ĐÁNH GIÁ KHÔNG GIAN EMBEDDING CỦA CÁC MÔ HÌNH NỀN TẢNG CHO ẢNH DA LIỄU NHIỄM TRÙNG NHẰM ĐỊNH HƯỚNG LỰA CHỌN BACKBONE CHO PIPELINE FINE-TUNING

Phạm Thị Anh Thư*, Nguyễn Gia Khôi, Trần Trung Hiếu, Trần Anh Đức, Bùi Thị Thanh Tú

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM

22dh113663@st.huflit.edu.vn, 22dh111723@st.huflit.edu.vn, 22dh114528@st.huflit.edu.vn, 23dh110826@st.huflit.edu.vn, tubtt@huflit.edu.vn

TÓM TẮT— Khi xây dựng pipeline fine-tuning ảnh da liễu trong điều kiện dữ liệu có nhãn khan hiếm, việc chọn đúng backbone embedding trước khi xây dựng pipeline fine-tuning sẽ ảnh hưởng đáng kể đến hiệu năng và chi phí của quá trình fine-tuning. Nghiên cứu so sánh 14 mô hình thuộc ba nhóm là phổ quát, y khoa tổng quát và chuyên biệt da liễu, trên cùng giao thức truy xuất ảnh theo ảnh với 1.300 ảnh lâm sàng bề mặt của 13 bệnh nhiễm trùng (leave-one-out), kết hợp ba kiểm định thống kê (bootstrap 95% CI, McNemar, paired permutation) và phân tích contamination bằng perceptual hashing. Kết quả cho thấy miền dữ liệu huấn luyện là yếu tố then chốt: các mô hình y khoa lệch miền chỉ đạt 76,92 đến 89,38% Top-1, trong khi các mô hình có ảnh da hoặc mô bệnh học đạt 90 đến 94%. Đáng chú ý, DermLIP với khoảng 86 triệu tham số đạt 94,15%, gần như ngang MetaCLIP H14 (94,08%) dù nhỏ hơn ∼10 lần (hiệu số +0,08 điểm phần trăm; CI [−0,77%, +0,92%]; McNemar và permutation p = 1,0). Ngược lại, RCLIP (X-quang/CT) chỉ đạt 76,92%, thấp hơn cả MobileCLIP S2 phổ quát (90,69%), cho thấy nhãn "y khoa" tự nó không mang lại lợi thế. Mức contamination với Derm1M chỉ ∼4,85% và không xáo trộn thứ hạng. Dựa trên hiệu năng, chi phí và cấu trúc gom cụm theo lớp bệnh, nghiên cứu đề xuất DermLIP làm điểm khởi đầu mặc định (94,15% Top-1, 8/13 lớp đạt 100% Top-1, nhanh hơn MetaCLIP ∼9 lần); các mô hình lệch miền như RCLIP và ClipMD nên được loại sớm; MetaCLIP H14 phù hợp khi cần khái quát đa miền, còn MedSigLIP khi cần bao phủ thêm các miền y tế khác. Kết quả cũng xác nhận cosine phù hợp hơn Euclid cho truy xuất đa phương thức.

Từ khóa— Embedding ảnh da liễu, đánh giá không gian embedding, mô hình nền tảng cho bệnh da, độ tương đồng Cosine trong truy xuất ảnh, lựa chọn backbone cho pipeline fine-tuning

I. GIỚI THIỆU

Trong ngành da liễu, hình ảnh đóng vai trò là nguồn thông tin quan trọng giúp bác sĩ thăm khám và điều trị các bệnh ngoài da. Phần lớn những nhận định ban đầu được đưa ra dựa trên việc quan sát trực tiếp đặc điểm của tổn thương như vị trí, kích thước, màu sắc và các thay đổi trên bề mặt da. Không chỉ hỗ trợ chẩn đoán, ảnh da liễu còn giúp theo dõi diễn tiến của bệnh từ giai đoạn khởi phát cho đến khi điều trị ổn định [1]. Vì vậy, việc khai thác hiệu quả thông tin từ ảnh da liễu có ý nghĩa thiết thực trong cả lâm sàng và nghiên cứu.

Những năm gần đây, sự phát triển của các phương pháp xử lý ảnh và học máy đã tạo điều kiện cho việc xây dựng nhiều công cụ tự động hỗ trợ phân tích ảnh y tế. Một hướng tiếp cận phổ biến là chuyển ảnh thành các véc-tơ đặc trưng để so sánh và tìm kiếm những hình ảnh có đặc điểm gần giống nhau. Trong phân tích ảnh da, các Mô hình thị giác embedding (Vision Embedding Model) – dựa trên kiến trúc học sâu như CNN hoặc Transformer – được sử dụng để trích xuất đặc trưng, phục vụ cho bài toán phân loại và truy vấn ảnh tương tự, với hiệu quả đã được chứng minh qua nhiều ứng dụng thực tế [1]. Tuy nhiên, khi áp dụng trực tiếp vào ảnh da liễu, các mô hình dùng chung này vẫn bộc lộ một số hạn chế như suy giảm hiệu năng khi gặp dữ liệu nằm ngoài phân phối huấn luyện, chưa bao quát đầy đủ sự đa dạng về tông da hoặc các bệnh hiếm gặp, đồng thời phụ thuộc nhiều vào các bộ dữ liệu đã được chuẩn hóa sẵn [2].

