung so sánh **phân bố nhãn trước và sau ROS** trên tập train.

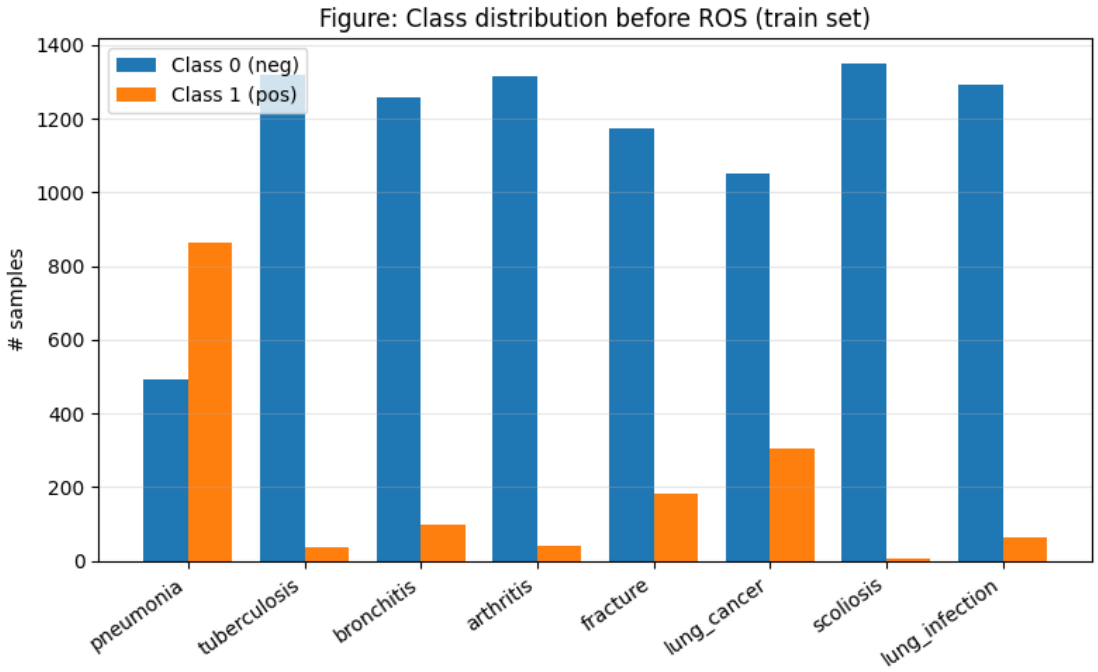


Figure 7: Trước khi áp dụng ROS.

Figure 7 cho thấy dữ liệu huấn luyện trước khi áp dụng ROS bị **mất cân bằng lớp rất rõ rệt** ở hầu hết các bệnh. Cụ thể, số mẫu thuộc lớp âm tính (Class 0 – neg) thường ở mức khoảng hơn 1,000 mẫu mỗi bệnh, trong khi lớp dương tính (Class 1 – pos) chỉ chiếm một tỷ lệ rất nhỏ (nhiều bệnh chỉ vài chục đến vài trăm mẫu). Sự chênh lệch này đặc biệt đáng chú ý ở các bệnh như *tuberculosis*, *arthritis* và *scoliosis*, nơi lớp dương tính gần như rất hiếm so với lớp âm tính. Trạng thái này dễ khiến mô hình học bị thiên lệch về dự đoán âm tính, dẫn đến Recall thấp cho lớp dương tính trong giai đoạn đánh giá.

