ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

□★□

Báo cáo Đồ án

RoTTA-ML: Generalizing Robust Test-Time Adaptation for Multi-Label Classification

Tổng quát hóa Phương pháp Test-Time Adaptation cho Bài toán Phân loại Đa nhãn

CS2311.CH190: Chuyên đề nghiên cứu về một số vấn đề chọn lọc trong khoa học máy tính

Giảng viên: TS. Dương Việt Hằng

Học viên thực hiện:

Văn Đức Ngọ 240101020

Phạm Thăng Long 240101016

Nguyễn Hoàng Hải 240101008

Tóm tắt

Khoảng cách giữa hiệu suất trong phòng thí nghiệm và độ tin cậy trong thực tiễn lâm sàng vẫn là một rào cản lớn đối với việc ứng dụng rộng rãi các mô hình AI chẩn đoán qua ảnh X-quang lồng ngực (CXR). Nguyên nhân cốt lõi là hiện tượng distribution shift, nơi các mô hình được huấn luyện trên một nguồn dữ liệu gặp khó khăn nghiêm trọng khi phải đối mặt với sự đa dạng của dữ liệu từ các bệnh viện và thiết bị khác nhau. Test-Time Adaptation (TTA) mang đến một giải pháp hấp dẫn, cho phép mô hình tự thích ứng mà không cần gán nhãn lại, nhưng cộng đồng nghiên cứu lại gần như bỏ qua một thách thức kép: làm thế nào để các phương pháp này hoạt động trong các kịch bản vừa động, vừa đa nhãn của y tế thực tế?

Để giải quyết lỗ hổng này, nhóm nghiên cứu đề xuất RoTTA-ML, một sự tổng quát hóa đầu tiên của RoTTA—một thuật toán TTA bền vững hàng đầu—cho bài toán phân loại đa nhãn. Thay vì chỉ đơn thuần điều chỉnh, nhóm đã tái thiết kế hai thành phần cốt lõi của nó: (1) một memory bank đa nhãn mới, MultiLabel-CSTU, với một thuật toán cân bằng nhãn động có khả năng bảo vệ các lớp bệnh hiếm; và (2) một mục tiêu học nhất quán (consistency objective) dựa trên Binary Cross-Entropy được xây dựng riêng cho không gian đầu ra đa nhãn.

Trên một benchmark đầy thách thức do nhóm thiết lập—nơi một mô hình MobileNetV3-small huấn luyện trên CheXpert phải thích ứng với luồng dữ liệu nhiễu động từ NIH-14—RoTTA-ML đã chứng tỏ sự vượt trội. Nó không chỉ cải thiện đáng kể chỉ số Mean AUC so với baseline không thích ứng mà còn cho thấy sự bền vững vượt xa TENT, một phương pháp phổ biến khác bị sụp đổ hoàn toàn trong kịch bản của chúng tôi. Phân tích sâu về động lực học của memory bank không chỉ xác nhận hiệu quả của cơ chế cân bằng mà còn phơi bày một hạn chế cố hữu của các phương pháp tự huấn luyện: một "điểm mù" đối với các lớp bệnh mà mô hình ban đầu không thể nhận diện. Phát hiện này cung

cấp những định hướng quan trọng cho việc phát triển các hệ thống TTA an toàn và đáng tin cậy hơn trong tương lai.

1. Giới thiệu

1.1. Bối cảnh: Trí tuệ Nhân tạo trong Y tế

Trong thập kỷ qua, lĩnh vực Trí tuệ Nhân tạo (Artificial Intelligence - AI), đặc biệt là học sâu (deep learning), đã tạo ra một cuộc cách mạng trong nhiều ngành công nghiệp, và y tế là một trong những lĩnh vực hưởng lợi sâu sắc nhất. Các mô hình học sâu, đặc biệt là mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs), đã chứng tỏ khả năng phi thường trong việc phân tích hình ảnh y tế, đạt đến và đôi khi vượt qua độ chính xác của các chuyên gia con người trong các tác vụ cụ thể. Một trong những ứng dụng nổi bật nhất là phân tích ảnh X-quang lồng ngực (CXR), một công cụ chẩn đoán tuyến đầu, phổ biến và có chi phí thấp. Các mô hình AI có thể tự động phát hiện nhiều bệnh lý như viêm phổi, tràn dịch màng phổi, hay các khối u, hứa hẹn giảm tải công việc cho các bác sĩ X-quang, tăng tốc độ chẩn đoán và cải thiện khả năng tiếp cận chăm sóc sức khỏe ở những khu vực thiếu chuyên gia.

Động lực thúc đẩy việc ứng dụng AI vào chẩn đoán CXR không chỉ đến từ tiềm năng công nghệ mà còn từ những thách thức cố hữu của phương pháp chẩn đoán thủ công. Các bác sĩ X-quang hàng ngày phải đối mặt với một khối lượng công việc khổng lồ, dẫn đến nguy cơ mệt mỏi và sai sót. Hơn nữa, nhiều dấu hiệu bệnh lý trên ảnh CXR rất tinh vi và khó phát hiện, đòi hỏi trình độ chuyên môn và kinh nghiệm cao. Sự khác biệt trong diễn giải giữa các bác sĩ (inter-observer variability) cũng là một vấn đề đã được ghi nhận, có thể dẫn đến sự không nhất quán trong chẩn đoán. Trong bối cảnh đó, AI không đóng vai trò thay thế con người, mà là một công cụ hỗ trợ đắc lực, hoạt động như một cặp mắt thứ hai giúp sàng lọc các ca bệnh, phát hiện các vùng bất thường tiềm ẩn, và ưu tiên các trường hợp khẩn cấp, qua đó nâng cao chất lượng chẩn đoán tổng thể.

Sức mạnh của các mô hình CNNs nằm ở khả năng tự động học các biểu diễn đặc trưng phân cấp (hierarchical feature representations) trực tiếp từ dữ liệu pixel. Ở các lớp đầu tiên, mạng học cách nhận diện các đặc trưng cơ bản như cạnh, góc, và kết cấu (texture). Càng đi sâu vào các lớp sau, những đặc trưng này được tổ hợp lại để hình thành các cấu trúc phức tạp hơn, chẳng hạn như hình dạng của tim, đường viền của phổi, và cuối cùng là các dấu hiệu bệnh lý như vùng đông đặc (consolidation) hay các nốt mờ (nodules). Quá trình học này đòi hỏi một lượng dữ liệu cực lớn, và sự ra đời của các bộ dữ liệu công khai quy mô lớn như CheXpert, NIH-14, và MIMIC-CXR chính là chất xúc tác cho những bước tiến vượt bậc trong lĩnh vực này. Các bộ dữ liệu này cung cấp hàng trăm ngàn hình ảnh có gán nhãn, cho phép các mô hình học được sự đa dạng và biến thiên của các bệnh lý lồng ngực.

Tuy nhiên, sự thành công ấn tượng của các mô hình này trong môi trường học thuật có kiểm soát lại che giấu một thách thức nền tảng cản trở việc ứng dụng rộng rãi: làm thế nào để đảm bảo các mô hình này hoạt động một cách đáng tin cậy và bền vững khi đối mặt với sự đa dạng và không thể đoán trước của dữ liệu lâm sàng trong thế giới thực? Câu hỏi này chính là tiền đề dẫn đến vấn đề cốt lõi mà nghiên cứu của chúng tôi tập trung giải quyết: distribution shift.

1.2. Vấn đề cốt lõi: Distribution Shift

Mặc dù các mô hình AI đạt được thành công vang dội trong môi trường nghiên cứu có kiểm soát, việc triển khai chúng vào thực tiễn lâm sàng lại vấp phải một rào cản lớn: vấn đề distribution shift. Về mặt kỹ thuật, distribution shift là hiện tượng phân phối xác suất của dữ liệu trong thực tế (target domain) khác biệt so với phân phối của dữ liệu được dùng để huấn luyện mô hình (source domain). Nói một cách đơn giản, những gì mô hình nhìn thấy lúc triển khai không còn giống với những gì nó đã học trong quá trình huấn luyện. Hậu quả trực tiếp và nguy hiểm nhất là sự suy giảm hiệu năng một cách thầm lặng và khó lường, nơi một mô hình từng đạt độ chính xác trên 90% trong phòng thí nghiệm có thể hoạt

động không hơn gì việc đoán ngẫu nhiên khi đưa vào sử dụng thực tế.

Trong lĩnh vực y tế, distribution shift không phải là một ngoại lệ mà là một quy luật cố hữu và không thể tránh khỏi. Nó xuất phát từ một hệ sinh thái phức tạp và đa dạng, bao gồm:

- **Sự đa dạng của Thiết bị (Hardware Variance):** Các bệnh viện và phòng khám sử dụng máy chụp X-quang từ nhiều nhà sản xuất khác nhau (ví dụ: Siemens, Philips, GE). Mỗi dòng máy có các đặc tính cảm biến, bộ lọc, và thuật toán tái tạo hình ảnh riêng, dẫn đến sự khác biệt tinh vi về độ tương phản, độ sắc nét và mức độ nhiễu nền trong các ảnh X-quang. Một mô hình được huấn luyện chủ yếu trên dữ liệu từ máy Siemens có thể gặp khó khăn khi diễn giải các đặc trưng từ ảnh chụp bằng máy GE.
- **Đặc điểm Bệnh nhân (Population Shift)**: Phân phối của các bệnh lý và đặc điểm nhân khẩu học không đồng nhất trên toàn cầu. Một mô hình được huấn luyện trên dữ liệu từ một bệnh viện ở Bắc Mỹ, nơi có tỷ lệ béo phì cao, có thể hoạt động kém hiệu quả khi áp dụng cho một quần thể bệnh nhân ở châu Á, nơi có thể trạng và các bệnh lý phổ biến khác. Sự khác biệt về tuổi tác, giới tính, và các bệnh lý nền cũng tạo ra những thay đổi trong biểu hiện hình ảnh mà mô hình có thể chưa từng học qua.
- **Quy trình Chụp (Protocol Variance):** Ngay cả trong cùng một bệnh viện, quy trình chụp cũng có thể không hoàn toàn đồng nhất. Tư thế của bệnh nhân (đứng, nằm), mức độ hít vào của phổi tại thời điểm chụp, các thông số phơi sáng (kVp, mAs) do kỹ thuật viên lựa chọn, và góc chụp (trước-sau, sau-trước) đều ảnh hưởng đến hình ảnh cuối cùng. Sự thiếu nhất quán này tạo ra một dạng shift liên tục mà mô hình phải đối mặt.
- **Nhiễu và Suy giảm Chất lượng (Corruptions and Degradations):** Ngoài các shift mang tính hệ thống, dữ liệu thực tế còn phải chịu đựng các dạng suy giảm chất lượng ngẫu nhiên. Hình ảnh có thể bị nhiễu do lỗi cảm biến (sensor noise), bị mờ do chuyển động của bệnh nhân (motion

blur), hoặc bị thay đổi độ sáng/tương phản do các yếu tố môi trường. Các dị vật như dây điện cực, ống nội khí quản, hay các thiết bị y tế khác cũng là những yếu tố không xuất hiện trong các bộ dữ liệu sạch nhưng lại rất phổ biến trong thực tế.

Hậu quả của distribution shift trong y tế là đặc biệt nghiêm trọng. Một mô hình AI đưa ra chẩn đoán sai không chỉ là một lỗi thống kê, mà còn có thể dẫn đến việc bỏ sót bệnh lý nguy hiểm hoặc chẩn đoán nhầm gây ra các can thiệp y tế không cần thiết, ảnh hưởng trực tiếp đến sức khỏe và tính mạng của bệnh nhân. Sự thiếu tin cậy này là rào cản lớn nhất ngăn cản việc chấp nhận và tích hợp rộng rãi các công cụ AI vào quy trình làm việc lâm sàng. Do đó, việc phát triển các mô hình không chỉ chính xác mà còn phải bền vững (robust) và có khả năng thích ứng (adaptive) với sự thay đổi của môi trường là một yêu cầu cấp thiết.