Khác với phần lớn các bài toán thị giác thông thường, việc phân biệt các bệnh da liễu hiếm khi dựa trên những đặc điểm dễ nhận biết. Nhiều bệnh có biểu hiện khá giống nhau ở giai đoạn sớm hoặc khi tổn thương xuất hiện tại những vị trí không điển hình, làm cho độ tương đồng giữa các lớp (inter-class similarity) cao trong khi mức biến thiên bên trong cùng một lớp lại rất lớn. Đây là một thách thức đã được ghi nhận trong nhiều nghiên cứu khi áp dụng các mạng học sâu truyền thống để phân tách các nhóm bệnh [1]. Những dấu hiệu mang giá trị chẩn đoán thường thể hiện qua các chi tiết nhỏ trong ảnh như ranh giới tổn thương rõ hay mờ, bề mặt da khô bong vảy hay ẩm rỉ dịch, hoặc sắc tố phân bố đồng đều hay loang lổ. Đây đều là những đặc trưng vi mô xuất hiện ở mức chi tiết nhỏ của tổn thương da. Đối với các mô hình vốn được huấn luyện chủ yếu trên ảnh đời sống, những đặc điểm này không thuộc nhóm đặc trưng mà mô hình được thiết kế để chú ý, từ đó làm giảm khả năng phân biệt các tổn thương phức tạp [3].

Trước những hạn chế này, một số nhóm nghiên cứu gần đây đã chuyển sang phát triển các mô hình được huấn luyện trực tiếp trên dữ liệu y tế, bao gồm cả ảnh da liễu. Việc đưa các mô hình chuyên biệt này vào so sánh nhằm đánh giá khả năng học được những đặc trưng mang ý nghĩa lâm sàng rõ ràng hơn, thay vì chỉ phản ánh mức độ

* Corresponding Author

tương đồng về mặt thị giác. Một số kết quả ban đầu cho thấy các mô hình này không chỉ giúp gom cụm ảnh theo đúng bản chất bệnh lý tốt hơn mà còn giảm nhầm lẫn giữa những bệnh có biểu hiện gần giống nhau [4]. Tuy nhiên, phần lớn các nghiên cứu hiện nay vẫn tập trung chủ yếu vào độ chính xác ở bước phân loại đầu ra. Trong khi đó, chất lượng của chính embedding ảnh vẫn chưa được khảo sát một cách hệ thống, dù đây là nền tảng quan trọng cho nhiều bước tiếp theo như fine-tuning mô hình, truy xuất ảnh y khoa hay xây dựng các hệ thống hỗ trợ ra quyết định lâm sàng.

Nghiên cứu này được thực hiện nhằm bổ sung khoảng trống còn tồn tại trong các nghiên cứu trước bằng cách đặt ba nhóm mô hình gồm mô hình phổ quát, mô hình y khoa tổng quát và mô hình chuyên biệt da liễu vào cùng một giao thức truy xuất ảnh theo ảnh sử dụng embedding không tinh chỉnh (frozen embeddings) cho nhóm bệnh nhiễm trùng. Cách thiết kế này phản ánh khá sát với những kịch bản triển khai thực tế, nơi các mô hình nền tảng thường được dùng như bộ trích xuất đặc trưng để đánh giá nhanh, xây dựng nguyên mẫu (prototype) cho hệ thống truy xuất ảnh hoặc làm bước sàng lọc ban đầu trước khi quyết định đầu tư huấn luyện chuyên sâu [5]. Nghiên cứu nhằm trả lời ba câu hỏi nghiên cứu sau:

RQ1: Khi dùng trực tiếp các mô hình nền tảng mà không tinh chỉnh thêm, mức độ gần nhau giữa dữ liệu huấn luyện ban đầu và ảnh da liễu lâm sàng bề mặt ảnh hưởng ra sao đến chất lượng embedding của mô hình, thể hiện qua hiệu năng truy xuất k-NN trên nhóm bệnh nhiễm trùng?

RQ2: Trong điều kiện tài nguyên còn hạn chế, việc huấn luyện trực tiếp trên ảnh da liễu có giúp mô hình học được tri thức miền đủ tốt để bù lại lợi thế về quy mô kiến trúc và dữ liệu của các mô hình tổng quát hay không, và nếu có thì bù được đến mức nào?

RQ3: Việc sử dụng dữ liệu tiền huấn luyện mang nhãn "y khoa" có tạo ra lợi thế cho bài toán ảnh da liễu lâm sàng bề mặt hay không, hay hiệu năng của mô hình vẫn phụ thuộc vào mức độ bao phủ cụ thể của ảnh da trong tập huấn luyện?

Đóng góp của bài báo bao gồm:

(1) Nghiên cứu xây dựng một khung đánh giá dùng để so sánh ba nhóm mô hình embedding gồm mô hình phổ quát, mô hình y khoa tổng quát và mô hình chuyên biệt da liễu trên cùng một giao thức truy xuất ảnh theo ảnh với 14 mô hình đại diện. Khung đánh giá này đi kèm ba phương pháp kiểm định thống kê gồm bootstrap 95% CI với 10.000 lần lấy mẫu, McNemar exact test và paired permutation test với 10.000 lần hoán vị, đồng thời kết hợp phân tích nhầm lẫn theo từng lớp bệnh. Cách thiết kế này có thể được tái sử dụng cho các bài toán da liễu khác hoặc các miền y tế có đặc thù tương tự.