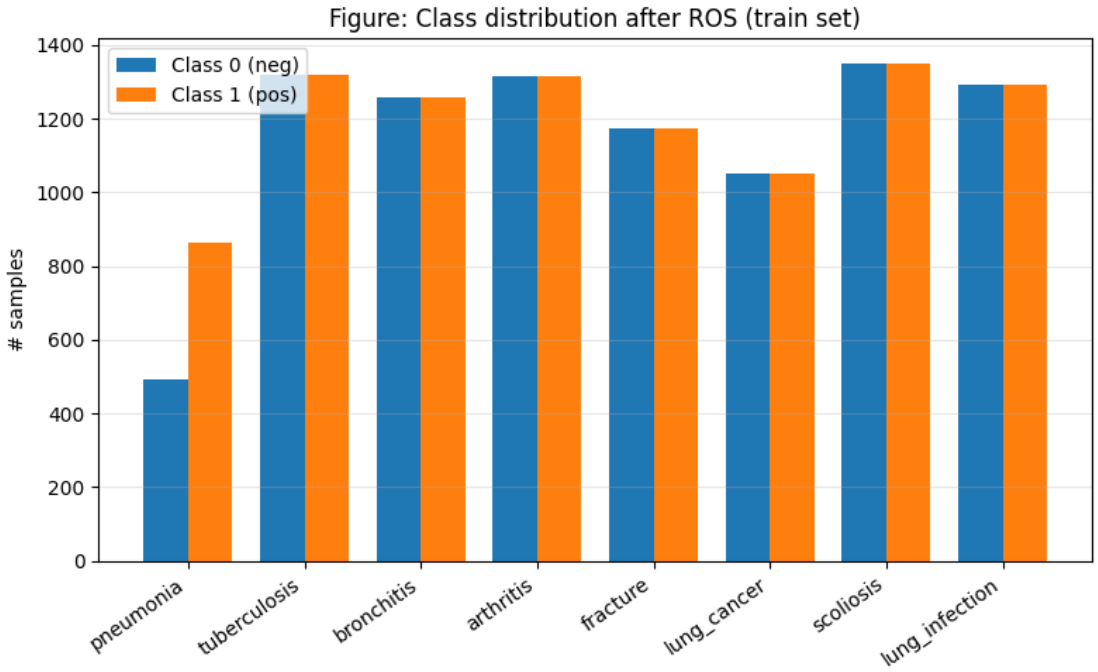


Figure 8: Sau khi áp dụng ROS.

Figure 8 minh họa phân bố dữ liệu sau khi áp dụng ROS với `sampling_strategy=1.0`. Có thể quan sát rằng ROS đã **tăng số lượng mẫu lớp dương tính** bằng cách lặp lại các mẫu dương tính trong tập train, qua đó đưa phân bố hai lớp về trạng thái **xấp xỉ cân bằng** cho phần lớn các bệnh: số mẫu Class 1 (pos) tăng lên gần bằng số mẫu Class 0 (neg). Điều này cho thấy ROS đạt đúng mục tiêu của giai đoạn tăng cường dữ liệu: giảm mức độ mất cân bằng lớp và tạo điều kiện để mô hình nhìn thấy nhiều mẫu dương tính hơn khi huấn luyện, từ đó kỳ vọng cải thiện khả năng phát hiện bệnh (Recall) ở giai đoạn đánh giá.

4.2 Giai đoạn 2: Đánh giá

4.2.1 Mô hình 1: XGBoost

Model & Disease		Image features only			Image features with EHR features		
Model	Disease	Accuracy	Precision	Recall	Accuracy	Precision	Recall
XGBoost	pneumonia	70.4%	33.6%	9.1%	78.5%	70.1%	36.6%
	tuberculosis	97.3%	9.0%	0.1%	97.8%	95.7%	19.8%
	bronchitis	94.8%	19.7%	0.4%	95.2%	77.9%	10.9%
	arthritis	95.9%	2.9%	0.0%	96.5%	93.7%	16.1%
	fracture	92.7%	11.1%	0.3%	93.2%	71.2%	8.9%
	lung cancer	71.0%	33.9%	7.8%	77.2%	66.3%	31.6%
	scoliosis	99.4%	0.0%	0.0%	99.6%	97.5%	18.9%
	lung infection	96.7%	9.7%	0.2%	97.1%	92.2%	11.8%

Table 1: Kết quả đánh giá khi không có Data Augmentation

Model & Disease		Image features only			Image features with EHR features		
Model	Disease	Accuracy	Precision	Recall	Accuracy	Precision	Recall
XGBoost	pneumonia	58.4%	30.5%	42.5%	78.8%	70.0%	37.8%
	tuberculosis	88.5%	3.7%	12.9%	97.8%	96.1%	19.5%
	bronchitis	81.4%	6.7%	20.6%	95.3%	79.4%	10.0%
	arthritis	83.5%	4.9%	16.1%	96.4%	92.6%	14.7%
	fracture	76.5%	8.6%	24.6%	93.3%	69.3%	8.7%
	lung cancer	58.6%	30.6%	43.1%	77.1%	65.4%	30.8%
	scoliosis	97.8%	0.7%	2.3%	99.6%	100.0%	24.8%
	lung infection	86.3%	4.0%	13.3%	96.9%	94.8%	8.8%

Table 2: Kết quả đánh giá khi thêm Data Augmentation

Model & Disease		Image features only			Image with EHR features		
Model	Disease	Accuracy	Precision	Recall	Accuracy	Precision	Recall
XGBoost	pneumonia	-12.0	-3.1	33.4	0.3	-0.1	1.2
	tuberculosis	-8.8	-5.3	12.8	0.0	0.4	-0.3
	bronchitis	-13.4	-13.0	20.2	0.1	1.5	-0.9
	arthritis	-12.4	2.0	16.1	-0.1	-1.1	-1.4
	fracture	-16.2	-2.5	24.3	0.1	-1.9	-0.2
	lung cancer	-12.4	-3.3	35.3	-0.1	-0.9	-0.8
	scoliosis	-1.6	0.7	2.3	0.0	2.5	5.9
	lung infection	-10.4	-5.7	13.1	-0.2	2.6	-3.0