1.3. Test-Time Adaptation: Một hướng giải quyết tiềm năng

Trước thách thức của distribution shift, các giải pháp truyền thống như thu thập và gán nhãn lại dữ liệu từ miền đích để fine-tune lại mô hình thường không khả thi trong thực tế. Quá trình này không chỉ tốn kém về mặt tài chính và nhân lực chuyên môn mà còn tạo ra độ trễ đáng kể, làm cho mô hình luôn đi sau sự thay đổi của môi trường. Một hướng tiếp cận linh hoạt và thực tế hơn đã nổi lên, đó là Thích ứng tại Thời điểm Kiểm thử (Test-Time Adaptation - TTA).

Triết lý cốt lõi của TTA là trao cho mô hình khả năng tự học và tự điều chỉnh ngay trong quá trình hoạt động. Thay vì là một thực thể tĩnh, mô hình trở thành một hệ thống động, có khả năng cập nhật các tham số của mình một cách on-the-fly (tức thời) bằng cách học từ chính luồng dữ liệu kiểm thử không nhãn mà nó đang xử lý. Bằng cách tận dụng các tín hiệu tự giám sát (self-supervised signals) có sẵn trong dữ liệu đích, chẳng hạn như sự tự tin trong dự đoán hoặc tính nhất quán dưới các phép biến đổi, TTA có thể điều chỉnh mô hình để phù hợp hơn với các đặc điểm của phân phối mới mà không cần bất kỳ sự can thiệp

nào từ con người.

Tuy nhiên, các kịch bản TTA trong thế giới thực, đặc biệt là trong các ứng dụng như giám sát y tế liên tục hoặc xe tự lái, còn phức tạp hơn nhiều so với các giả định trong phòng thí nghiệm. Kịch bản này, thường được gọi là Practical TTA (PTTA), phải đối mặt với hai thách thức lớn xảy ra đồng thời:

- **Phân phối Thay đổi Liên tục (Continual Distribution Shift):** Môi trường không chỉ thay đổi một lần mà thay đổi không ngừng. Ví dụ, trong một ngày, chất lượng ảnh X-quang có thể thay đổi giữa các ca làm việc của kỹ thuật viên, hoặc khi một máy chụp cũ được thay thế bằng một máy mới. Mô hình phải có khả năng liên tục học cái mới mà không bị quên đi kiến thức cũ một cách thảm khốc (catastrophic forgetting).
- **Dữ liệu có Tương quan (Correlated Data):** Dữ liệu không đến một cách độc lập và ngẫu nhiên (i.i.d). Trong một khoảng thời gian ngắn, các mẫu có thể rất giống nhau. Ví dụ, một khoa cấp cứu có thể tiếp nhận một loạt các ca chấn thương ngực trong một buổi tối, dẫn đến một luồng dữ liệu tạm thời bị mất cân bằng nghiêm trọng. Nếu không cẩn thận, mô hình có thể học lệch (overfit) vào xu hướng tạm thời này và sụp đổ (model collapse), quên đi cách nhận diện các bệnh lý khác.

RoTTA (Robust Test-Time Adaptation) \cite{yuan2023robust} là một phương pháp state-of-the-art được thiết kế đặc biệt để giải quyết chính xác kịch bản PTTA đầy thách thức này. Tuy nhiên, một lỗ hổng nghiên cứu lớn vẫn còn tồn tại: RoTTA, giống như hầu hết các phương pháp TTA khác, được xây dựng và đánh giá trên các bài toán phân loại đơn nhãn (single-label classification). Trong khi đó, các bài toán thực tế quan trọng như chẩn đoán CXR lại có bản chất là đa nhãn (multi-label), nơi một hình ảnh có thể chứa nhiều bệnh lý cùng lúc. Việc áp dụng trực tiếp các cơ chế của RoTTA vào bối cảnh này là không thể, đòi hỏi một sự tái thiết kế và tổng quát hóa sâu sắc.

1.4. Đóng góp của Nghiên cứu: Xây dựng RoTTA-ML

Nghiên cứu này giải quyết trực tiếp lỗ hổng đã nêu bằng cách đề xuất RoTTA-ML, một sự tổng quát hóa toàn diện và đầu tiên của framework RoTTA cho bài toán phân loại đa nhãn trong các kịch bản động. Chúng tôi không chỉ đơn thuần điều chỉnh một vài tham số, mà tái thiết kế các thành phần cốt lõi của thuật toán để phù hợp với các đặc thù của không gian dự đoán đa nhãn. Các đóng góp chính của công trình này bao gồm:

- Đề xuất một Framework TTA Đa nhãn Hoàn chỉnh: Chúng tôi xây dựng một hệ thống end-to-end có khả năng thích ứng liên tục cho các mô hình chẩn đoán đa nhãn. Framework này kế thừa triết lý bền vững của RoTTA nhưng được trang bị các cơ chế mới để xử lý hiệu quả các luồng dữ liệu y tế phức tạp.
- Thiết kế Memory Bank Đa nhãn (MultiLabel-CSTU): Đây là đóng góp kỹ thuật trọng tâm. Chúng tôi đề xuất một kiến trúc memory bank mới với cấu trúc phẳng, loại bỏ ràng buộc mỗi mẫu chỉ thuộc một lớp. Quan trọng hơn, chúng tôi giới thiệu một thuật toán cân bằng thông minh, có khả năng theo dõi và duy trì sự đa dạng của từng nhãn bệnh riêng lẻ. Cơ chế bảo vệ lớp thiểu số được thiết kế để chủ động chống lại sự mất cân bằng nhãn nghiêm trọng, một vấn đề cố hữu trong dữ liệu y tế.
- Xây dựng Mục tiêu Học nhất quán cho Đa nhãn: Chúng tôi xây dựng một mục tiêu học (learning objective) và các thành phần phụ trợ phù hợp. Điều này bao gồm việc định nghĩa lại cách tạo pseudo-label (dựa trên sigmoid và ngưỡng), một thước đo uncertainty mới (dựa trên tổng của binary entropy), và một hàm consistency loss (bce_entropy) dựa trên Binary Cross-Entropy để hướng dẫn quá trình học của mô hình Student từ Teacher.
- Thiết lập Quy trình Đánh giá Nghiêm ngặt: Chúng tôi thiết lập một quy trình thực nghiệm chi tiết và có tính tái lập để đánh giá phương pháp của mình. Kịch bản domain shift được tạo ra một cách có kiểm soát giữa hai

bộ dữ liệu CXR quy mô lớn là CheXpert và NIH-14, kết hợp với một luồng nhiễu động. Chúng tôi tiến hành so sánh công bằng giữa RoTTA-ML và một phiên bản TENT đã được điều chỉnh cho đa nhãn, cung cấp những phân tích định lượng đầu tiên về hiệu quả của các chiến lược TTA phức tạp trong lĩnh vực này.

2. Related Work

Nghiên cứu của chúng tôi được xây dựng dựa trên nền tảng của nhiều lĩnh vực liên quan, chủ yếu nằm ở giao điểm của (1) học sâu cho phân loại hình ảnh y tế đa nhãn, (2) các phương pháp giải quyết distribution shift, và (3) các kỹ thuật thích ứng tại thời điểm kiểm thử (Test-Time Adaptation).

2.1. Học sâu cho Phân loại ảnh X-quang Đa nhãn

Việc ứng dụng các mạng nơ-ron tích chập (CNNs) đã trở thành tiêu chuẩn vàng trong phân tích ảnh CXR. Các công trình tiên phong như CheXNet \cite{rajpurkar2017chexnet} đã chứng minh rằng các mô hình, chẳng hạn như DenseNet, có thể đạt được hiệu suất ngang tầm với các bác sĩ X-quang trong việc phát hiện viêm phổi. Thành công này đã khởi đầu cho một làn sóng nghiên cứu, được thúc đẩy mạnh mẽ bởi sự ra đời của các bộ dữ liệu công khai quy mô lớn như CheXpert \cite{irvin2019chexpert}, NIH ChestX-ray14 \cite{wang2017chestx}, và MIMIC-CXR \cite{johnson2019mimic}.

Các nghiên cứu sau đó đã tập trung vào việc giải quyết các thách thức đặc thù của dữ liệu CXR. Nhiều công trình đã khám phá các kiến trúc mạng phức tạp hơn, sử dụng các cơ chế chú ý (attention mechanisms) để mô hình tập trung vào các vùng bệnh lý quan trọng. Một hướng đi khác là xử lý sự mất cân bằng nghiêm trọng của các lớp bệnh bằng các kỹ thuật như trọng số hóa hàm loss (loss re-weighting) hoặc các phương pháp lấy mẫu tinh vi. Để giải quyết bản chất đa nhãn, các nhà nghiên cứu đã khám phá việc mô hình hóa sự phụ thuộc

giữa các bệnh lý, ví dụ như sử dụng Mạng Đồ thị (Graph Neural Networks) để học mối tương quan rằng một số bệnh thường xuất hiện cùng nhau.

Tuy nhiên, một hạn chế chung của hầu hết các công trình này là chúng được huấn luyện và đánh giá dựa trên giả định rằng dữ liệu huấn luyện và kiểm thử được lấy từ cùng một phân phối (i.i.d. assumption). Giả định này hiếm khi đúng trong thực tế lâm sàng, nơi các mô hình thường xuyên phải đối mặt với dữ liệu từ các nguồn hoàn toàn mới, dẫn đến sự cần thiết của các kỹ thuật thích ứng.

2.2. Các hướng giải quyết Distribution Shift

Để vượt qua rào cản của distribution shift, nhiều phương pháp đã được đề xuất.

- Domain Adaptation (DA) & Domain Generalization (DG):
    - Unsupervised Domain Adaptation (UDA) là một hướng tiếp cận phổ biến, trong đó mô hình cố gắng học các đặc trưng bất biến (invariant features) bằng cách giảm thiểu sự khác biệt thống kê giữa miền nguồn có nhãn và miền đích không nhãn. Các kỹ thuật thường được sử dụng bao gồm Maximum Mean Discrepancy (MMD) hoặc học đối nghịch (adversarial learning).
    - Domain Generalization (DG) còn tham vọng hơn, cố gắng học một mô hình có khả năng tổng quát hóa tốt trên một miền đích hoàn toàn chưa từng thấy bằng cách học từ nhiều miền nguồn đa dạng.
    - Hạn chế: Cả DA và DG đều yêu cầu quyền truy cập vào dữ liệu (hoặc ít nhất là một lượng lớn dữ liệu) từ miền nguồn trong quá trình huấn luyện hoặc thích ứng. Trong nhiều kịch bản thực tế, đặc biệt là y tế, dữ liệu nguồn thường là độc quyền, nhạy cảm về quyền riêng tư và không thể truy cập được tại thời điểm triển khai, khiến các phương pháp này không khả thi.
- Semi-Supervised và Self-Supervised Learning (SSL):
    - Các ý tưởng cốt lõi của TTA bắt nguồn sâu sắc từ SSL. Pseudo-Labeling là một trong những kỹ thuật đơn giản nhất, sử

dụng chính các dự đoán tự tin của mô hình làm nhãn tạm thời để huấn luyện tiếp.

- Consistency Regularization là một ý tưởng mạnh mẽ hơn, buộc mô hình phải đưa ra các dự đoán nhất quán cho cùng một đầu vào dưới các phép biến đổi (augmentation) khác nhau. Các framework như Mean Teacher, nơi một mô hình học trò (student) học từ một mô hình người thầy (teacher) có trọng số được cập nhật mượt mà (qua EMA), đã trở thành nền tảng cho nhiều phương pháp TTA hiện đại.
- Sự khác biệt chính: SSL thường được áp dụng trong một giai đoạn huấn luyện ngoại tuyến (offline), nơi có cả dữ liệu có nhãn và không nhãn. Ngược lại, TTA hoạt động hoàn toàn trực tuyến (online) trong giai đoạn inference, chỉ với dữ liệu không nhãn.

2.3. Test-Time Adaptation (TTA)

TTA giải quyết trực tiếp hạn chế về quyền truy cập dữ liệu nguồn. Các phương pháp TTA có thể được phân loại thành nhiều trường phái chính, và các phương pháp chúng tôi đánh giá là những đại diện tiêu biểu.