(2) Nghiên cứu cung cấp bằng chứng thực nghiệm cho thấy miền dữ liệu huấn luyện có ảnh hưởng quyết định đến chất lượng embedding, thậm chí quan trọng hơn cả quy mô kiến trúc mô hình trong một số trường hợp. Cụ thể, DermLIP với kiến trúc ViT-B/16 và khoảng 86 triệu tham số đạt Top-1 = 94,15%, không có khác biệt thống kê đáng kể so với MetaCLIP H14 có khoảng 1 tỷ tham số với Top-1 = 94,08% ở cùng chỉ số đánh giá. Ngược lại, các mô hình y khoa được huấn luyện trên dữ liệu lệch miền như X-quang hoặc giải phẫu bệnh lại cho hiệu năng thấp hơn cả các mô hình phổ quát quy mô nhỏ.

(3) Nghiên cứu cũng thực hiện phân tích contamination định lượng nhằm đánh giá mức độ trùng lặp giữa tập đánh giá gồm 1.300 ảnh và tập huấn luyện Derm1M bằng perceptual hashing (pHash). Mức độ trùng lặp được chia thành ba nhóm gồm YES, SAME_LESION và SAME_DISEASE, đồng thời tiến hành thêm phân tích độ nhạy sau khi loại bỏ các trường hợp trùng lặp. Kết quả cho thấy mức contamination thực tế khá thấp, chỉ khoảng 4,85%, và không làm thay đổi thứ hạng giữa các mô hình, qua đó củng cố độ tin cậy cho các kết luận chính của nghiên cứu.

(4) Nghiên cứu cung cấp cơ sở định lượng cho việc lựa chọn backbone trong pipeline fine-tuning ảnh da liễu nhiễm trùng thông qua việc so sánh đồng thời ba khía cạnh gồm chất lượng embedding (Top-1, Top-5, Top-20), chi phí tính toán như thời gian trích xuất đặc trưng và cấu trúc gom cụm ở cấp độ lớp bệnh, thể hiện qua số lượng lớp đạt 100% Top-1.

II. TỔNG QUAN TÀI LIỆU

Tổng quan tài liệu được tổ chức theo ba hướng nghiên cứu có liên quan trực tiếp đến câu hỏi đặt ra trong bài báo: embedding tiền huấn luyện cho ảnh y tế; mô hình thị giác–ngôn ngữ trong y sinh; và truy xuất ảnh dựa trên nội dung trong da liễu. Cuối mục, một bảng tổng hợp đối chiếu các công trình tiêu biểu với nghiên cứu hiện tại làm rõ khoảng trống mà bài báo nhắm tới.

A. MÔ HÌNH EMBEDDING TIỀN HUẤN LUYỆN CHO ẢNH Y TẾ

Hướng sử dụng các mô hình nền tảng như bộ trích xuất đặc trưng cố định cho ảnh y tế đã được nghiên cứu khá rộng rãi trong khoảng mười năm trở lại đây. Raghu et al. [6] là một trong những công trình đầu tiên cho thấy hiệu

quả của học chuyển giao từ ImageNet sang ảnh y tế không chỉ phụ thuộc vào quy mô mô hình mà còn chịu ảnh hưởng lớn từ kiến trúc mô hình và đặc thù của miền dữ liệu đích. Nói cách khác, mô hình lớn hơn chưa chắc đã cho kết quả tốt hơn trong mọi trường hợp. Tiếp nối hướng nghiên cứu này, Azizi et al. [7] khai thác dữ liệu y tế không nhãn thông qua học tự giám sát và cho thấy các mô hình quy mô lớn có thể vượt trội hơn đáng kể so với cách học chuyển giao truyền thống từ ImageNet trong các bài toán phân loại y tế. Từ hai công trình này có thể thấy rằng chất lượng embedding không chỉ phụ thuộc vào kiến trúc mô hình mà còn bị chi phối mạnh bởi dữ liệu tiền huấn luyện và phương pháp học được sử dụng.

Theo hướng tiếp cận đó, Khun Jush et al. [5] đánh giá trực tiếp embedding tiền huấn luyện trên bài toán truy xuất ảnh y tế và cho thấy không gian embedding từ các mô hình lớn vẫn duy trì hiệu quả truy xuất tốt ngay cả khi không huấn luyện lại. Kết quả này cho thấy tiềm năng triển khai các mô hình nền tảng như bộ trích xuất đặc trưng trong bối cảnh dữ liệu y tế có nhãn còn hạn chế, đồng thời cũng là cơ sở thực nghiệm cho thiết kế đánh giá embedding không tinh chỉnh được áp dụng trong nghiên cứu này.

B. MÔ HÌNH THỊ GIÁC–NGÔN NGỮ TRONG Y SINH

Sự xuất hiện của các mô hình thị giác–ngôn ngữ như CLIP (Contrastive Language–Image Pretraining) [8] và GLoRIA [9] đã mở ra một hướng tiếp cận mới cho embedding ảnh y tế, trong đó mô hình được huấn luyện đồng thời trên cả ảnh và mô tả văn bản nhằm đưa thêm thông tin ngữ nghĩa lâm sàng vào không gian embedding. Theo hướng này, Zhang et al. [10] giới thiệu BiomedCLIP, một mô hình nền tảng được huấn luyện trên khoảng 15 triệu cặp ảnh–văn bản từ PubMed Central và đạt kết quả tốt trên nhiều bài toán y sinh, bao gồm zero-shot và truy xuất ảnh y tế. Khảo sát của Zhao et al. [11] về các biến thể CLIP cho y sinh như PubMedCLIP, BiomedCLIP và một số mô hình khác cũng cho thấy các mô hình được huấn luyện trên dữ liệu y sinh thường đạt hiệu năng cao hơn so với các mô hình phổ quát trong nhiều tác vụ y tế.