Table 3: Đánh giá độ lệch giữa không Data Augmentation và có Data Augmentation

Kết quả từ Table 1 cho thấy khi chỉ sử dụng các đặc trưng hình ảnh từ ảnh X-quang ngực (CXR), mô hình XGBoost đạt độ chính xác tổng thể rất cao, dao động từ 71% đến gần 100% tùy thuộc vào từng bệnh. Tuy nhiên, các chỉ số Precision và Recall đều tương đối thấp, đặc biệt với những bệnh có sự mất cân bằng lớp nghiêm trọng như tuberculosis, arthritis và scoliosis. Điều này phản ánh xu hướng của mô hình khi học từ dữ liệu mất cân bằng: mô hình thường dự

đoán nhãn âm tính phổ biến hơn, dẫn đến việc bỏ sót nhiều ca dương tính, trong khi các bệnh như pneumonia và lung infection lại có Precision và Recall cao hơn nhờ mức độ mất cân bằng lớp ít nghiêm trọng hơn.

Table 2 minh họa hiệu quả của phương pháp Data Augmentation thông qua SMOTE. Sau khi áp dụng SMOTE trên 18 đặc trưng hình ảnh, Recall của các bệnh như pneumonia tăng đáng kể từ 9,1% lên 42,5%, lung cancer từ 7,8% lên 43,1%. Một số chỉ số Precision giảm nhẹ, nhưng tổng thể, mô hình học được nhiều mẫu dương tính hơn, cho thấy SMOTE giúp giải quyết hiệu quả vấn đề mất cân bằng lớp trong dữ liệu ảnh. Khi kết hợp các đặc trưng EHR vào mô hình, cả Precision và Recall đều được cải thiện rõ rệt, ví dụ pneumonia đạt Precision 70,0% và Recall 37,8%, tuberculosis Precision 96,1% và Recall 19,5%, scoliosis Precision 100% và Recall 24,8%. Điều này chứng tỏ rằng thông tin từ EHR bổ sung giá trị dự đoán, giúp mô hình nhận diện chính xác hơn các ca bệnh dương tính.

Đánh giá sự thay đổi giữa các thiết lập có và không có Data Augmentation (Table 3) cho thấy SMOTE giúp tăng Recall mạnh mẽ cho hầu hết các bệnh dựa trên đặc trưng ảnh, trong khi ảnh hưởng đến Accuracy và Precision là tương đối nhỏ. Ngược lại, khi kết hợp thêm các đặc trưng EHR, sự thay đổi các chỉ số gần như không đáng kể, hoặc chỉ dương nhẹ, cho thấy rằng EHR duy trì hiệu suất ổn định, nhưng không tác động mạnh mẽ đến cải thiện Recall như dữ liệu ảnh đã được tăng cường. Kết quả này nhấn mạnh vai trò quan trọng của việc cân bằng dữ liệu trong không gian đặc trưng hình ảnh đối với việc cải thiện khả năng nhận diện các ca bệnh dương tính, đồng thời cho thấy việc tích hợp EHR hỗ trợ ổn định kết quả mà không tạo ra ảnh hưởng quá lớn, phản ánh sự phối hợp cân bằng giữa hai nguồn dữ liệu trong quá trình học của mô hình XGBoost.