- **Thích ứng dựa trên Chuẩn hóa (Normalization-based Adaptation)**: Đây là cách tiếp cận đơn giản nhất, tập trung vào việc điều chỉnh các thống kê trong các lớp chuẩn hóa. BN-Adapt (hoặc NORM) buộc các lớp Batch Normalization (BN) hoạt động ở chế độ train(), sử dụng thống kê của batch dữ liệu hiện tại thay vì các thống kê chạy (running statistics) đã lỗi thời từ miền nguồn.
- **Tối thiểu hóa Entropy (Entropy Minimization):** Dựa trên giả định rằng một mô hình tốt nên đưa ra các dự đoán tự tin (có entropy thấp). TENT \cite{wang2020tent} là một phương pháp kinh điển thuộc trường phái này. Nó chỉ cập nhật các tham số affine (γ, β) của lớp BN để tối thiểu hóa entropy của dự đoán đầu ra. Mặc dù hiệu quả và nhẹ, TENT có thể dễ bị tích lũy lỗi nếu các dự đoán tự tin ban đầu là sai.

- **Tự huấn luyện và Học nhất quán (Self-training and Consistency-based)**: Đây là các phương pháp phức tạp hơn, kết hợp các cơ chế phòng vệ để tăng cường sự ổn định. Các phương pháp như CoTTA \cite{wang2022continual} cải tiến TENT bằng cách sử dụng mô hình Teacher-Student và một cơ chế khôi phục ngẫu nhiên (stochastic restoration) để chống lại hiện tượng quên. RoTTA \cite{yuan2023robust}, phương pháp nền tảng của chúng tôi, tiến một bước xa hơn bằng cách giới thiệu một memory bank để lưu trữ một tập hợp con các mẫu đáng tin cậy và đa dạng từ luồng kiểm thử. Memory bank này hoạt động như một bộ đệm tri thức, giúp ổn định quá trình tạo pseudo-label và cho phép mô hình học từ một cái nhìn tổng thể và cân bằng hơn về phân phối của miền đích, thay vì từ các batch riêng lẻ có thể bị nhiễu và mất cân bằng.

Trong khi các ý tưởng nền tảng này đã được khám phá, việc áp dụng và so sánh một cách có hệ thống các phương pháp TTA đa dạng này cho bối cảnh đa nhãn, đặc biệt trong lĩnh vực y tế đầy thách thức, vẫn là một hướng đi cần thiết và là trọng tâm của nghiên cứu này.

3. Dữ liệu (Data)

Việc lựa chọn dữ liệu phù hợp là yếu tố then chốt để có thể xây dựng và đánh giá một cách có ý nghĩa các phương pháp giải quyết distribution shift. Dự án của chúng tôi sử dụng hai bộ dữ liệu X-quang ngực công khai lớn và phổ biến hàng đầu thế giới, CheXpert và NIH Chest X-ray14, để tạo ra một kịch bản domain shift thực tế và đầy thách thức.

3.1. CheXpert: Miền Nguồn (Source Domain)

- **Nguồn gốc và Quy mô:** Bộ dữ liệu CheXpert \cite{irvin2019chexpert} được phát hành bởi Đại học Stanford, là một trong những bộ dữ liệu CXR lớn nhất hiện có, chứa 224,316 ảnh X-quang ngực từ 65,240 bệnh nhân. Quy mô lớn của nó cung cấp một nguồn dữ liệu phong phú và đa dạng, lý

tưởng cho việc huấn luyện các mô hình học sâu mạnh mẽ có khả năng học được các biểu diễn đặc trưng phức tạp.

- **Quá trình Gán nhãn và Xử lý:** Một trong những đặc điểm nổi bật của CheXpert là việc sử dụng một công cụ xử lý ngôn ngữ tự nhiên (NLP) tiên tiến để tự động trích xuất các nhãn bệnh lý từ các báo cáo X-quang đi kèm. Công cụ này có khả năng nhận diện và phân loại 14 quan sát phổ biến, đồng thời xử lý các trường hợp không chắc chắn (uncertainty). Ví dụ, một báo cáo ghi không thể loại trừ viêm phổi sẽ được gán nhãn không chắc chắn cho bệnh viêm phổi. Trong nghiên cứu của chúng tôi, theo thông lệ chung, các nhãn không chắc chắn này được coi là nhãn dương tính để tối đa hóa việc thu nhận các ca bệnh tiềm năng.
- **Vai trò trong dự án:** Chúng tôi sử dụng bộ dữ liệu này làm miền nguồn (source domain) để huấn luyện (fine-tune) mô hình cơ sở. Việc huấn luyện trên một bộ dữ liệu quy mô lớn và chất lượng cao như CheXpert đảm bảo rằng mô hình ban đầu của chúng tôi có một nền tảng kiến thức vững chắc về các đặc điểm hình ảnh của các bệnh lý lồng ngực. Chúng tôi tập trung vào 5 bệnh lý có tầm quan trọng lâm sàng và tỷ lệ xuất hiện hợp lý: \textit{Atelectasis (Xẹp phổi), Cardiomegaly (Tim to), Consolidation (Đông đặc), Pleural Effusion (Tràn dịch màng phổi),} và \textit{Pneumothorax (Tràn khí màng phổi)}.

3.2. NIH Chest X-ray14: Miền Đích (Target Domain)

- Nguồn gốc và Quy mô: Bộ dữ liệu NIH Chest X-ray14 \cite{wang2017chestx} được Viện Y tế Quốc gia Hoa Kỳ (NIH) công bố, bao gồm 112,120 ảnh X-quang ngực từ 30,805 bệnh nhân. Mặc dù nhỏ hơn CheXpert, đây vẫn là một bộ dữ liệu rất lớn và có giá trị, được sử dụng rộng rãi trong cộng đồng nghiên cứu.
- Đặc điểm và Sự khác biệt: Tương tự như CheXpert, nhãn của NIH-14 cũng được trích xuất từ các báo cáo y khoa bằng phương pháp NLP. Tuy

nhiên, dữ liệu này được thu thập từ một hệ thống bệnh viện hoàn toàn khác, với các thiết bị, quy trình và quần thể bệnh nhân khác biệt so với Stanford. Chính sự khác biệt này tạo ra một domain shift tự nhiên, thực tế. Ví dụ, các nghiên cứu đã chỉ ra rằng có sự khác biệt về phân phối pixel, độ tương phản trung bình, và thậm chí cả tỷ lệ mắc các bệnh lý giữa hai bộ dữ liệu này.

- Vai trò trong dự án: Chúng tôi sử dụng tập kiểm thử của bộ dữ liệu này làm miền đích (target domain). Đây là môi trường xa lạ mà mô hình đã huấn luyện của chúng tôi phải đối mặt. Dữ liệu được lọc để chỉ giữ lại các ảnh và nhãn tương ứng với 5 bệnh lý mục tiêu để đảm bảo sự tương thích với mô hình. Việc đánh giá trên NIH-14 cho phép chúng tôi đo lường một cách khách quan khả năng của các phương pháp TTA trong việc thích ứng với một môi trường lâm sàng mới mà không cần huấn luyện lại từ đầu.

3.3. Tiền xử lý Dữ liệu

Để đảm bảo rằng các so sánh giữa các phương pháp là công bằng và kết quả nghiên cứu có tính tái lập, việc xây dựng một quy trình tiền xử lý dữ liệu (preprocessing pipeline) nhất quán và được định nghĩa rõ ràng là cực kỳ quan trọng. Cả hai bộ dữ liệu CheXpert và NIH-14, mặc dù có nguồn gốc khác nhau, đều được đưa qua cùng một chuỗi các bước xử lý trước khi được đưa vào mô hình.

3.3.1. Lọc và Lựa chọn Nhãn Bệnh (Label Filtering and Selection)

Bước đầu tiên và quan trọng nhất là xác định phạm vi của bài toán. Cả hai bộ dữ liệu gốc đều chứa 14 bệnh lý khác nhau. Tuy nhiên, nhiều bệnh lý trong số đó có tỷ lệ xuất hiện rất thấp (extreme imbalance), gây khó khăn cho việc huấn luyện và đánh giá một cách đáng tin cậy. Do đó, chúng tôi đã đưa ra quyết định chiến lược là tập trung vào một tập hợp con gồm 5 bệnh lý lồng ngực phổ biến và có tầm quan trọng lâm sàng cao:

- Atelectasis (Xẹp phổi)
- Cardiomegaly (Tim to)
- Consolidation (Đông đặc)
- Pleural Effusion (Tràn dịch màng phổi)
- Pneumothorax (Tràn khí màng phổi)

Sau khi xác định được tập nhãn mục tiêu, chúng tôi tiến hành lọc cả hai bộ dữ liệu. Tất cả các cột nhãn không thuộc 5 bệnh lý này đều được loại bỏ. Quá trình này đảm bảo rằng mô hình chỉ học và được đánh giá trên các lớp bệnh có liên quan, tạo ra một bài toán tập trung và rõ ràng hơn.

3.3.2. Xử lý Nhãn không chắc chắn (Handling Uncertainty Labels)

Một đặc điểm chung của các bộ dữ liệu được gán nhãn tự động từ báo cáo y khoa là sự tồn tại của các nhãn không chắc chắn. Trong bộ dữ liệu CheXpert, công cụ gán nhãn có thể đưa ra giá trị U (Uncertain) cho một bệnh lý khi ngôn ngữ trong báo cáo không đủ để khẳng định hay phủ định một cách chắc chắn. Theo các thông lệ phổ biến trong nhiều nghiên cứu trước đây về CheXpert, chúng tôi áp dụng chiến lược coi các trường hợp không chắc chắn (U) là dương tính (1). Lý do đằng sau quyết định này là để tăng độ nhạy (sensitivity) của mô hình, ưu tiên việc không bỏ sót các ca bệnh tiềm năng, điều này phù hợp hơn với các ứng dụng sàng lọc trong lâm sàng.

3.3.3. Xử lý Hình ảnh (Image Processing)

Tất cả các hình ảnh từ cả hai bộ dữ liệu sau đó được xử lý qua các bước sau:

- Chuyển đổi sang ảnh xám: Mặc dù hầu hết ảnh X-quang về bản chất là ảnh đơn sắc, các định dạng file (như JPG) có thể lưu chúng dưới dạng 3 kênh màu (RGB) lặp lại. Để đảm bảo tính nhất quán, tất cả các ảnh đều được chuyển đổi thành định dạng ảnh xám 1 kênh.
- Thay đổi kích thước (Resizing): Các mô hình CNN hiện đại thường yêu cầu đầu vào có kích thước cố định. Chúng tôi thay đổi kích thước của tất

cả các ảnh thành 224x224 pixels. Kích thước này là một sự cân bằng hợp lý giữa việc giữ lại đủ chi tiết lâm sàng và tối ưu hóa chi phí tính toán, đồng thời cũng là kích thước đầu vào tiêu chuẩn cho nhiều kiến trúc được pre-train trên ImageNet.

- Chuẩn hóa (Normalization): Đây là một bước quan trọng để ổn định quá trình huấn luyện. Giá trị của mỗi pixel (thường từ 0-255) trước tiên được chuẩn hóa về khoảng [0, 1]. Sau đó, chúng tôi áp dụng một bước chuẩn hóa thứ hai, trừ đi giá trị trung bình và chia cho độ lệch chuẩn của bộ dữ liệu ImageNet. Việc sử dụng các giá trị thống kê của ImageNet là cần thiết vì mô hình của chúng tôi sử dụng các trọng số đã được pre-train trên bộ dữ liệu này, và bước chuẩn hóa này đảm bảo rằng phân phối của dữ liệu đầu vào mới phù hợp với những gì mô hình đã học từ ImageNet.

3.4. Nhiễu Tổng hợp (Synthetic Corruptions)

Natural domain shift giữa CheXpert và NIH-14 cung cấp một baseline quan trọng để đánh giá khả năng thích ứng với sự khác biệt mang tính hệ thống giữa các cơ sở y tế. Tuy nhiên, trong môi trường lâm sàng thực tế, các mô hình AI còn phải đối mặt với những thách thức mang tính tạm thời và không thể đoán trước, đó là sự suy giảm chất lượng hình ảnh do nhiễu. Để tạo ra một kịch bản đánh giá toàn diện hơn, mô phỏng gần nhất với các điều kiện triển khai thực tế, chúng tôi đã xây dựng một luồng dữ liệu kiểm thử động bằng cách áp dụng một tập hợp các nhiễu tổng hợp (synthetic corruptions) lên bộ dữ liệu NIH-14.