Tuy nhiên, các benchmark embedding y sinh hiện nay chủ yếu tập trung vào ảnh X-quang, CT hoặc giải phẫu bệnh, đồng thời phần lớn chỉ đánh giá thông qua zero-shot hoặc linear probing; trong khi đó, bài toán truy xuất ảnh theo ảnh lại ít được sử dụng để đánh giá chất lượng embedding. Những nghiên cứu so sánh hiện có cũng mới dừng ở hai nhóm mô hình là phổ quát và y sinh tổng quát, còn các mô hình chuyên biệt cho da liễu như DermLIP [4], WhyLesionCLIP [12] hay các mô hình nền tảng thế hệ mới như PanDerm [13] và SkinGPT-4 [14] vẫn chưa được đặt trong cùng một khung đánh giá. Cần lưu ý rằng PanDerm và SkinGPT-4 được thiết kế cho những tác vụ phức tạp hơn như chẩn đoán đa phương thức hoặc sinh báo cáo lâm sàng nên không thuộc phạm vi so sánh trực tiếp của nghiên cứu này. Trong nghiên cứu hiện tại, chỉ các mô hình embedding theo kiểu CLIP với cùng giao diện trích xuất đặc trưng mới được đưa vào đánh giá. Nhìn chung, ảnh da liễu lâm sàng với những đặc thù riêng về màu sắc, cấu trúc bề mặt và bối cảnh chụp vẫn chưa được khảo sát đầy đủ trong khung của các mô hình embedding hiện đại.

C. TRUY XUẤT ẢNH DỰA TRÊN NỘI DUNG TRONG DA LIỄU

Các hệ thống truy xuất ảnh dựa trên nội dung (Content-Based Image Retrieval – CBIR) đã được nghiên cứu trong lĩnh vực da liễu từ hơn một thập kỷ trước, chủ yếu phục vụ cho hỗ trợ chẩn đoán và giáo dục y khoa. Tschandl et al. [15] cho thấy truy xuất ảnh dựa trên nội dung sử dụng đặc trưng từ Convolutional Neural Network (CNN) có thể đạt độ chính xác chẩn đoán tương đương với dự đoán softmax trên ảnh nội soi da, đồng thời mang lại khả năng diễn giải tốt hơn thông qua việc hiển thị các ca bệnh tương tự cho bác sĩ tham khảo. Gassner et al. [16] tiếp tục phát triển hệ thống truy xuất ảnh dựa trên nội dung có tăng cường bản đồ nổi bật mang tên Saliency-Enhanced Content-Based Image Retrieval (SE-CBIR), trong đó bản đồ nổi bật được sử dụng để tăng khả năng diễn giải kết quả. Nghiên cứu này ghi nhận mức cải thiện khoảng 22% độ chính xác phân loại khi bác sĩ sử dụng kết quả truy xuất như một công cụ hỗ trợ. Trong khi đó, Barhoumi và Khelifa [17] đề xuất một hệ thống truy xuất ảnh da liễu có khả năng lựa chọn động độ đo khoảng cách cho từng truy vấn, từ đó cho thấy không tồn tại một độ đo duy nhất tối ưu cho mọi loại tổn thương da.

Những nghiên cứu gần đây chủ yếu tập trung vào hai hướng chính là cải thiện phương pháp trích xuất đặc trưng và mở rộng truy vấn. Huang et al. [18] đề xuất phương pháp học tương phản (contrastive learning) kết hợp nhận biết tổn thương cho bài toán truy xuất và phân loại ảnh da, mở ra hướng kết hợp giữa biểu diễn đặc trưng và metric learning chuyên biệt. Rashad et al. [19] phát triển phương pháp Retrieval based on Query Expansion (RbQE) nhằm mở rộng truy vấn trong truy xuất ảnh y tế dựa trên nội dung, qua đó cải thiện hiệu năng trong các trường hợp ảnh truy vấn ban đầu chưa chứa đủ đặc trưng cần thiết. Bên cạnh đó, Liu et al. [20] xây dựng hệ thống chẩn đoán phân biệt sử dụng học sâu cho 26 bệnh da phổ biến và cho thấy tiềm năng lớn của các biểu diễn đặc trưng sâu trong việc phân biệt những tổn thương có hình thái gần giống nhau.

Tuy nhiên, phần lớn các hệ thống CBIR trong da liễu hiện nay vẫn tập trung vào ảnh nội soi da từ các bộ dữ liệu tiêu chuẩn như ISIC, chủ yếu xoay quanh bài toán phân loại các nhóm u da như melanoma, nevus hoặc carcinoma với đặc trưng được trích xuất từ các CNN truyền thống. Trong khi đó, việc đánh giá embedding từ các mô hình nền

tảng thị giác–ngôn ngữ trên ảnh da liễu lâm sàng bề mặt, đặc biệt đối với nhóm bệnh nhiễm trùng có mức độ chồng lấn hình thái cao, vẫn còn là khoảng trống chưa được nghiên cứu đầy đủ.

D. KHOẢNG TRỐNG NGHIÊN CỨU

Bảng 1 đối chiếu các nghiên cứu tiêu biểu trong ba hướng nêu trên với nghiên cứu hiện tại theo bốn tiêu chí: nhóm mô hình được so sánh, miền dữ liệu đánh giá, tác vụ đánh giá và mức độ kiểm định thống kê. Đối chiếu này cho thấy ba khoảng trống cụ thể: (i) chưa có nghiên cứu nào đặt đồng thời ba nhóm mô hình (phổ quát, y khoa tổng quát, chuyên biệt da liễu) trên cùng một giao thức đánh giá; (ii) ảnh da liễu lâm sàng bề mặt thuộc nhóm bệnh nhiễm trùng chưa được dùng làm miền kiểm tra cho các mô hình embedding hiện đại; (iii) các kết luận về hiệu năng tương đối giữa các mô hình thường chỉ dựa vào chênh lệch điểm số mà thiếu kiểm định thống kê chặt chẽ. Nghiên cứu này nhằm lấp đồng thời cả ba khoảng trống.