4.2.2 Mô hình 2: XGBClassifier + ROS

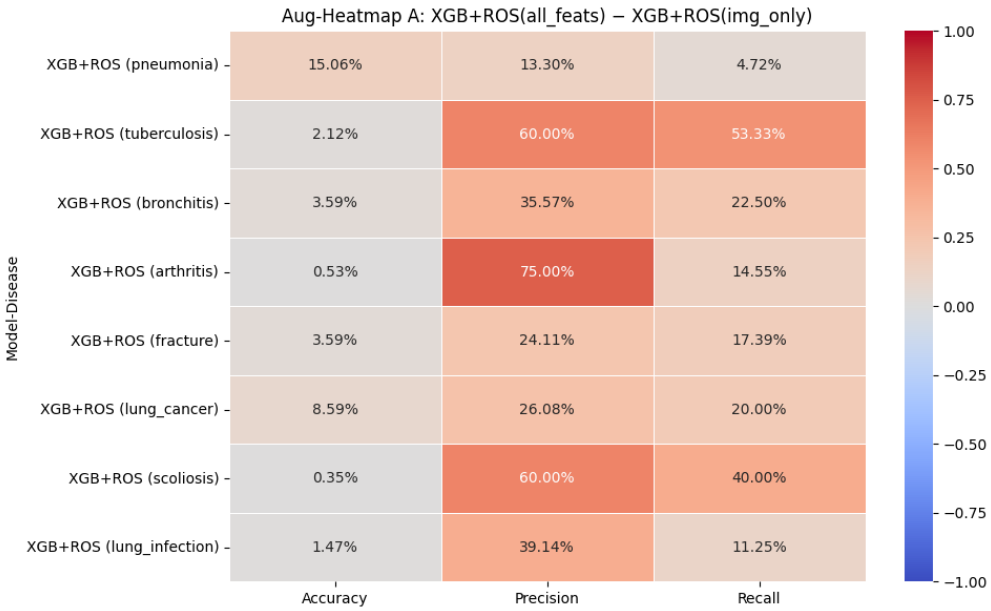


Figure 9: Performance gain of XGB+ROS after incorporating MIMIC-IV EHR features

Figure 9 cho thấy việc **bổ sung đặc trưng EHR (all_feats)** giúp cải thiện rất rõ rệt chất lượng dự đoán so với chỉ dùng đặc trưng ảnh (img_only), đặc biệt ở các chỉ số liên quan trực tiếp đến phát hiện bệnh (Precision/Recall). Cụ thể, *tuberculosis* tăng mạnh Precision (+60.00%) và Recall (+53.33%), *scoliosis* tăng Precision (+60.00%) và Recall (+40.00%), *arthritis* tăng Precision (+75.00%) và Recall (+14.55%). Các bệnh khác cũng cải thiện đáng kể như *bronchitis* (+35.57% Precision, +22.50% Recall), *lung_cancer* (+26.08% Precision, +20.00% Recall) và *lung_infection* (+39.14% Precision, +11.25% Recall). Accuracy nhìn chung tăng nhẹ đến vừa (ví dụ pneumonia +15.06%, lung_cancer +8.59%), cho thấy EHR giúp mô hình phân biệt tốt hơn lớp dương tính mà không làm giảm mạnh độ chính xác tổng thể. Nhìn chung, kết quả này khẳng định EHR đóng vai trò quan trọng trong việc tăng độ tin cậy dự đoán dương tính và giảm bỏ sót bệnh khi đã cân bằng lớp bằng ROS.

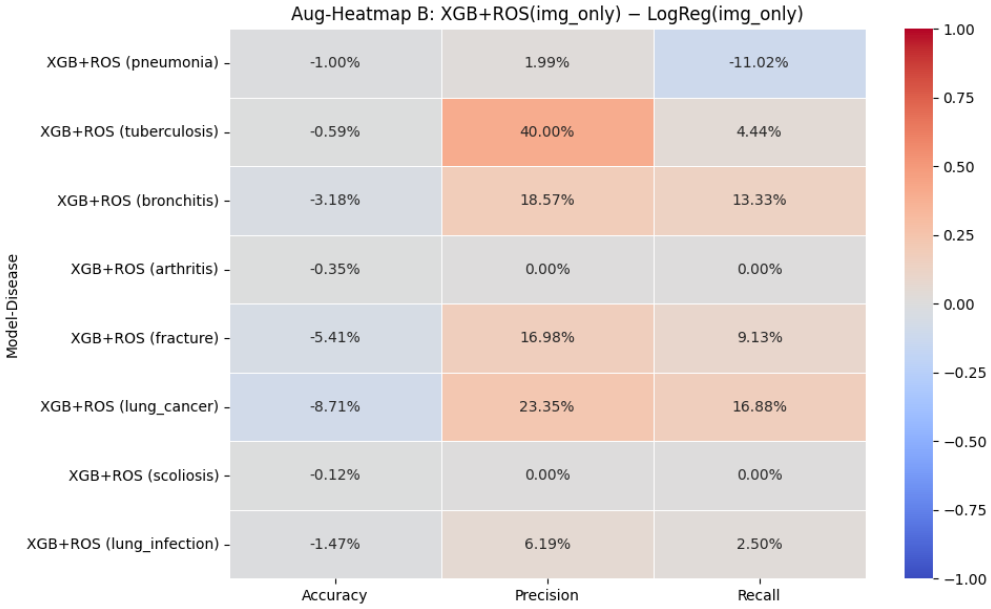


Figure 10: Performance gain of XGB+ROS over Logistic Regression using image-only features

Figure 10 so sánh XGB+ROS với baseline Logistic Regression khi **chỉ sử dụng đặc trưng ảnh**. Kết quả cho thấy Accuracy của XGB+ROS giảm nhẹ ở hầu hết bệnh (ví dụ fracture -5.41%, lung_cancer -8.71%), tuy nhiên Precision/Recall lại thường tăng ở nhiều bệnh: *tuberculosis* tăng Precision +40.00% và Recall +4.44%, *bronchitis* tăng Precision +18.57% và Recall +13.33%, *lung_cancer* tăng Precision +23.35% và Recall +16.88%, *fracture* tăng Precision +16.98% và Recall +9.13%. Điều này phản ánh trade-off thường gặp: mô hình boosting kết hợp oversampling có xu hướng **nhạy hơn với lớp dương tính** (tăng khả năng phát hiện bệnh) nhưng có thể làm thay đổi hành vi dự đoán trên lớp âm tính dẫn tới Accuracy không tăng tương ứng. Riêng *pneumonia* có Recall giảm (-11.02%), cho thấy trong một số trường hợp việc chỉ dựa vào đặc trưng ảnh và ROS chưa đủ ổn định hoặc phân bố nhãn ban đầu không còn theo dạng "dương tính hiếm". Ngoài ra, *arthritis* và *scoliosis* có độ lệch Precision/Recall bằng 0, gợi ý rằng trong thiết lập chỉ ảnh, cả hai mô hình đều gặp khó khăn trong việc dự đoán lớp dương tính cho các bệnh này.