Việc sử dụng nhiễu tổng hợp cho phép chúng tôi kiểm soát và định lượng được mức độ shift một cách chính xác, đồng thời đánh giá được sự bền vững (robustness) của các phương pháp TTA khi đối mặt với nhiều loại suy giảm chất lượng khác nhau. Chúng tôi đã lựa chọn cẩn thận 6 loại nhiễu phổ biến, mỗi loại mô phỏng một dạng lỗi thực tế có thể xảy ra trong quy trình chẩn đoán hình ảnh:

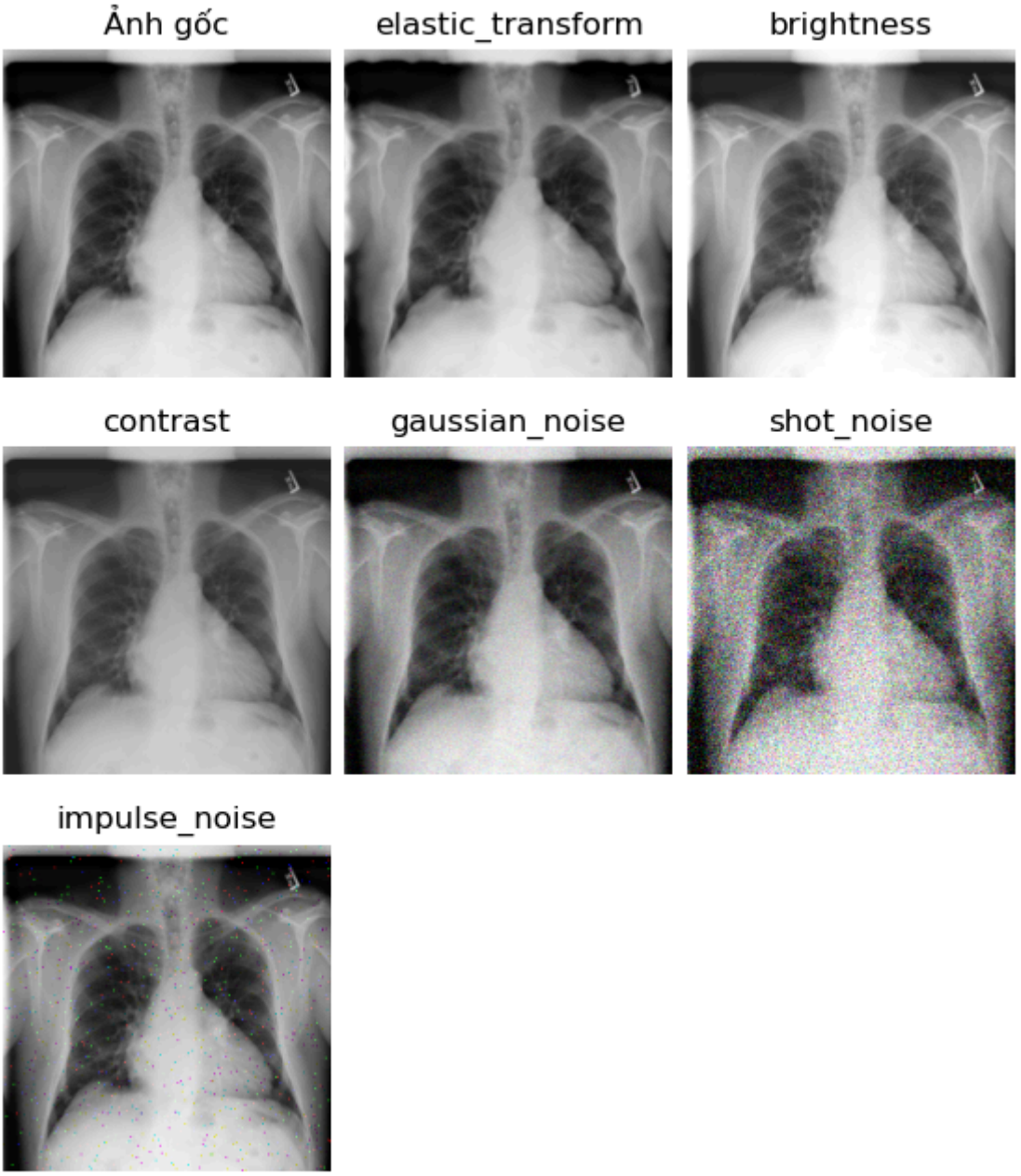
- Elastic Transform (Biến dạng Đàn hồi):

    - Bản chất: Phép biến đổi này áp dụng một trường dịch chuyển cục bộ lên hình ảnh, làm cho các cấu trúc bị co giãn và cong vênh một cách phi tuyến tính.
    - Mô phỏng thực tế: Đây là một loại nhiễu rất phù hợp với dữ liệu y tế. Nó không chỉ mô phỏng sự khác biệt nhỏ về mặt giải phẫu giữa các bệnh nhân mà còn tái tạo được sự biến dạng của các mô mềm do tư thế chụp không hoàn toàn chuẩn hoặc do các chuyển động sinh lý như hít thở.
- Contrast (Độ tương phản) & Brightness (Độ sáng):
    - Bản chất: Hai phép biến đổi này thay đổi trực tiếp sự phân bố cường độ của các pixel. Contrast điều chỉnh sự khác biệt giữa các vùng sáng và tối, trong khi Brightness làm cho toàn bộ hình ảnh trở nên sáng hơn hoặc tối hơn.
    - Mô phỏng thực tế: Đây là những vấn đề rất phổ biến, mô phỏng trực tiếp sự thay đổi trong các thông số phơi sáng của máy chụp X-quang. Một hình ảnh bị cháy sáng (over-exposed) hoặc quá tối (under-exposed) có thể làm ẩn đi hoặc làm sai lệch các chi tiết bệnh lý quan trọng.
- Gaussian Noise (Nhiễu Gauss):
    - Bản chất: Thêm nhiễu ngẫu nhiên được lấy từ phân phối Gauss vào mỗi pixel, tạo ra hiệu ứng lấm tấm trên toàn ảnh.
    - Mô phỏng thực tế: Đây là loại nhiễu điện tử phổ biến nhất, phát sinh từ các thành phần điện tử của cảm biến hình ảnh (image sensor) do các yếu tố như nhiệt độ hoặc mức độ khuếch đại tín hiệu thấp.
- Shot Noise & Impulse Noise (Nhiễu Xung):
    - Bản chất: Shot Noise (hay nhiễu Poisson) mô phỏng sự biến thiên thống kê của các photon tia X. Impulse Noise (hay nhiễu muối

tiêu) thay thế ngẫu nhiên một số pixel bằng các giá trị cực đại (trắng) hoặc cực tiểu (đen).

- Mô phỏng khi đã thích ứng, để yêu cầu sự xem xét của chuyên gia con người, tăng cường sự an toàn khi triển khai trong lâm sàng.

Ảnh minh họa ảnh gốc và các loại nhiễu tương ứng



3.5. Phân tích Định lượng về Distribution Shift

Để xác thực tính thực tế và độ khó của các kịch bản shift tổng hợp, chúng tôi đã tiến hành phân tích các thuộc tính thống kê ở cấp độ pixel của các hình ảnh bị nhiễu. Bảng \ref{tab:data_distribution_vi} so sánh phân phối pixel (Giá trị Trung bình - Mean và Độ lệch chuẩn - Std) của các bộ dữ liệu y tế trong nghiên cứu này với các benchmark tiêu chuẩn như CIFAR-10-C.

Bảng: Phân phối thống kê (Mean và Std) của các bộ dữ liệu. CIFAR-10-C được dùng làm tham chiếu về độ lớn của distribution shift.

Dataset / Kịch bản	Mean	Std
Natural Image Benchmarks (Tham chiếu):		
CIFAR-10 (Sạch)	0.4731	0.2563
CIFAR-10-C (Tất cả nhiễu, Mức 5.0)	0.5298	0.2393
Kịch bản Hình ảnh Y tế của chúng tôi:		
CheXpert (Nguồn, Sạch)	0.5034	0.2888
NIH-14 (Đích, Sạch)	0.4885	0.2473
NIH-14 + Tất cả nhiễu (Mức 5.0)	0.5008	0.2825

Phân tích này cho thấy rằng việc áp dụng các corruption nhân tạo lên ảnh CXR gây ra những distribution shift có độ lớn tương đương, thậm chí vượt qua cả những shift được quan sát trong các benchmark tiêu chuẩn. Ví dụ, chỉ riêng nhiễu Shot Noise đã làm tăng độ lệch chuẩn của ảnh CXR lên một lượng lớn hơn khoảng 26 lần so với tác động của nó trên CIFAR-10. Điều này khẳng định rằng thiết lập thực nghiệm của chúng tôi cung cấp một môi trường đầy thách thức nhưng vẫn thực tế để đánh giá các phương pháp TTA.

Bảng \ref{tab:per_corruption_stats_full_vi} cung cấp một cái nhìn chi tiết hơn về tác động đa dạng của từng loại nhiễu.

- Nhiễu Cộng tính (Additive Noise) như Shot Noise chủ yếu làm tăng phương sai của pixel, thể hiện qua sự gia tăng đáng kể của Std trong khi Mean thay đổi rất ít.

- Nhiễu về Độ sáng/Tương phản (Brightness/Contrast) gây ra những thay đổi thống kê mạnh mẽ nhất. Brightness làm tăng mạnh giá trị Mean (gây cháy sáng), trong khi Contrast làm giảm Std, làm mất chi tiết.
- Nhiễu làm mờ (Blur) thường làm giảm cả Mean và Std khi các chi tiết hình ảnh được làm mịn.
- Nhiễu Kỹ thuật số & Biến dạng (Digital & Distortion) như Elastic Transform có tác động không đáng kể đến các thống kê bậc một này. Đây là một phát hiện quan trọng, cho thấy hạn chế của việc chỉ dựa vào Mean/Std để đo lường domain shift. Các loại nhiễu này phá vỡ cấu trúc không gian của hình ảnh mà không làm thay đổi đáng kể phân phối pixel tổng thể, nhưng vẫn có thể ảnh hưởng nặng nề đến hiệu suất của mô hình.

Phân tích này xác nhận sự phức tạp của các distribution shift được đánh giá, đòi hỏi các phương pháp TTA phải đủ bền vững để xử lý nhiều loại thay đổi khác nhau, không chỉ là những thay đổi đơn giản về Mean và Std.

Bảng: Phân phối thống kê chi tiết (Mean và Std) cho các loại nhiễu được sử dụng trên NIH-14, so sánh với CIFAR-10 (Mức 5.0).

Nhóm Nhiễu	Loại Nhiễu	NIH-14 (CXR)		CIFAR-10	
		Mean	Std	Mean	Std
	Baseline Sạch	0.4885	0.2473	0.4766	0.2504
Độ sáng & Tương phản	Brightness	0.8922	0.1562	0.6962	0.2333
	Contrast	0.4885	0.0950	0.4746	0.1391
Nhiễu Cộng tính	Gaussian Noise	0.4909	0.2594	0.4744	0.2628
	Shot Noise	0.4883	0.5366	0.4717	0.2615
	Impulse Noise	0.4888	0.2655	0.4782	0.2755
Biến dạng	Elastic Transform	0.4885	0.2468	0.4741	0.2402

4. Phương pháp (Methods)

Trọng tâm của nghiên cứu này là phát triển và đánh giá một phương pháp thích ứng mới, RoTTA-ML. Tuy nhiên, để có một phép so sánh công bằng và ý nghĩa,

điều kiện tiên quyết là phải bắt đầu từ một mô hình cơ sở (base model) mạnh mẽ, được huấn luyện một cách bài bản. Phần này mô tả chi tiết quy trình xây dựng và huấn luyện mô hình cơ sở của chúng tôi trên miền dữ liệu nguồn.