Bảng 1. *Đối chiếu các nghiên cứu tiêu biểu với nghiên cứu hiện tại*

Nghiên cứu	Nhóm mô hình so sánh	Miền dữ liệu đánh giá	Tác vụ đánh giá	Kiểm định thống kê
Raghu et al. [6]	Phổ quát (ImageNet) vs. y khoa	X-quang, võng mạc	Phân loại	Không
Azizi et al. [7]	ImageNet vs. tự giám sát y khoa	X-quang, da liễu	Phân loại	Có (CI)
Khun Jush et al. [5]	Embedding tiền huấn luyện đa miền	Ảnh y tế đa phương thức	Truy xuất	Không
Zhao et al. [11]	CLIP, PubMedCLIP, BiomedCLIP	Ảnh y sinh đa miền	Phân loại zero-shot, linear probing	Không
Tschandl et al. [15]	CNN truyền thống	Ảnh nội soi da (ISIC)	CBIR + chẩn đoán	Có (ROC)
Gassner et al. [16]	CNN + saliency map	Ảnh nội soi da	CBIR + reader study	Có
Huang et al. [18]	Contrastive + lesion-aware	Ảnh nội soi da	CBIR + phân loại	Không
Nghiên cứu này	**3 nhóm: phổ quát, y khoa tổng quát, chuyên biệt da liễu (14 mô hình)**	**Ảnh lâm sàng bề mặt, 13 bệnh nhiễm trùng**	**Truy xuất ảnh theo ảnh + phân tích nhầm lẫn theo lớp**	**Có (bootstrap CI, McNemar, permutation)**

Trên cơ sở đối chiếu này, nghiên cứu sử dụng giao thức truy xuất k-NN với embedding không tinh chỉnh trên 1.300 ảnh thuộc 13 bệnh nhiễm trùng. Đây là bước sàng lọc trước khi tinh chỉnh trên dữ liệu đích: mô hình có không gian embedding tốt ngay từ đầu là ứng viên đáng đầu tư thêm, mô hình kém có thể được loại sớm để tiết kiệm chi phí phát triển. Kết quả thu được không thay thế đánh giá sau fine-tuning mà cung cấp một cơ sở định lượng cho việc chọn mô hình nền tảng làm backbone trong pipeline fine-tuning da liễu nhiễm trùng.

III. PHƯƠNG PHÁP NGHIÊN CỨU

A. MỤC TIÊU NGHIÊN CỨU

Mục tiêu nghiên cứu là đánh giá thực chất mức độ phù hợp của các mô hình embedding hiện có khi áp dụng trực tiếp cho ảnh da liễu thuộc nhóm bệnh nhiễm trùng, trong bối cảnh gần như không có dữ liệu để huấn luyện thêm. Hướng phân loại truyền thống đòi hỏi nhãn đầy đủ và chi phí huấn luyện không nhỏ; nghiên cứu đặt ra một câu hỏi sát thực tế triển khai hơn: liệu mô hình có tự sắp xếp được các ảnh cùng bệnh lý vào những vùng lân cận trong không gian embedding hay không. Đây là yêu cầu tiên quyết: nếu không gian embedding không phản ánh quan hệ bệnh học, mọi ứng dụng kế tiếp – từ tìm kiếm ảnh tương tự đến hỗ trợ chẩn đoán – đều bị giới hạn. Yêu cầu này đặc biệt liên quan đến hệ thống truy xuất ảnh dựa trên nội dung trong y học, vốn không dùng nhãn bệnh có sẵn mà chuyển mỗi ảnh thành một biểu diễn số và so sánh khoảng cách giữa các biểu diễn để truy về những ca tương đồng cho bác sĩ tham khảo.

Nhiều nghiên cứu trong y tế đã cho thấy cách tiếp cận này vẫn hiệu quả ngay cả khi không huấn luyện lại, miễn là embedding đủ tốt để phản ánh các đặc điểm chính của ảnh y khoa [5]. Đây là quan sát có ý nghĩa thực tiễn lớn, bởi

dữ liệu y tế có nhãn vừa khan hiếm vừa tốn kém để thu thập: xây dựng một tập có nhãn đáng tin cậy đòi hỏi chuyên môn lâm sàng và nguồn lực gán nhãn đáng kể. Trong điều kiện đó, dùng mô hình nền tảng làm bộ trích xuất đặc trưng sẵn có khả thi hơn nhiều so với tự huấn luyện từ đầu [6]. Theo hướng đó, nghiên cứu sử dụng các mô hình embedding ở trạng thái gốc, đóng vai trò bộ trích xuất đặc trưng cố định, không huấn luyện lại hay tinh chỉnh tham số [21]. Đây cũng chính là kịch bản triển khai phổ biến trên thực tế: mô hình nền tảng thường được dùng để đánh giá nhanh dữ liệu, dựng nguyên mẫu hệ thống truy xuất hoặc làm bước tiền xử lý trước khi quyết định có đầu tư huấn luyện chuyên sâu hay không [19]. Các công trình trong y học cũng cho thấy embedding tiền huấn luyện có thể được dùng trực tiếp mà vẫn cho kết quả có ý nghĩa, nhất là ở các bài toán truy xuất và phân tích độ tương đồng [5].