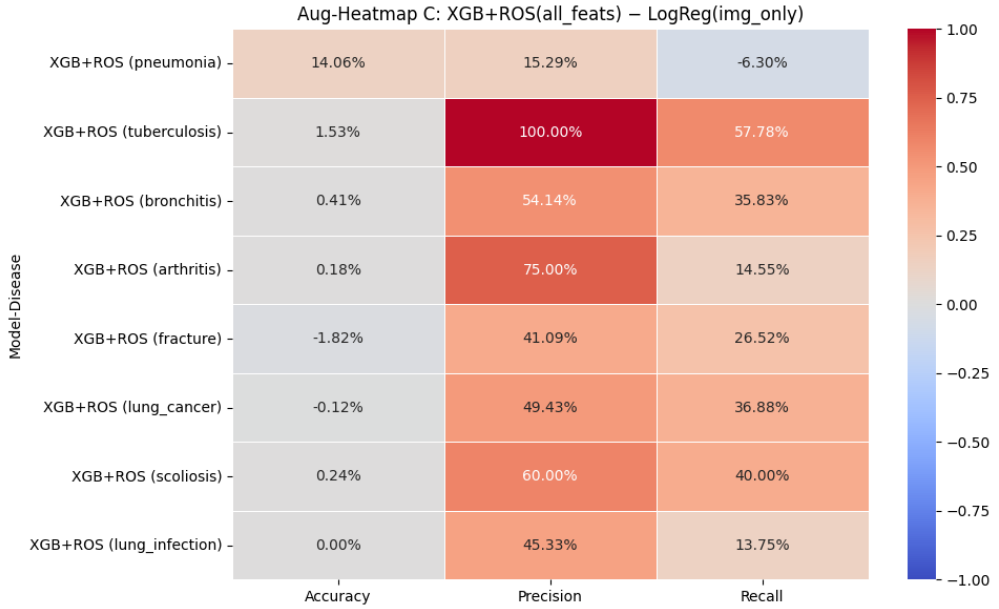


Figure 11: Performance gain of XGB+ROS with image+EHR features over Logistic Regression with image-only features

Figure 11 thể hiện mức cải thiện của pipeline đầy đủ (XGB+ROS với ảnh+EHR) so với baseline đơn giản (LogReg chỉ ảnh). Kết quả nổi bật nhất là Precision và Recall tăng rất mạnh ở nhiều bệnh hiếm: *tuberculosis* tăng Precision +100.00% và Recall +57.78%, *bronchitis* tăng Precision +54.14% và Recall +35.83%, *lung_ cancer* tăng Precision +49.43% và Recall +36.88%, *scoliosis* tăng Precision +60.00% và Recall +40.00%, *fracture* tăng Precision +41.09% và Recall +26.52%. Accuracy nhìn chung thay đổi nhỏ: một số bệnh tăng nhẹ (tuberculosis +1.53%, bronchitis +0.41%, scoliosis +0.24%), trong khi một vài bệnh giảm nhẹ (fracture -1.82%, lung_cancer -0.12%). Điều này cho thấy pipeline đầy đủ ưu tiên cải thiện khả năng phát hiện bệnh (Recall) và độ chính xác khi dự đoán dương tính (Precision) – vốn quan trọng trong y khoa – dù Accuracy tổng thể không phải lúc nào cũng tăng. Riêng *pneumonia* có Accuracy/Precision tăng (lần lượt +14.06% và +15.29%) nhưng Recall giảm (-6.30%), gợi ý rằng với bệnh này mô hình có xu hướng dự đoán "thận trọng" hơn (ít dự đoán dương tính hơn), từ đó làm giảm bỏ sót theo hướng ngược lại và cần cân nhắc điều chỉnh ngưỡng dự đoán nếu mục tiêu ưu tiên Recall.