4.1. Huấn luyện Mô hình Cơ sở (Base Model Fine-tuning)

Việc lựa chọn kiến trúc mô hình là một sự cân bằng giữa hiệu suất (performance) và hiệu quả tính toán (computational efficiency). Thay vì chọn các kiến trúc lớn và phức tạp như ResNet-50 hay DenseNet-121, chúng tôi đã quyết định sử dụng MobileNetV3-Small \cite{howard2019searching}. Lựa chọn này dựa trên hai lý do chính:

- Hiệu quả tính toán: MobileNetV3 được thiết kế đặc biệt cho các thiết bị có tài nguyên hạn chế, sử dụng các kỹ thuật như depth-wise separable convolutions và các khối inverted residual bottleneck với cơ chế squeeze-and-excitation. Điều này làm cho nó trở thành một ứng cử viên lý tưởng cho các ứng dụng y tế trong thực tế, nơi mô hình có thể cần được triển khai trên các máy tính cục bộ hoặc các hệ thống nhúng, và tốc độ inference là một yếu tố quan trọng.
- Khả năng Biểu diễn Mạnh mẽ: Mặc dù nhẹ, MobileNetV3 vẫn duy trì được khả năng học các biểu diễn đặc trưng mạnh mẽ. Việc sử dụng kỹ thuật tìm kiếm kiến trúc mạng (Neural Architecture Search - NAS) giúp nó đạt được sự cân bằng tối ưu giữa độ trễ và độ chính xác.

Mô hình được khởi tạo với các trọng số đã được pre-train trên bộ dữ liệu ImageNet-1K. Kỹ thuật học chuyển giao (transfer learning) này cho phép mô hình tận dụng các đặc trưng hình ảnh bậc thấp (như cạnh, góc, màu sắc) đã được học từ hàng triệu ảnh tự nhiên, giúp quá trình huấn luyện trên dữ liệu y tế hội tụ nhanh hơn và đạt được kết quả tốt hơn.

Mô hình MobileNetV3-Small sau đó được fine-tune trên bộ dữ liệu CheXpert (miền nguồn). Lớp classifier cuối cùng của mô hình gốc, vốn được thiết kế cho

1000 lớp của ImageNet, đã được loại bỏ và thay thế bằng một lớp fully-connected mới với 5 nơ-ron đầu ra, tương ứng với 5 bệnh lý mục tiêu của chúng tôi. Quá trình huấn luyện tuân theo các thiết lập sau:

- Hàm loss: Chúng tôi sử dụng hàm BCEWithLogitsLoss (Binary Cross-Entropy with Logits). Đây là hàm loss tiêu chuẩn và phù hợp nhất cho bài toán phân loại đa nhãn. Nó tính toán loss cho từng nhãn một cách độc lập, cho phép mô hình học cách nhận diện sự hiện diện hoặc vắng mặt của nhiều bệnh cùng lúc.
- Trình tối ưu hóa (Optimizer): Chúng tôi sử dụng trình tối ưu hóa Adam, một lựa chọn phổ biến và hiệu quả, với tốc độ học (learning rate) ban đầu được đặt là 1e-5 và kích thước batch là 64.
- Chiến lược Huấn luyện: Mô hình được huấn luyện trong một số lượng lớn các bước (steps). Để tránh overfitting và tìm ra điểm hội tụ tốt nhất, chúng tôi sử dụng một chiến lược giảm tốc độ học linh hoạt và dừng sớm (early stopping) dựa trên hiệu suất trên tập validation.

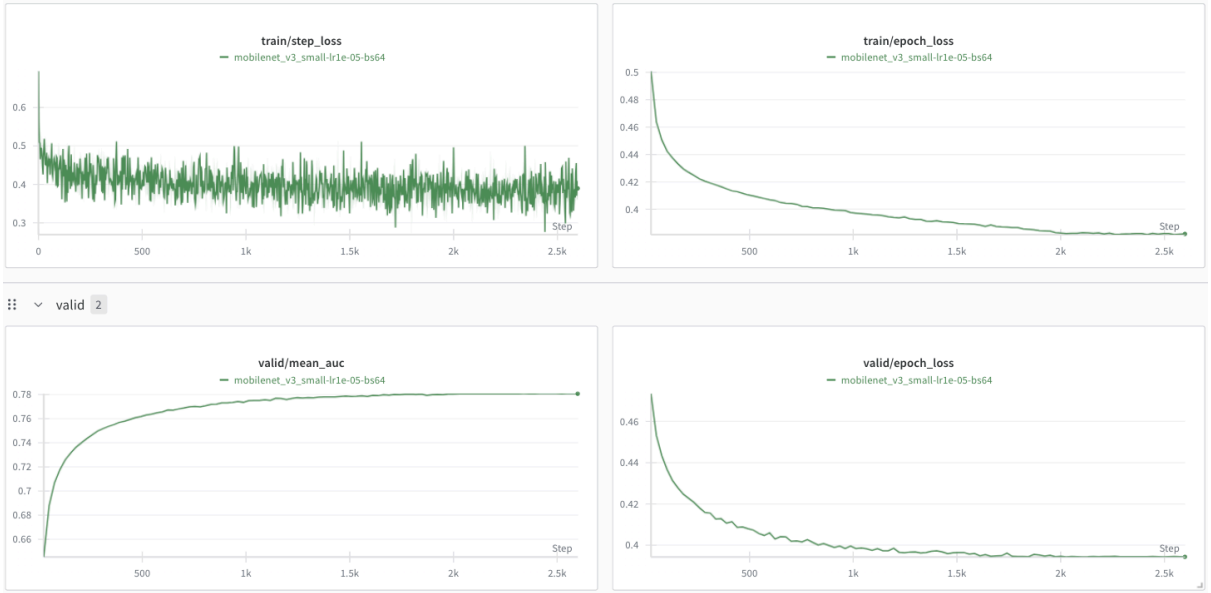
Hình dưới đây minh họa chi tiết quá trình huấn luyện và kiểm định (validation) của mô hình cơ sở. Các biểu đồ này cung cấp những bằng chứng rõ ràng về một quá trình huấn luyện thành công và ổn định.

- Loss trên tập Huấn luyện (Train Loss): Biểu đồ train/step_loss cho thấy loss trên từng batch dao động trong một khoảng hẹp và không có xu hướng tăng, cho thấy quá trình học ổn định. Biểu đồ train/epoch_loss thể hiện sự hội tụ rõ ràng, với loss giảm mạnh trong khoảng 500 steps đầu tiên và sau đó giảm từ từ, cho thấy mô hình đang học hiệu quả các đặc trưng từ dữ liệu.
- Loss và AUC trên tập Kiểm định (Validation Loss & AUC): Các đường cong trên tập validation là chỉ báo quan trọng nhất về khả năng tổng quát hóa của mô hình. Biểu đồ valid/epoch_loss cho thấy loss giảm nhanh và hội tụ về một giá trị rất thấp (dưới 0.40), song song với đó,

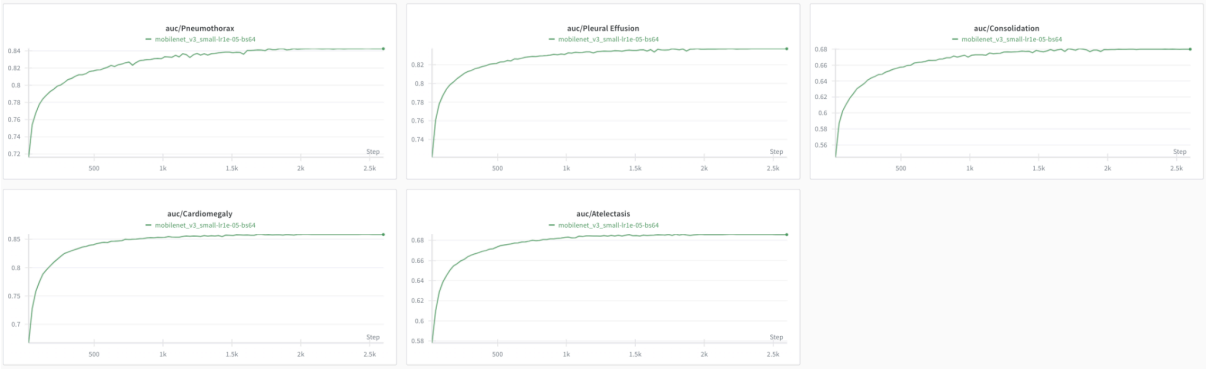
valid/mean_auc tăng nhanh và đạt đến trạng thái bão hòa (plateau) ở mức xấp xỉ 0.78. Sự hội tụ mượt mà của cả hai chỉ số này, không có dấu hiệu của overfitting (ví dụ: validation loss tăng trở lại), khẳng định rằng mô hình đã học được các đặc trưng có ý nghĩa và có khả năng tổng quát hóa tốt.

Lựa chọn mô hình cơ sở: Dựa trên các kết quả này, chúng tôi đã chọn checkpoint (trọng số) của mô hình tại step có chỉ số Mean AUC trên tập validation cao nhất làm mô hình cơ sở cuối cùng. Mô hình này đại diện cho trạng thái tốt nhất mà chúng ta có thể đạt được chỉ bằng cách sử dụng dữ liệu từ miền nguồn, và nó sẽ là điểm xuất phát cho tất cả các thí nghiệm thích ứng sau này. Phân tích chi tiết trên từng lớp bệnh (Hình \ref{fig:per_class_auc}) cũng cho thấy sự hội tụ ổn định và hiệu suất tốt trên cả 5 bệnh lý mục tiêu.

Hình 1: Các biểu đồ thể hiện quá trình huấn luyện và kiểm định. Hàng trên: loss trên tập huấn luyện theo từng bước và epoch. Hàng dưới: Mean AUC và loss trên tập kiểm định theo epoch.



Hình 2: Hiệu suất AUC trên từng lớp bệnh trong quá trình kiểm định



4.2. RoTTA-ML: Tái thiết kế RoTTA cho Đa nhãn

Mặc dù triết lý của RoTTA là rất mạnh mẽ, việc áp dụng nó vào bài toán đa nhãn đòi hỏi một sự tái thiết kế sâu sắc thay vì chỉ là một sự điều chỉnh bề mặt. RoTTA-ML giữ lại kiến trúc Teacher-Student và Robust Batch Normalization (RBN)—hai thành phần vốn đã độc lập với dạng bài toán—nhưng thay đổi hoàn toàn hai trụ cột còn lại: memory bank và mục tiêu học (learning objective).

4.2.1. Memory Bank Đa nhãn (MultiLabel-CSTU): Trái tim Thích ứng Mới

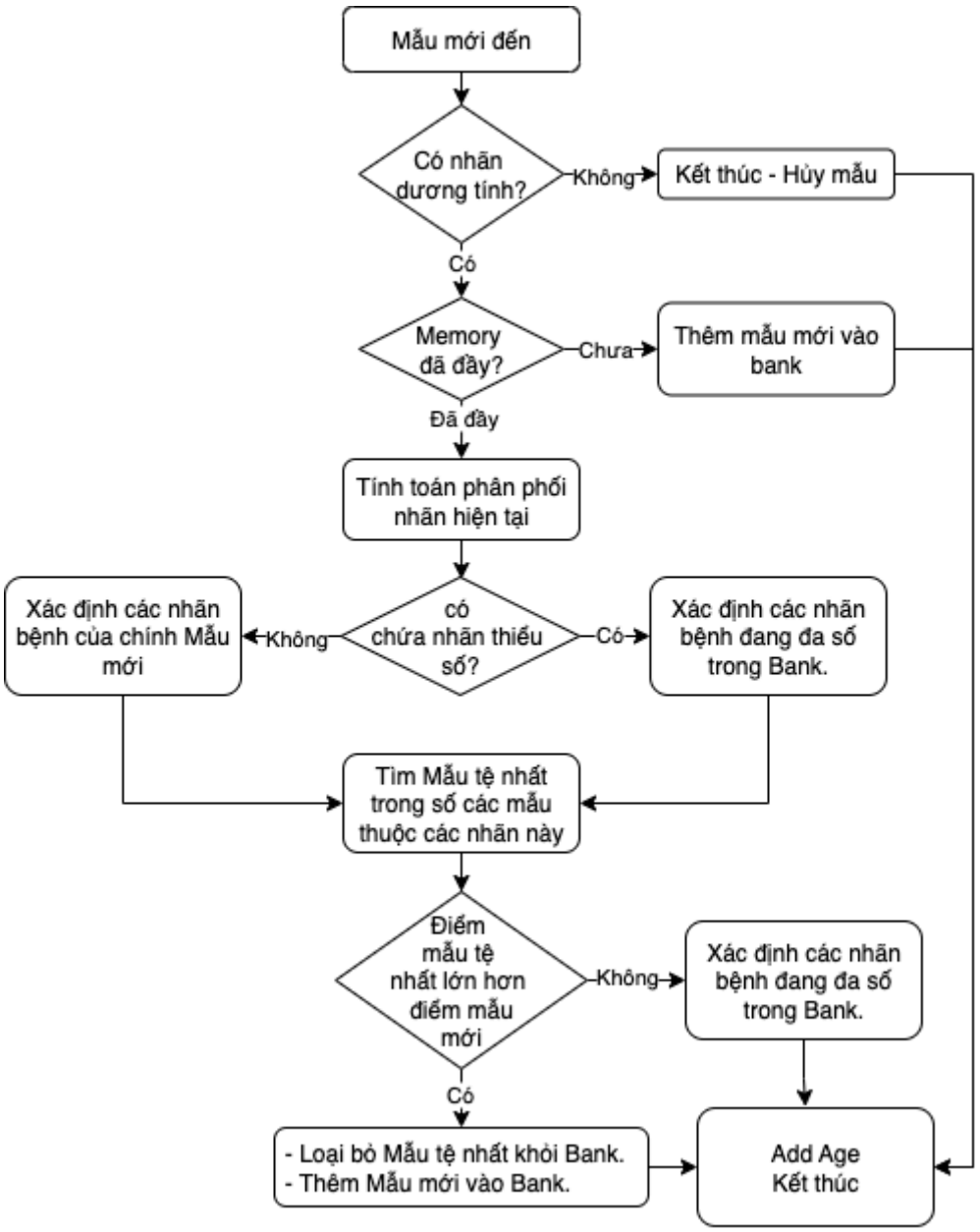
Memory bank là trái tim của RoTTA, hoạt động như một bộ đệm tri thức để lọc nhiễu và cân bằng phân phối. Cơ chế CSTU gốc, với cấu trúc phân cấp theo lớp, hoàn toàn sụp đổ trong bối cảnh đa nhãn. Do đó, chúng tôi đã thiết kế MultiLabel-CSTU từ đầu để giải quyết các thách thức của dữ liệu đa nhãn.