Trên cơ sở đó, nghiên cứu so sánh các mô hình thuộc ba nhóm: mô hình phổ quát, mô hình y khoa tổng quát và mô hình chuyên biệt cho da liễu, nhằm làm rõ ảnh hưởng của miền dữ liệu huấn luyện – vốn được xem là yếu tố then chốt chi phối khả năng chuyển giao biểu diễn sang các tác vụ mới. Kết quả đánh giá đóng vai trò bước sàng lọc ban đầu, làm cơ sở cho việc chọn mô hình embedding phù hợp trước khi bước vào các giai đoạn phát triển sâu hơn như xây dựng hệ thống truy xuất ảnh, phân tích độ tương đồng giữa các biểu hiện bệnh lý hay tinh chỉnh cho các bài toán chẩn đoán da liễu cụ thể.

B. TẬP DỮ LIỆU

Tập dữ liệu trong nghiên cứu gồm 1.300 ảnh da liễu lâm sàng bề mặt, phân bổ đều 100 ảnh cho mỗi trong 13 lớp bệnh nhiễm trùng. Quy trình xây dựng tập dữ liệu gồm ba bước: (i) thu thập ảnh từ các kho công khai có cấp phép cho mục đích nghiên cứu; (ii) sàng lọc theo tiêu chí chất lượng và rà soát thủ công; (iii) lấy mẫu ngẫu nhiên đến đủ 100 ảnh mỗi lớp để bảo đảm cân bằng giữa các lớp. Mỗi ảnh trong tập đánh giá được truy ngược về nguồn gốc cụ thể trên các nền tảng công khai (Kaggle, Roboflow, Mendeley) để phục vụ tính tái lập. Bảng 2 báo cáo chi tiết phân bổ ảnh theo từng nguồn [22]–[30] kèm nhãn cấp phép tương ứng.

Bảng 2. *Phân bổ số lượng ảnh từ mỗi nguồn cho từng lớp bệnh*

Lớp bệnh	Skin-Disease-Dataset (Kaggle) [22]	Dermatology_hair [23]	human monkey pox [24]	Skin-Disease-Dataset (Roboflow) [25]	Skin-Disease-detection [26]	MSID [27]	skin disease [28]	Skin [29]	Candida intertrigo [30]	Tổng
Chickenpox	100									100
Folliculitis		100								100
Head Lice		100								100
Herpes					80		20			100
Herpes Zoster	100									100
Infectious	28			12			24	24	12	100
Infestation/ Bites	100									100
Monkeypox			87			13				100
Nail Fungus	100									100
Ringworm	100									100
Sarampion			66			34				100
Tinea Capitis		100								100
Warts				100						100
Tổng	528	300	153	112	80	47	44	24	12	1.300
License	**CC0**	**Apache 2.0**	**Public Domain**	**CC BY 4.0**	**CC BY 4.0**	**CC BY 4.0**	**CC BY 4.0**	**CC BY 4.0**	**CC BY 4.0**	

Hình 1 và Hình 2 minh họa ảnh đại diện cho từng lớp bệnh trong tập dữ liệu, làm rõ thách thức về chồng lấn hình thái lâm sàng, đặc biệt giữa nhóm bệnh có biểu hiện mụn nước/sẩn (Chickenpox, Herpes Zoster, Herpes,

Monkeypox) và biểu hiện ban đỏ lan tỏa (Sarampion, Infectious), yếu tố quyết định mức độ khó của bài toán phân biệt và độ nhạy của không gian embedding.

Hình 1. *Ảnh đại diện minh họa các bệnh nhiễm trùng có biểu hiện mụn nước/sẩn hoặc ban đỏ lan tỏa, dễ chồng lấn hình thái: (a) Chickenpox, (b) Herpes Zoster, (c) Herpes, (d) Monkeypox, (e) Sarampion, (f) Infectious*

Hình 2. *Ảnh đại diện các nhóm bệnh ngoài da còn lại trong tập dữ liệu: (a) Folliculitis, (b) Infestation/Bites, (c) Head Lice, (d) Tinea Capitis, (e) Ringworm, (f) Nail Fungus, (g) Warts*

Tập ảnh được giới hạn ở một phương thức duy nhất là ảnh lâm sàng bề mặt da chụp bằng thiết bị quang học thông thường (điện thoại, máy ảnh kỹ thuật số); ảnh nội soi da và ảnh giải phẫu bệnh đều bị loại để kiểm soát biến số – khi trộn nhiều phương thức, mô hình có xu hướng gom cụm theo loại thiết bị thay vì theo bệnh. Về nhãn, nhãn được kế thừa trực tiếp từ tập gốc, chưa qua xác nhận lại bởi bác sĩ da liễu độc lập; tác động của giới hạn này được phân tích định lượng trong phần Kết quả (Phân tích nhầm lẫn) và Giới hạn nghiên cứu. Ba biện pháp giảm thiểu được áp dụng: (i) ưu tiên các lớp lấy hoàn toàn từ một nguồn (9/13 lớp); (ii) báo cáo kết quả theo từng lớp thay vì chỉ trung bình tổng; (iii) phân tích loại trừ lớp có hiệu năng thấp bất thường.