5 Thảo luận

5.1 Nhận xét kết quả

5.1.1 SMOTE và XGBClassifier/XGBRegressor

Kết quả thực nghiệm cho thấy khi mô hình XGBoost chỉ sử dụng các đặc trưng hình ảnh trích xuất từ CXR, độ chính xác tổng thể (Accuracy) đạt mức cao. Tuy nhiên, Precision và Recall lại thấp đối với các bệnh có mức độ mất cân bằng lớp nghiêm trọng, cho thấy mô hình có xu hướng dự đoán phần lớn các mẫu là âm tính. Hiện tượng này phản ánh hạn chế phổ biến của các mô hình học máy khi làm việc với dữ liệu y sinh không cân bằng, trong đó các ca bệnh dương tính hiếm bị mô hình bỏ sót, làm giảm khả năng phát hiện bệnh trong thực tế lâm sàng.

Việc áp dụng kỹ thuật SMOTE trên các đặc trưng hình ảnh đã giúp cải thiện đáng kể Recall ở hầu hết các bệnh, đặc biệt là các bệnh có tỷ lệ dương tính thấp. Nhờ việc tạo thêm các mẫu tổng hợp cho lớp thiểu số, mô hình học được các đặc trưng đại diện tốt hơn cho các ca bệnh hiếm, từ đó giảm hiện tượng thiên lệch về lớp âm tính. Mặc dù Precision có xu hướng giảm trong một số trường hợp, sự gia tăng Recall cho thấy SMOTE giúp cải thiện khả năng phát hiện ca bệnh, vốn là yếu tố quan trọng trong các bài toán chẩn đoán y khoa.

Tổng thể, kết quả cho thấy việc kết hợp tăng cường dữ liệu ảnh và tích hợp dữ liệu EHR có thể giúp cải thiện hiệu suất dự đoán CRD ở một mức độ nhất định, đặc biệt đối với các lớp thiểu số, đồng thời nâng cao tính phù hợp của mô hình trong các ứng dụng hỗ trợ chẩn đoán lâm sàng.

5.1.2 ROS + XGBClassifier (Mô hình 2)

Kết quả từ các heatmap cho thấy ROS kết hợp XGBClassifier mang lại cải thiện đáng kể về khả năng phát hiện lớp dương tính khi mô hình được cung cấp thêm đặc trưng EHR. Cụ thể, so với trường hợp chỉ dùng đặc trưng ảnh, việc bổ sung EHR giúp tăng mạnh Precision và Recall ở nhiều bệnh hiếm như tuberculosis, scoliosis và arthritis. Điều này gợi ý rằng các biến lâm sàng có cấu trúc trong EHR đóng vai trò quan trọng trong việc giảm nhiễu và tăng khả năng phân biệt dương tính, đặc biệt khi dữ liệu ảnh bị hạn chế về tín hiệu hoặc mất cân bằng nghiêm trọng.

Khi chỉ sử dụng đặc trưng ảnh, XGB+ROS không luôn tối ưu Accuracy so với Logistic Regression, nhưng thường cải thiện Precision/Recall ở một số bệnh. Điều này phản ánh trade-off phổ biến: mô hình phi tuyến mạnh hơn (boosting) kết hợp oversampling có xu hướng "nhạy" hơn với lớp dương tính, tăng khả năng bắt bệnh nhưng có thể thay đổi hành vi dự đoán trên lớp âm tính, làm Accuracy không tăng tương ứng.

5.2 Hạn chế và khuyến nghị

5.2.1 SMOTE và XGBClassifier/XGBRegressor

Mặc dù các đặc trưng hình ảnh (CXR) sau khi được tăng cường bằng SMOTE được sử dụng như một dạng dữ liệu bổ sung để huấn luyện mô hình XGBoost cùng với các đặc trưng EHR, kết quả cho thấy hiệu quả cải thiện là không đáng kể. Cụ thể, các chỉ số Precision, Recall và

Accuracy của mô hình sử dụng EHR gần như không thay đổi so với trường hợp không áp dụng SMOTE, hoặc chỉ tăng ở mức rất nhỏ.

Điều này cho thấy việc tăng cường dữ liệu chỉ dựa trên các đặc trưng hình ảnh không đủ để bù đắp cho sự khan hiếm và thiếu thông tin trong EHR. Mặc dù số lượng mẫu huấn luyện tăng lên, các mẫu tổng hợp được tạo bởi SMOTE không cung cấp thêm thông tin lâm sàng mới cho EHR, khiến mô hình không học được các mối quan hệ có ý nghĩa hơn giữa đặc trưng hình ảnh và đặc trưng lâm sàng.

Ngoài ra, việc Recall tăng khi chỉ sử dụng đặc trưng hình ảnh sau SMOTE nhưng đi kèm với sự suy giảm Precision cho thấy mô hình có xu hướng dự đoán dương tính quá mức. Khi kết hợp với EHR, hiện tượng này không được cải thiện rõ rệt, dẫn đến việc chiến lược augmentation này chưa thực sự hữu dụng cho mô hình đa mô thức trong bối cảnh dữ liệu EHR thưa thớt.