- Cấu trúc Dữ liệu Phẳng và Linh hoạt: Thay vì một danh sách các danh sách, MultiLabel-CSTU sử dụng một danh sách phẳng duy nhất chứa các đối tượng MemoryItem. Mỗi MemoryItem là một cấu trúc dữ liệu lưu trữ không chỉ tensor ảnh mà còn cả vector pseudo-label đa nhãn, điểm uncertainty và tuổi của nó. Cấu trúc này loại bỏ hoàn toàn ràng buộc mỗi mẫu chỉ thuộc một lớp, cho phép lưu trữ bất kỳ tổ hợp nhãn nào.
- Thuật toán Cân bằng Nhãn Động: Thách thức lớn nhất là làm thế nào để duy trì sự cân bằng khi không còn khái niệm lớp của một mẫu.

MultiLabel-CSTU giải quyết vấn đề này bằng một thuật toán cân bằng động:

- Theo dõi theo Từng nhãn: Thay vì theo dõi số lượng mẫu theo lớp, bank duy trì một vector đếm số lượng cho từng nhãn bệnh riêng lẻ. Vector này được tính toán lại một cách linh hoạt mỗi khi cần đưa ra quyết định.
- Điểm số Heuristic: Quyết định loại bỏ mẫu nào vẫn dựa trên một điểm số heuristic, là tổng trọng số của tuổi (Timeliness) và độ bất định (Uncertainty). Một điểm số cao cho thấy một mẫu vừa cũ vừa không chắc chắn, là ứng cử viên hàng đầu để bị loại bỏ.
- Chiến lược Thay thế Thông minh: Khi memory bank đã đầy và một mẫu mới đến, thuật toán sẽ thực hiện một chiến lược thay thế tinh vi (minh họa trong Hình \ref{fig:cstu_logic}):
  - Bảo vệ Thiểu số: Nếu mẫu mới chứa ít nhất một nhãn bệnh đang có số lượng thấp trong bank (dưới một ngưỡng mong muốn), mẫu này được coi là quý giá vì nó giúp tăng sự đa dạng. Thuật toán sẽ tìm kiếm trong số các mẫu thuộc các nhãn đang chiếm đa số để loại bỏ mẫu có điểm heuristic cao nhất, nhằm nhường chỗ.
  - Kiểm soát Đa số: Ngược lại, nếu mẫu mới chỉ chứa các nhãn đã được biểu diễn tốt, thuật toán sẽ chỉ tìm kiếm và loại bỏ một mẫu cũ từ chính các nhãn đó, ngăn chặn sự dư thừa.

Chiến lược này không chỉ cân bằng số lượng mẫu trên từng nhãn mà còn chủ động tìm kiếm và bảo vệ sự đa dạng của các lớp bệnh hiếm, một vấn đề cực kỳ quan trọng trong dữ liệu y tế.



4.2.2. Mục tiêu Học nhất quán cho Đa nhãn

Để hướng dẫn quá trình học của Student model một cách hiệu quả trong không gian đa nhãn, chúng tôi đã xây dựng lại toàn bộ quy trình tạo tín hiệu giám sát.

- Tạo Nhãn giả và Uncertainty: Mô hình Teacher (phiên bản ổn định, cập nhật chậm qua EMA) được sử dụng để tạo tín hiệu giám sát. Với một batch dữ liệu mới, đầu ra logits của Teacher được đưa qua hàm Sigmoid để tạo ra một vector xác suất p_T

    - Nhãn giả Mềm (Soft Pseudo-label): Vector xác suất p_T này đóng vai trò là nhãn giả mềm, thể hiện độ tự tin của Teacher trên từng bệnh lý. Nó được sử dụng trực tiếp để tính toán consistency loss.
    - Nhãn giả Cứng (Hard Pseudo-label): Một ngưỡng (threshold, ví dụ 0.5) được áp dụng lên p_T để tạo ra một vector nhị phân (0 hoặc 1). Vector này được dùng làm nhãn để cập nhật trạng thái cân bằng của MultiLabel-CSTU.
    - Đo lường Uncertainty: Độ bất định của một dự đoán không còn có thể được đo bằng softmax entropy. Thay vào đó, chúng tôi định nghĩa nó là tổng của binary entropy trên tất cả các nhãn. Giá trị này đạt cực đại khi các xác suất dự đoán gần 0.5 (không chắc chắn) và cực tiểu khi chúng gần 0 hoặc 1 (tự tin), phản ánh chính xác độ tự tin tổng thể của mô hình trên toàn bộ các bệnh lý.
- Hàm Loss nhất quán (Consistency Loss): Khi huấn luyện Student model trên dữ liệu lấy từ memory bank, chúng tôi sử dụng một hàm loss dựa trên Binary Cross-Entropy (BCE). Mục tiêu là tối thiểu hóa sự khác biệt giữa dự đoán (logits) của Student model (nhận đầu vào là ảnh đã được strong augmentation) và nhãn giả mềm (xác suất) được tạo ra bởi Teacher model. Toàn bộ loss của một mẫu sau đó được trọng số hóa theo tuổi (timeliness reweighting), giúp giảm ảnh hưởng của các mẫu cũ và có thể đã lỗi thời trong memory bank. Cách tiếp cận này buộc Student model phải học các biểu diễn đặc trưng bền vững, bất biến với các phép biến đổi, thay vì chỉ học vẹt các mẫu.

4.3. Phương pháp So sánh: TENT-ML

Để đánh giá một cách khách quan hiệu quả của RoTTA-ML, việc so sánh nó với một baseline mạnh và có liên quan là rất cần thiết. Thay vì chỉ so sánh với trường hợp không thích ứng (Source-only), chúng tôi đã triển khai và điều chỉnh TENT (Test-Time Entropy Minimization) \cite{wang2020tent}, một trong

những phương pháp TTA kinh điển và có ảnh hưởng nhất, cho bài toán đa nhãn. Phiên bản này được chúng tôi gọi là TENT-ML.

4.3.1. Nguyên lý của TENT

TENT hoạt động dựa trên một giả định đơn giản nhưng mạnh mẽ: một mô hình được huấn luyện tốt nên đưa ra các dự đoán có độ tự tin cao (tức là có entropy thấp) trên dữ liệu kiểm thử. Do đó, mục tiêu của TENT là tối ưu hóa các tham số của mô hình tại test time để tối thiểu hóa entropy của các dự đoán đầu ra.
Một ưu điểm lớn của TENT là sự đơn giản và hiệu quả. Nó không cần một mô hình Teacher phức tạp hay một memory bank tốn bộ nhớ. Thay vào đó, nó học trực tiếp từ từng batch dữ liệu online. Để giữ cho quá trình thích ứng nhẹ nhàng và tránh làm thay đổi các đặc trưng cốt lõi đã học, TENT chỉ cập nhật các tham số affine (γ và β) của các lớp Batch Normalization (BN) trong mạng.

4.3.2. Thách thức và Giải pháp Điều chỉnh cho Đa nhãn (TENT-ML)

TENT gốc được thiết kế cho bài toán đa lớp, sử dụng hàm loss softmax_entropy để tối ưu hóa. Hàm loss này không thể áp dụng cho đầu ra sigmoid của bài toán đa nhãn. Thách thức chính là tìm một hàm loss thay thế có thể đóng vai trò tương tự: khuyến khích sự tự tin trên từng nhãn một cách độc lập.
Để giải quyết vấn đề này, chúng tôi đề xuất một hàm loss có dạng hình chữ U cho TENT-ML. Thay vì đo lường entropy trên một phân phối xác suất, hàm loss này đo lường sự thiếu tự tin trên từng đầu ra nhị phân. Công thức được định nghĩa cho một dự đoán đầu ra P như sau:

$$\mathcal{L}_{Tent\_ML}(P) = \sum_{c=1}^{C} 4 \cdot p_c \cdot (1 - p_c)$$

Trong đó:

- C là tổng số lớp bệnh (trong trường hợp này là 5).
- P_C là xác suất dự đoán của mô hình cho lớp bệnh C, được tính bằng hàm sigmoid.

Hàm loss này có một đặc tính toán học rất phù hợp: giá trị của nó đạt cực tiểu (bằng 0) khi xác suất P_C tiến về 0 hoặc 1 (tức là mô hình rất tự tin rằng bệnh đó vắng mặt hoặc hiện diện), và đạt cực đại (bằng 1) khi xác suất P_C bằng 0.5 (mô hình không chắc chắn nhất). Bằng cách tối thiểu hóa tổng của các giá trị này trên tất cả các lớp bệnh và trên toàn bộ batch, chúng tôi buộc mô hình phải đưa ra quyết định dứt khoát hơn cho mỗi bệnh, đẩy các xác suất dự đoán ra xa khỏi vùng không chắc chắn 0.5. Tương tự như TENT gốc, quá trình tối ưu hóa này chỉ được áp dụng lên các tham số affine (γ và β) của các lớp Batch Normalization.

Việc lựa chọn TENT-ML làm phương pháp so sánh là rất có ý nghĩa. Nó đại diện cho trường phái TTA nhẹ, không có bộ nhớ, và dựa trên giả định về sự tự tin của dự đoán. So sánh RoTTA-ML với TENT-ML sẽ cho phép chúng tôi trả lời câu hỏi: Liệu các cơ chế phức tạp hơn như memory bank và Teacher-Student có thực sự cần thiết và mang lại lợi ích vượt trội so với một phương pháp tối ưu hóa trực tiếp và đơn giản hơn trong bối cảnh chẩn đoán CXR đa nhãn hay không?

5. Thí nghiệm (Experiments)

Để đánh giá một cách toàn diện và khách quan hiệu quả của RoTTA-ML, chúng tôi đã thiết kế một loạt các thí nghiệm có kiểm soát. Phần này sẽ trình bày chi tiết các kịch bản đánh giá, các kết quả định lượng, và những phân tích chuyên sâu rút ra từ các kết quả đó. Mục tiêu của chúng tôi không chỉ là chứng minh sự vượt trội của phương pháp đề xuất mà còn là để hiểu rõ hơn về hành vi, điểm mạnh và điểm yếu của các chiến lược TTA khác nhau khi áp dụng vào một bài toán thực tế phức tạp.

5.1. Kịch bản 1: Thích ứng với Natural Domain Shift (Dữ liệu Sạch)

Thí nghiệm đầu tiên tập trung vào việc đánh giá khả năng của các phương pháp TTA trong việc xử lý natural domain shift — sự khác biệt tự nhiên, vốn có giữa hai bộ dữ liệu được thu thập độc lập.