Trong số 13 lớp, 9 lớp được lấy hoàn toàn từ một nguồn duy nhất (Chickenpox, Folliculitis, Head Lice, Herpes Zoster, Infestation/Bites, Nail Fungus, Ringworm, Tinea Capitis, Warts), bảo đảm tính nhất quán nội bộ về phong cách ảnh và tiêu chuẩn gán nhãn. Bốn lớp còn lại (Herpes, Infectious, Monkeypox, Sarampion) được tổng hợp từ 2–5 nguồn để đủ 100 ảnh. Riêng lớp Infectious có nhãn mang tính tổng quát, không đại diện cho một bệnh lý cụ thể; chồng lấn hình thái với các lớp khác và tác động lên hiệu năng truy xuất ở lớp này được phân tích chi tiết

trong phần Kết quả. Lớp Infectious được giữ lại có chủ đích nhằm phục vụ phân tích định lượng ảnh hưởng của chất lượng nhãn đến hiệu năng truy xuất; cụ thể, khi loại trừ lớp này khỏi đánh giá, Top-1 của tất cả mô hình đều tăng đồng đều khoảng 3,7–4,8 điểm phần trăm mà không làm thay đổi thứ hạng giữa các mô hình (xem mục Phân tích nhầm lẫn). Việc giữ lại lớp tổng quát này còn cho phép phản ánh kịch bản triển khai thực tế, nơi các tập dữ liệu y khoa công khai thường chứa các nhãn chưa được chuẩn hóa.

C. CÁC MÔ HÌNH EMBEDDING ĐƯỢC ĐÁNH GIÁ

Các mô hình embedding trong nghiên cứu được chia thành ba nhóm dựa trên miền dữ liệu dùng để huấn luyện ban đầu: mô hình phổ quát, mô hình y khoa tổng quát và mô hình chuyên biệt cho da liễu. Cách phân chia này xuất phát từ giả định rằng mức độ phù hợp giữa miền huấn luyện và miền ứng dụng sẽ ảnh hưởng trực tiếp đến cách mô hình tổ chức không gian embedding.

Nhóm phổ quát gồm các mô hình được huấn luyện trên những tập dữ liệu Internet lớn và đa dạng, chủ yếu chứa ảnh đời sống và nội dung trực tuyến. Ưu điểm của nhóm này là khả năng học các đặc trưng thị giác mang tính tổng quát và hoạt động ổn định trên nhiều loại ảnh khác nhau. Tuy nhiên, vì không được thiết kế riêng cho y học, các mô hình này thường chú ý nhiều hơn đến hình dạng tổng thể, màu sắc nổi bật hoặc bố cục ảnh. Trong bài toán da liễu, nơi các dấu hiệu quan trọng thường nằm ở chi tiết nhỏ, cần kiểm tra xem các đặc trưng tổng quát đó có đủ để phân biệt và gom cụm các tổn thương bệnh lý hay không.

Nhóm y khoa tổng quát được phát triển nhằm thu hẹp khoảng cách này bằng cách huấn luyện trên các bộ dữ liệu y tế đa chuyên ngành. Giả thuyết đặt ra là mô hình sẽ nhạy hơn với các đặc điểm thường gặp trong ảnh y khoa như cấu trúc mô, thay đổi mật độ hoặc hình thái tổn thương. Tuy vậy, do dữ liệu huấn luyện trải rộng từ X-quang, CT, MRI đến nội soi, mức độ phù hợp của nhóm này với ảnh da liễu lâm sàng bề mặt vẫn cần được đánh giá thêm.

Cuối cùng là nhóm chuyên biệt cho da liễu, được huấn luyện trực tiếp trên ảnh tổn thương da để học các đặc điểm hình thái mang ý nghĩa chẩn đoán. Nhóm này được đưa vào nhằm kiểm tra xem tính chuyên biệt của dữ liệu huấn luyện có tạo ra lợi thế rõ rệt trong không gian embedding hay không, đặc biệt trong điều kiện không huấn luyện lại hay tinh chỉnh tham số. Nhóm này được kỳ vọng sẽ tổ chức không gian embedding dựa nhiều hơn vào quan hệ bệnh học thay vì chỉ dựa trên mức độ giống nhau về mặt thị giác.

Việc đặt cả ba nhóm vào cùng một khung so sánh giúp phân tích trực tiếp mối liên hệ giữa miền dữ liệu huấn luyện và chất lượng không gian embedding khi áp dụng cho ảnh da liễu nhiễm trùng. Thay vì chỉ nhìn vào kết quả cuối cùng, cách tiếp cận này giúp giải thích rõ hơn vì sao một mô hình hoạt động tốt hoặc kém trong từng bối cảnh cụ thể. Danh sách chi tiết các mô hình, phiên bản, số chiều biểu diễn và các thông số liên quan đến thời gian suy luận sẽ được trình bày ở phần kết quả để tiện đối chiếu và phân tích.

Bảng 3. *Tổng hợp 14 mô hình embedding được đánh giá*

Mô hình	Nhóm	Bộ mã hóa thị giác	Tham số (Tổng / VE)	Dim	Phiên bản / Nguồn (HuggingFace)	Dữ liệu huấn luyện
MetaCLIP (H14) [31]	Phổ quát	ViT-H/14	~1B / ~632M	1024	facebook/metaclip-h14-fullcc2.5b	FullCC2.5B — 2,5 tỷ cặp ảnh–văn bản từ Internet
OpenAI-CLIP (Large-P14) [8]	Phổ quát	ViT-L/14	~304M / ~304M	768	openai/clip-vit-large-patch14	WIT — 400 triệu cặp ảnh–văn bản từ Internet
OpenAI-CLIP (Base-P32) [8]	Phổ quát	ViT-B/32	~86M / ~88M	512	openai/clip-vit-base-patch32	WIT — 400 triệu cặp ảnh–văn bản từ Internet
SigLIP (SO400M) [32]	Phổ quát	SO400M	~400M / ~400M	1152	google/siglip-so400m-patch14-384	WebLI — dữ liệu Internet đa ngôn ngữ
AltCLIP (Base) [33]	Phổ quát	ViT-B/16	~86M / ~86M	768	BAAI/AltCLIP	Dữ liệu song ngữ Anh–Trung từ Internet
ConvNeXt-CLIP (Base) [34], [35]	Phổ quát	ConvNeXt-B	~89M / ~89M	640	laion_aesthetic_s13b_b82k	LAION — dữ liệu Internet quy mô lớn