Ngoài ra số lượng CRD và CXR mỗi bệnh nhân còn hạn chế, ground truth dựa vào mã ICD khi xuất viện có thể sai số, và thiếu sự tham gia của chuyên gia y tế để hướng dẫn mô hình học các đặc trưng quan trọng.

5.2.2 ROS và XGBClassifier

Mặc dù ROS giúp cân bằng số lượng mẫu dương tính và âm tính trong tập huấn luyện, bản chất của phương pháp này là **lặp lại các mẫu dương tính có sẵn** thay vì tạo ra thông tin mới. Do đó, ROS có thể làm mô hình **học quá khớp (overfitting)** lên một số mẫu dương tính đặc trưng, đặc biệt ở các bệnh rất hiếm như *tuberculosis* và *scoliosis*. Hiện tượng này có thể khiến các chỉ số cải thiện không đồng đều theo từng bệnh, và Accuracy đôi khi không tăng tương ứng với Precision/Recall trong một số thiết lập.

Kết quả so sánh cho thấy khi chỉ sử dụng đặc trưng ảnh, XGB+ROS không luôn vượt Logistic Regression về Accuracy, thậm chí có trường hợp giảm nhẹ. Điều này gợi ý rằng việc oversampling trong không gian đặc trưng ảnh có thể làm thay đổi biên quyết định theo hướng "nhạy" hơn với lớp dương tính, giúp tăng khả năng phát hiện bệnh ở một số bệnh nhưng cũng có thể làm giảm độ ổn định dự đoán trên lớp âm tính. Ngoài ra, ở một vài bệnh (ví dụ *pneumonia*), Recall có thể giảm khi so với baseline, cho thấy ROS không phải lúc nào cũng đem lại lợi ích nếu phân bố ban đầu không theo dạng "lớp dương tính hiếm" hoặc đặc trưng ảnh chưa đủ mạnh.

Khi kết hợp thêm đặc trưng EHR, các heatmap cho thấy Precision và Recall tăng mạnh ở nhiều bệnh hiếm, tuy nhiên điều này cũng đi kèm một hạn chế là **phụ thuộc vào chất lượng và mức độ đầy đủ của EHR**. Nếu EHR bị thiếu nhiều hoặc nhiễu, việc lặp mẫu bằng ROS không thể bổ sung thêm thông tin lâm sàng mới, và mô hình vẫn có thể bị giới hạn bởi tín hiệu đầu vào. Do đó, cải thiện từ ROS có thể không ổn định giữa các bệnh và giữa các lần chia dữ liệu.

Để khắc phục, cần (i) đánh giá mô hình trên nhiều seed hoặc cross-validation để kiểm tra tính ổn định; (ii) kết hợp ROS với regularization và tinh chỉnh siêu tham số của XGBClassifier (giới hạn độ sâu cây, tăng min_child_weight, subsample/colsample) nhằm giảm overfitting do dữ liệu lặp; và (iii) cân nhắc tối ưu ngưỡng dự đoán theo mục tiêu lâm sàng để kiểm soát trade-off Precision–Recall thay vì chỉ dựa vào ngưỡng mặc định. Ngoài ra, số lượng CRD và CXR mỗi bệnh nhân còn hạn chế, ground truth dựa vào mã ICD khi xuất viện có thể sai số, và thiếu sự tham gia của chuyên gia y tế để hướng dẫn mô hình học các đặc trưng quan trọng.

6 Kết luận

Dự án này đã phát triển một hệ thống chẩn đoán CRD tích hợp hình ảnh X-quang ngực (CXR) và các đặc trưng hồ sơ sức khỏe điện tử (EHR), nhằm giải quyết đồng thời bài toán đa mô thức và mất cân bằng lớp trong dữ liệu y sinh. Hai chiến lược tăng cường dữ liệu ở mức đặc trưng, bao gồm SMOTE và Random Over-Sampling (ROS), đã được khảo sát kết hợp với các mô hình dựa trên XGBoost để đánh giá tác động lên hiệu suất dự đoán.

Kết quả cho thấy SMOTE áp dụng trên các đặc trưng hình ảnh CXR giúp cải thiện đáng kể Recall, đặc biệt đối với các bệnh hiếm có tỷ lệ dương tính thấp, phản ánh khả năng giúp mô hình học tốt hơn các mẫu thuộc lớp thiểu số trong không gian đặc trưng ảnh. Tuy nhiên, khi kết hợp với các đặc trưng EHR, mức cải thiện là không đáng kể, cho thấy việc tạo mẫu tổng hợp từ đặc trưng ảnh không đủ để bổ sung thêm thông tin lâm sàng mới cho EHR vốn thưa thớt và thiếu dữ liệu. Mặc dù vậy, XGBClassifier/XGBRegressor vẫn cho thấy hiệu suất ổn định và khả năng giải thích thông qua tầm quan trọng đặc trưng, phản ánh các mẫu quyết định phù hợp với trực giác lâm sàng.