- Mục tiêu: Đo lường xem các phương pháp có thể cải thiện hiệu suất của mô hình khi đối mặt với dữ liệu từ miền đích (NIH-14) ở trạng thái sạch (không có nhiễu tổng hợp) hay không. Kịch bản này mô phỏng trường hợp triển khai mô hình từ bệnh viện A (CheXpert) sang bệnh viện B (NIH-14) với điều kiện hình ảnh lý tưởng.
- Thiết lập: Mô hình cơ sở (đã fine-tune trên CheXpert) được áp dụng trên toàn bộ tập kiểm thử của NIH-14. Các phương pháp RoTTA-ML và TENT-ML được kích hoạt để thích ứng liên tục trên luồng dữ liệu này.

Bảng 1 so sánh hiệu suất của các phương pháp trên bộ dữ liệu NIH-14 sạch.

Phương pháp	**Mean AUC**	**Δ%**	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Pleural E.**	**Pneumothorax**
Source-only	0.7174	---	0.6770	0.8003	0.6200	0.7363	0.7531
TENT-ML	0.6699	-6.62%	0.6404	0.7294	0.5912	0.7079	0.6805
RoTTA-ML	0.7373	+2.77%	0.6896	0.8299	0.6313	0.7551	0.7806

Phân tích Kết quả:

- **Xác nhận sự tồn tại của Natural Domain Shift:** Hiệu suất của Source-only trên NIH-14 (Mean AUC = 0.7174) thấp hơn đáng kể so với hiệu suất Oracle của nó trên CheXpert (Mean AUC = 0.7765). Khoảng cách hiệu suất ~0.06 AUC này khẳng định rằng có một sự khác biệt đáng

kể về mặt phân phối dữ liệu giữa hai bệnh viện, ngay cả khi không có nhiễu.

- **Hiệu quả vượt trội của RoTTA-ML:** Phương pháp RoTTA-ML đã chứng tỏ khả năng thích ứng rất hiệu quả. Nó không chỉ cải thiện Mean AUC tổng thể một cách đáng kể (tăng 2.77%) mà còn cải thiện hiệu suất trên tất cả 5 lớp bệnh riêng lẻ. Điều này cho thấy RoTTA-ML đã học được các đặc trưng phù hợp hơn với phân phối của NIH-14, giúp nó đưa ra chẩn đoán chính xác hơn trên toàn bộ các bệnh lý. Sự thành công này cho thấy cơ chế memory bank và Teacher-Student là rất hiệu quả trong việc định hướng quá trình học một cách ổn định và đúng đắn.
- **Sự Thất bại của TENT-ML:** Trái ngược hoàn toàn, TENT-ML lại cho thấy sự suy giảm hiệu năng nghiêm trọng, với Mean AUC giảm tới -6.62% so với Source-only. Kết quả này là một lời cảnh báo quan trọng: việc tối thiểu hóa entropy một cách ngây thơ trong bối cảnh đa nhãn có thể rất nguy hiểm. Có thể do sự mất cân bằng tự nhiên của các nhãn trong NIH-14, việc buộc mô hình phải tự tin trên từng batch đã dẫn đến việc nó học lệch, củng cố những dự đoán sai ban đầu và gây ra hiện tượng tích lũy lỗi, cuối cùng làm sụp đổ hiệu năng tổng thể.

Kết luận của Thí nghiệm 1: Ngay cả trong điều kiện lý tưởng (dữ liệu sạch), domain shift vẫn là một vấn đề lớn. Các phương pháp TTA phức tạp và bền vững như RoTTA-ML là cần thiết để thích ứng hiệu quả, trong khi các phương pháp đơn giản hơn như TENT-ML có thể phản tác dụng và gây hại cho hiệu suất.

5.2. Thí nghiệm 2: Thích ứng trong Môi trường Nhiễu động

Thí nghiệm này là bài kiểm tra khắc nghiệt nhất, được thiết kế để đánh giá sự bền vững (robustness) và khả năng thích ứng linh hoạt của các phương pháp TTA khi đối mặt với một môi trường thay đổi liên tục và không thể đoán trước.

- Mục tiêu: Đánh giá khả năng của các phương pháp trong việc duy trì hiệu suất khi phân phối dữ liệu đầu vào thay đổi đột ngột và thường xuyên. Kịch bản này mô phỏng các điều kiện triển khai thực tế nhất, nơi chất lượng hình ảnh có thể bị suy giảm do nhiều yếu tố khác nhau.
- Thiết lập: Chúng tôi tạo ra một luồng dữ liệu kiểm thử động bằng cách tuần tự áp dụng 6 loại nhiễu hình ảnh khác nhau (elastic transform, contrast, brightness, gaussian noise, shot noise, impulse noise) lên các batch dữ liệu liên tiếp từ tập kiểm thử NIH-14.

5.2.1. Phân tích Hiệu suất Tổng thể

Bảng 2 tóm tắt hiệu suất Mean AUC trung bình của các phương pháp trên toàn bộ luồng dữ liệu nhiễu.

Phương pháp	Mean AUC	Δ% so với Source-only
Source-only	0.6389	---
TENT-ML	0.5588	-12.54%
RoTTA-ML (BN only)	0.6608	+3.43%
RoTTA-ML (BN+FC)	0.6615	+3.54%

Phân tích Kết quả:

- Tác động Hủy diệt của Nhiễu: Sự hiện diện của nhiễu đã làm suy giảm nghiêm trọng hiệu suất của mô hình cơ sở. Source-only giảm từ 0.7174 (trên dữ liệu sạch) xuống chỉ còn 0.6389, mất gần 8 điểm AUC. Điều này khẳng định rằng các mô hình học sâu rất nhạy cảm với các corruption mà chúng chưa từng thấy trong quá trình huấn luyện.
- Sự Bền vững Vượt trội của RoTTA-ML: Trong môi trường khắc nghiệt này, RoTTA-ML không chỉ sống sót mà còn phát triển mạnh mẽ. Phiên bản cập nhật cả lớp BN và lớp classifier (BN+FC) đã cải thiện hiệu suất

thêm +3.54% so với Source-only, thu hẹp đáng kể khoảng cách hiệu suất. Điều này là minh chứng thuyết phục cho sức mạnh tổng hợp của cả ba trụ cột:

    - Memory Bank (MultiLabel-CSTU): Đóng vai trò như một bộ lọc, chỉ giữ lại các mẫu đáng tin cậy và cân bằng, bảo vệ mô hình khỏi việc học từ các batch dữ liệu bị nhiễu nặng hoặc mất cân bằng.
    - Teacher-Student: Cung cấp các pseudo-label ổn định, làm giảm nguy cơ học sai từ những dự đoán nhất thời của mô hình khi đối mặt với một loại nhiễu mới.
    - Robust Batch Normalization (RBN): Đảm bảo việc chuẩn hóa đặc trưng luôn chính xác bằng cách dựa vào thống kê từ memory bank đáng tin cậy, thay vì từ các batch dữ liệu nhiễu loạn.
- Sự Sụp đổ của TENT-ML: Một lần nữa, TENT-ML cho thấy sự thất bại hoàn toàn, với hiệu suất giảm tới -12.54%. Trong một môi trường mà dữ liệu đầu vào liên tục thay đổi và có chất lượng kém, việc tối ưu hóa một cách mù quáng để tăng sự tự tin đã trở thành một thảm họa. Mô hình nhanh chóng bị thuyết phục bởi những dự đoán sai của chính nó, dẫn đến một vòng xoáy tích lũy lỗi không thể cứu vãn.

5.2.2. Phân tích Chi tiết trên từng loại Nhiễu

Để hiểu rõ hơn về khả năng xử lý của RoTTA-ML, Bảng sau phân tích hiệu suất trên từng loại nhiễu riêng lẻ.

Loại Nhiễu	Source-only	RoTTA-ML (BN+FC)	Δ%
Elastic Transform	0.7043	0.7214	+2.43%
Contrast	0.7030	0.7181	+2.15%
Brightness	0.6498	0.6689	+2.94%

Gaussian Noise	0.6272	0.6532	+4.15%
Impulse Noise	0.6128	0.6434	+4.99%
Shot Noise	0.5363	0.5639	+5.15%

Phân tích:

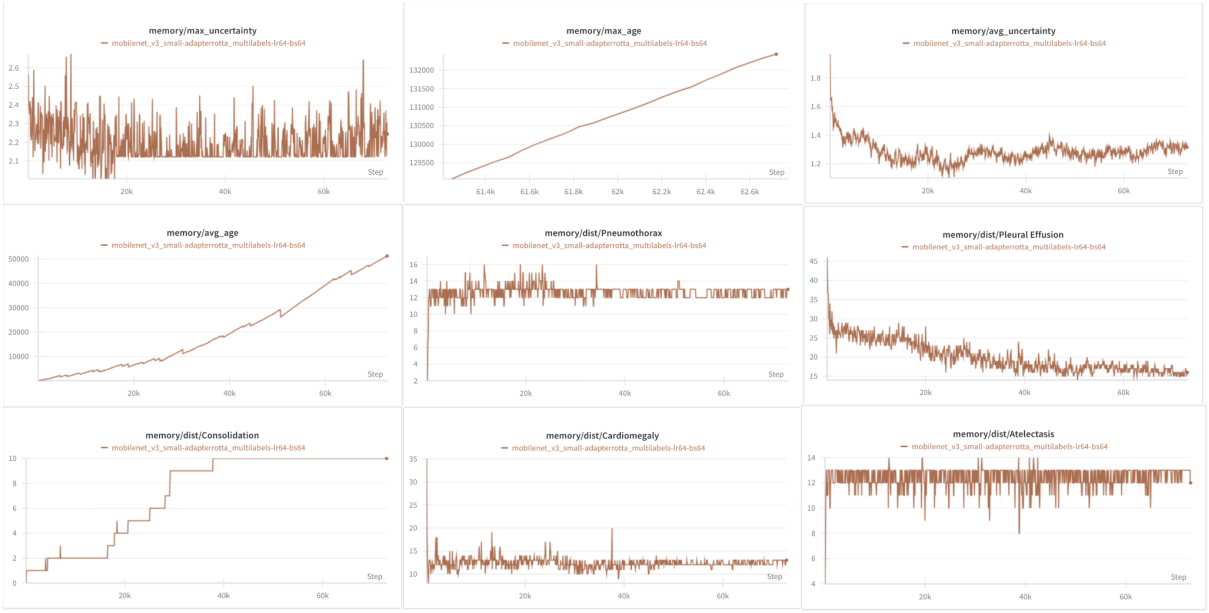
- **Cải thiện Nhất quán:** RoTTA-ML cho thấy sự cải thiện hiệu suất trên tất cả 6 loại nhiễu, chứng tỏ khả năng thích ứng toàn diện của nó.
- **Hiệu quả cao nhất trên Nhiễu nặng:** Mức độ cải thiện lớn nhất được ghi nhận ở các loại nhiễu gây suy giảm hiệu năng mạnh nhất cho mô hình Source-only, đó là Shot Noise (+5.15%) và Impulse Noise (+4.99%). Điều này cho thấy RoTTA-ML đặc biệt hiệu quả trong việc cứu vãn mô hình khỏi những cú sốc phân phối lớn. Cơ chế lọc của memory bank có thể đã đóng vai trò quan trọng trong việc loại bỏ các pixel nhiễu cực đoan, cung cấp một tín hiệu học sạch hơn cho mô hình.

Kết luận của Thí nghiệm 2: RoTTA-ML đã chứng tỏ là một phương pháp TTA rất bền vững và hiệu quả, có khả năng duy trì và cải thiện hiệu suất trong các môi trường động và đầy thách thức, nơi các phương pháp đơn giản hơn hoàn toàn thất bại.

5.3. Phân tích Chuyên sâu về Động lực học của Memory Bank

Để đi sâu hơn vào việc tại sao và như thế nào RoTTA-ML có thể thích ứng hiệu quả, chúng tôi đã tiến hành một phân tích chi tiết về các chỉ số nội tại của memory bank MultiLabel-CSTU trong suốt quá trình thích ứng. Các biểu đồ trong Hình \ref{fig:memory_analysis_vi} cung cấp một cửa sổ nhìn vào động lực học phức tạp bên trong bộ não của hệ thống.