Đối với chiến lược ROS kết hợp XGBClassifier, việc cân bằng lớp bằng cách lặp lại các mẫu dương tính giúp giảm thiên lệch của mô hình về lớp âm tính và cải thiện rõ rệt Precision và Recall ở nhiều bệnh, đặc biệt khi bổ sung thêm các đặc trưng EHR. Trong khi thiết lập chỉ sử dụng đặc trưng ảnh thể hiện trade-off rõ ràng giữa Accuracy và khả năng phát hiện bệnh, pipeline XGB+ROS+EHR cho thấy hiệu suất ổn định hơn và ưu tiên tốt hơn cho việc phát hiện lớp dương tính, phù hợp với yêu cầu của các ứng dụng chẩn đoán y khoa. Tuy nhiên, do bản chất lặp mẫu, ROS vẫn tiềm ẩn nguy cơ overfitting ở các bệnh rất hiếm, dẫn đến mức cải thiện không đồng đều giữa các bệnh

Mặc dù còn tồn tại các hạn chế như số lượng CRD và CXR mỗi bệnh nhân còn hạn chế, ground truth dựa trên mã ICD khi xuất viện có thể gây sai lệch, và thiếu sự tham gia trực tiếp của chuyên gia y tế trong quá trình gán nhãn và đánh giá, hệ thống đề xuất vẫn cho thấy tiềm năng đáng kể. Các hướng nghiên cứu tiếp theo nên mở rộng tập dữ liệu, kết hợp ý kiến chuyên gia lâm sàng, đánh giá trên nhiều lần chia dữ liệu hoặc cross-validation, đồng thời tinh chỉnh siêu tham số và ngưỡng dự đoán theo mục tiêu lâm sàng để nâng cao độ tin cậy và khả năng áp dụng thực tế của hệ thống.

PHỤ LỤC

Link github: https://github.com/tronglinux123/EHR_Transfer_Learning

References

[1] Anthony L. Byrne, Ben J. Marais, Carole D. Mitnick, Leonid Lecca, and Guy B. Marks. Tuberculosis and chronic respiratory disease: a systematic review. *International Journal of Infectious Diseases*, 32:138–146, March 2015.

[2] Joan B. Soriano, Parkes J. Kendrick, Katherine R. Paulson, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 2020.

[3] Anthony Chapron, Emilie Andres, Laure Fiquet, Fabienne Pelé, Emmanuel Allory, Estelle Le Pabic, Aurélie Veislinger, Lisa Le Guillou, Stéphanie Guillot, Bruno Laviolle, and Stéphane Jouneau. Early detection of chronic obstructive pulmonary disease in primary care: a randomised controlled trial. *British Journal of General Practice*, 73(737):e876–e884, July 2023.

[4] Liqa A. Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with covid-19 pneumonia. *BMC Pulmonary Medicine*, 20(1), September 2020.

[5] R.M Hopstaken, T Witbraad, J.M.A van Engelshoven, and G.J Dinant. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. *Clinical Radiology*, 59(8):743–752, August 2004.

[6] Adnane Ait Nasser and Moulay A. Akhloufi. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics*, 13(1):159, January 2023.

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 618–626. IEEE, October 2017.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[9] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002. Accessed: 2026-01-05.

[11] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2023.

[12] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), January 2023.

[13] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), December 2019.

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019.

[15] Daniel Aletaha and Josef S. Smolen. Diagnosis and management of rheumatoid arthritis. *JAMA*, 320(13):1360, October 2018.

[16] Guillermo Stegen, Kenneth Jones, and Patricio Kaplan. CRITERIA FOR GUIDANCE IN THE DIAGNOSIS OF TUBERCULOSIS. *Pediatrics*, 43(2):260–263, February 1969.

[17] Joseph A Janicki and Benjamin Alman. Scoliosis: Review of diagnosis and treatment. *Paediatrics & Child Health*, 12(9):771–776, November 2007.

[18] Dawn E. Jaroszewski, Brandon J. Webb, and Kevin O. Leslie. Diagnosis and management of lung infections. *Thoracic Surgery Clinics*, 22(3):301–324, August 2012.

[19] Samuel N. Grief and Julie K. Loza. Guidelines for the evaluation and treatment of pneumonia. *Primary Care: Clinics in Office Practice*, 45(3):485–503, September 2018.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[21] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.

[22] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[23] 2. over-sampling — version 0.14.1. https://imbalanced-learn.org/stable/over_sampling.html. Accessed: 2026-1-5.

[24] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, August 2016. ACM.

[25] Cynthia Yang, Egill A Fridgeirsson, Jan A Kors, Jenna M Reps, and Peter R Rijnbeek. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J. Big Data*, 11(1), January 2024.