Hình dưới đây ghi nhận các biểu đồ thể hiện động lực học của Memory Bank trong quá trình TTA.



Phân tích các biểu đồ này mang lại ba hiểu biết quan trọng

5.3.1 Quá trình Học tập và Thích ứng Thành công:

Biểu đồ memory/avg_uncertainty là minh chứng rõ ràng nhất cho một quá trình học tập hiệu quả. Trong khoảng 20,000 steps đầu tiên, độ bất định trung bình của các mẫu trong bank giảm mạnh từ mức cao (khoảng 1.9) xuống một vùng ổn định thấp hơn (khoảng 1.2-1.4). Điều này cho thấy mô hình đang nhanh chóng chuyển từ trạng thái không chắc chắn khi lần đầu tiếp xúc với miền dữ liệu mới sang trạng thái tự tin hơn sau khi đã học được các đặc trưng của phân phối đích. Việc duy trì sự ổn định sau đó cho thấy memory bank đã đạt đến một trạng thái cân bằng động, liên tục làm mới mình với các mẫu mà mô hình có độ tin cậy cao.

5.3.2. Hiệu quả của Cơ chế Cân bằng Nhãn Động:

- **Kiềm chế Lớp đa số:** Biểu đồ của Pleural Effusion cho thấy số lượng mẫu ban đầu tăng vọt lên đến hơn 45 (do đây là một bệnh lý phổ biến trong các batch đầu), nhưng sau đó đã bị cơ chế cân bằng kiềm chế và giảm dần về một mức ổn định hơn (khoảng 15-20 mẫu). Điều này chứng

tỏ thuật toán đã hoạt động hiệu quả, ngăn chặn một lớp bệnh duy nhất chiếm lĩnh memory bank và bảo vệ mô hình khỏi nguy cơ học lệch.

- **Duy trì sự Đa dạng:** Các lớp khác như Atelectasis, Cardiomegaly, và Pneumothorax duy trì một số lượng mẫu tương đối ổn định trong suốt quá trình, dao động quanh ngưỡng mong muốn (~12-14 mẫu). Điều này cho thấy bank đang tích cực duy trì sự đa dạng, đảm bảo mô hình có đủ dữ liệu để tiếp tục học về nhiều loại bệnh lý khác nhau.

5.3.3. 3. Bằng chứng về Điểm mù và Hạn chế Cố hữu:

Phân tích này cũng bộc lộ một hạn chế nghiêm trọng và là một phát hiện quan trọng của nghiên cứu:

- Điểm mù của Lớp hiếm: Biểu đồ của Consolidation kể một câu chuyện hoàn toàn khác. Số lượng mẫu của lớp này bắt đầu từ 0 và tăng lên rất chậm chạp theo một dạng bậc thang, chỉ đạt đến 10 mẫu ở cuối quá trình. Đây là bằng chứng rõ ràng về một điểm mù: mô hình Teacher ban đầu, do tác động của domain shift, gần như không thể nhận diện được lớp Consolidation với độ tự tin đủ để vượt qua ngưỡng tạo pseudo-label. Do đó, rất ít mẫu của lớp này được đưa vào bank.
- Hậu quả: Vì không có (hoặc có quá ít) dữ liệu để học, mô hình không thể cải thiện khả năng nhận diện Consolidation. Điều này giải thích một cách hoàn hảo tại sao trong Bảng kết quả, Consolidation là một trong những lớp có hiệu suất thấp nhất và mức độ cải thiện không đáng kể. Phát hiện này nhấn mạnh một hạn chế cố hữu của các phương pháp dựa trên pseudo-labeling: chúng phụ thuộc vào một mức độ nhận dạng ban đầu tối thiểu để có thể khởi động vòng lặp tự cải thiện.
- Sự tồn tại của Mẫu Bất tử: Biểu đồ memory/max_age cho thấy tuổi của mẫu cũ nhất trong bank tăng gần như tuyến tính theo thời gian. Điều này chỉ ra rằng có những mẫu, có thể là những mẫu hiếm hoi của lớp Consolidation được thêm vào sớm, đã không bao giờ bị loại bỏ. Mặc dù

việc bảo vệ lớp hiếm là có chủ đích, nhưng nó cũng tiềm ẩn rủi ro nếu những mẫu bất tử này tình cờ bị nhiễu nặng hoặc có nhãn giả sai.

Kết luận của Phân tích: MultiLabel-CSTU là một thành phần cốt lõi và hiệu quả, giúp RoTTA-ML ổn định và thích ứng thành công. Tuy nhiên, phân tích sâu cũng đã vạch ra những giới hạn quan trọng, mở ra các hướng đi rõ ràng cho các cải tiến trong tương lai.

6. Kết luận

6.1. Tóm tắt Kết quả và Đóng góp

Trong nghiên cứu này, chúng tôi đã đối mặt với một thách thức kép: giải quyết vấn đề distribution shift trong một kịch bản động (Practical TTA) cho một bài toán phức tạp là phân loại đa nhãn (multi-label) trên ảnh X-quang lồng ngực. Chúng tôi đã đề xuất và xác thực thành công RoTTA-ML, một sự tổng quát hóa toàn diện của phương pháp TTA state-of-the-art RoTTA. Thông qua việc tái thiết kế các thành phần cốt lõi như memory bank và mục tiêu học, RoTTA-ML đã chứng tỏ khả năng thích ứng hiệu quả và bền vững.

Kết quả thực nghiệm của chúng tôi đã mang lại những kết luận quan trọng:

- **Sự cần thiết của TTA Bền vững:** Chúng tôi đã định lượng được tác động nghiêm trọng của cả natural domain shift và dynamic corruptions lên hiệu suất của một mô hình AI y tế. Đồng thời, chúng tôi cũng cho thấy các phương pháp TTA đơn giản như TENT-ML có thể phản tác dụng và làm sụp đổ hiệu năng trong môi trường đa nhãn, nhiễu động, nhấn mạnh sự cần thiết của các cơ chế phòng vệ phức tạp hơn.
- **Hiệu quả của RoTTA-ML:** Phương pháp của chúng tôi đã cải thiện đáng kể hiệu suất chẩn đoán (Mean AUC) so với baseline không thích ứng trong cả hai kịch bản dữ liệu sạch và nhiễu. Sự thành công này khẳng định rằng triết lý cốt lõi của RoTTA — học từ một snapshot nhỏ, ổn định và cân bằng của thế giới — là một nguyên tắc mạnh mẽ có thể được tổng quát hóa thành công sang miền bài toán đa nhãn.

- **Vai trò Trung tâm của Memory Bank:** Phân tích chuyên sâu về động lực học của MultiLabel-CSTU đã chứng minh nó là một thành phần then chốt, hoạt động hiệu quả như một bộ lọc thông minh giúp kiềm chế các lớp đa số và duy trì sự đa dạng của các lớp bệnh trong quá trình thích ứng.

6.2. Hạn chế và Hướng phát triển Tương lai

Mặc dù đạt được những kết quả tích cực, nghiên cứu của chúng tôi cũng đã làm sáng tỏ những hạn chế cố hữu và mở ra các hướng đi quan trọng cho tương lai.

- Hạn chế Cốt lõi: Điểm mù của Pseudo-labeling: Phát hiện quan trọng nhất của chúng tôi là sự tồn tại của một điểm mù đối với các lớp bệnh cực hiếm hoặc khó (như Consolidation). Nếu mô hình Teacher ban đầu không thể nhận diện một lớp bệnh với độ tự tin tối thiểu, vòng lặp tự huấn luyện sẽ không bao giờ được khởi động cho lớp đó. Đây là một hạn chế nền tảng của các phương pháp dựa trên pseudo-labeling và là một rào cản lớn đối với việc triển khai an toàn trong y tế, nơi việc bỏ sót các bệnh hiếm có thể gây ra hậu quả nghiêm trọng.

Điều này mở ra nhiều hướng phát triển quan trọng và đầy hứa hẹn:

- Tích hợp Cơ chế Khám phá (Exploration Mechanisms):
    - Vấn đề: Quá trình hiện tại quá phụ thuộc vào khai thác (exploitation) các dự đoán tự tin.
    - Giải pháp: Các nghiên cứu trong tương lai có thể tích hợp các chiến lược khám phá (exploration). Ví dụ, hệ thống có thể định kỳ đưa các mẫu có uncertainty cao nhất vào memory bank bất kể nhãn giả của chúng, hoặc sử dụng một ngưỡng tự tin linh hoạt (hạ thấp ngưỡng cho các lớp đang bị bỏ quên). Điều này có thể giúp mô hình thoát khỏi điểm mù và học cách nhận diện các lớp bệnh mới hoặc khó.
- Cải tiến Chiến lược Quản lý Memory Bank:

    - Vấn đề: Phân tích của chúng tôi cho thấy sự tồn tại của các mẫu bất tử.
    - Giải pháp: Thiết kế các thuật toán loại bỏ mẫu thông minh hơn. Thay vì chỉ dựa trên tuổi và uncertainty, có thể xem xét các yếu tố khác như sự đa dạng về đặc trưng (feature diversity). Ví dụ, sử dụng thuật toán gom cụm trong không gian đặc trưng để đảm bảo rằng memory bank không chỉ cân bằng về nhãn mà còn bao phủ được nhiều dạng biểu hiện hình ảnh khác nhau của cùng một bệnh.
- Xây dựng Hệ thống TTA An toàn Tương tác với Con người:
    - Vấn đề: Một mô hình TTA tự động hoàn toàn vẫn tiềm ẩn rủi ro.
    - Giải pháp: Một hướng đi thực tế và có giá trị là xây dựng các hệ thống TTA có khả năng nhận thức được sự không chắc chắn của chính mình. Thay vì chỉ sử dụng uncertainty như một tín hiệu nội tại, hệ thống có thể sử dụng nó để tương tác với người dùng. Ví dụ, hệ thống có thể giơ cờ (flagging) các trường hợp mà nó vẫn còn độ uncertainty cao ngay cả sau khi đã thích ứng, và đề nghị một chuyên gia con người xem xét. Cách tiếp cận AI-in-the-loop này sẽ tăng cường đáng kể sự an toàn và độ tin cậy khi triển khai các hệ thống AI trong môi trường lâm sàng.

Bằng cách giải quyết những thách thức này, chúng ta có thể tiếp tục cải thiện sự bền vững và độ tin cậy của các hệ thống AI y tế, từng bước thu hẹp khoảng cách giữa tiềm năng nghiên cứu và ứng dụng thực tiễn.

Tài liệu tham khảo

1. N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, “Test-time adaptable neural networks for robust medical image segmentation,” Medical Image Analysis, vol. 68, p. 101907, 2021.
2. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning mod

to detect pneumonia in chest radiographs: a cross-sectional study," PLoS medicine, vol. 15, no. 11, p. e1002683, 2018.

3. D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," arXiv preprint arXiv:1903.12261, 2019.
4. D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," arXiv preprint arXiv:2006.10726, 2020.
5. L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15 922–15 932.
6. Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7201–7211.
7. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arxiv," arXiv preprint arXiv:1711.05225, vol. 10, 2017.
8. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 590–597.
9. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

10. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific data, vol. 6, no. 1, p. 317, 2019.
11. F. Liu, Y. Tian, Y. Chen, Y. Liu, V. Belagiannis, and G. Carneiro, “Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 20 697–20 706.
12. Z. Ren, X. Kong, Y. Zhang, and S. Wang, “Ukssl: Underlying knowledge based semi-supervised learning for medical image classification,” IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 459–466, 2023.
13. M. Kim, K.-R. Moon, and B.-D. Lee, “Unsupervised anomaly detection for posteroanterior chest x-rays using multiresolution patch-based self-supervised learning,” Scientific Reports, vol. 13, no. 1, p. 3415, 2023.
14. H. Guan and M. Liu, “Domain adaptation for medical image analysis: a survey,” IEEE Transactions on Biomedical Engineering, vol. 69, no. 3, pp. 1173–1185, 2021.
15. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
16. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
17. S. Tayebi Arasteh, M. Lotfinia, T. Nolte, M.-J. S¨ahn, P. Isfort, C. Kuhl, S. Nebelung, G. Kaissis, and D. Truhn, “Securing collaborative medical ai by using differential privacy: Domain transfer for classification of chest radiographs,” Radiology: Artificial Intelligence, vol. 6, no. 1, p. e230212, 2023