MobileCLIP (S2) [36]	Phổ quát	ViT-S (mobile)	~35M / ~35M	512	apple/MobileCLIP-S2	DataCompDR — kiến trúc tối ưu tốc độ
MedSigLIP (448) [37]	Y khoa	SO400M	~400M / ~400M	1152	google/medsiglip-448	Dữ liệu y văn đa chuyên khoa (X-quang, CT, da liễu, nhãn khoa)
BiomedCLIP (PubMed) [10]	Y khoa	ViT-B/16	~86M / ~86M	512	microsoft/BiomedCLIP-PubMedBERT_256	PMC-15M — 15 triệu cặp ảnh–văn bản từ PubMed Central
PLIP (Base) [38]	Y khoa	ViT-B/16	~86M / ~86M	512	vinid/plip	Ảnh giải phẫu bệnh từ Twitter y tế
ClipMD (Base) [39]	Y khoa	ViT-B/16	~86M / ~86M	512	Idan0405/ClipMD	Dữ liệu y khoa đa dạng, quy mô nhỏ
RCLIP (Base) [40]	Y khoa	ViT-B/16	~86M / ~86M	512	kaveh/rclip	Ảnh X-quang và CT
DermLIP (ViT-B/16) [4]	Da liễu	ViT-B/16	~86M / ~86M	512	redlessone/DermLIP_ViT-B-16	Ảnh da liễu lâm sàng bề mặt
WhyLesionCLIP (ViT-L/14) [12]	Da liễu	ViT-L/14	~304M / ~304M	768	yyupenn/WhyLesionCLIP	Ảnh nội soi da ISIC + tri thức từ sách giáo khoa y khoa

Ghi chú: Cột "Tham số (Tổng/ VE)" trong Bảng 3 báo cáo hai con số ngăn cách bởi dấu "/": tổng số tham số của toàn bộ mô hình (bao gồm cả vision encoder và text encoder), và số tham số của riêng vision encoder (viết tắt VE). Cách trình bày này thống nhất với các báo cáo gốc trên HuggingFace Hub. Do nghiên cứu chỉ sử dụng vision encoder để trích xuất embedding ảnh, các phân tích về thời gian suy luận và chi phí tính toán ở phần sau sẽ trích cụ thể quy mô tham số của vision encoder khi cần thiết.

D. LỰA CHỌN ĐỘ ĐO TƯƠNG ĐỒNG

Lựa chọn độ đo tương đồng có ảnh hưởng trực tiếp đến chất lượng truy xuất. Khoảng cách Euclid được dùng phổ biến trong các hệ truy xuất truyền thống nhờ tính trực quan, nhưng chỉ phù hợp khi cả hướng và độ lớn của véc-tơ đều mang ý nghĩa ngữ nghĩa. Với các mô hình dùng học tương phản [41] hoặc đa phương thức như CLIP [8], mục tiêu huấn luyện căn chỉnh các véc-tơ theo hướng và chuẩn hóa độ lớn về cùng giá trị, nên Euclid dễ gây sai lệch: hai véc-tơ cùng hướng (cùng loại tổn thương) nhưng khác độ lớn (không có ý nghĩa y học) vẫn cho khoảng cách Euclid lớn. Cosine xử lý vấn đề này bằng cách chỉ đo góc giữa các véc-tơ, bỏ qua độ lớn. Về mặt toán học, với hai véc-tơ u và v trong không gian n chiều, khoảng cách Euclid được tính bằng công thức [42]:

$$d_{Euclid}(u, v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$

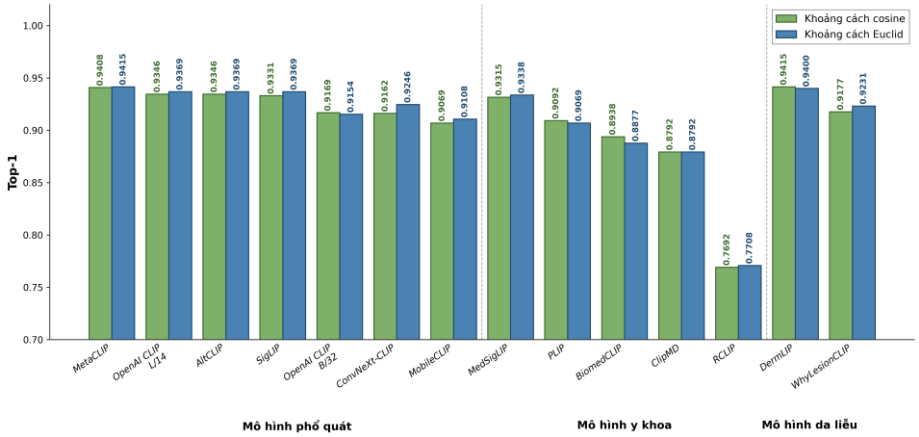
Trong đó, u_i và v_i lần lượt là giá trị đặc trưng tại chiều thứ i của véc-tơ u và v. Độ tương đồng cosine được định nghĩa là tích vô hướng của hai véc-tơ chia cho tích chuẩn L2 của chúng [42]:

$$Cosine(u, v) = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \cdot \sqrt{\sum_{i=1}^{n} v_i^2}}$$

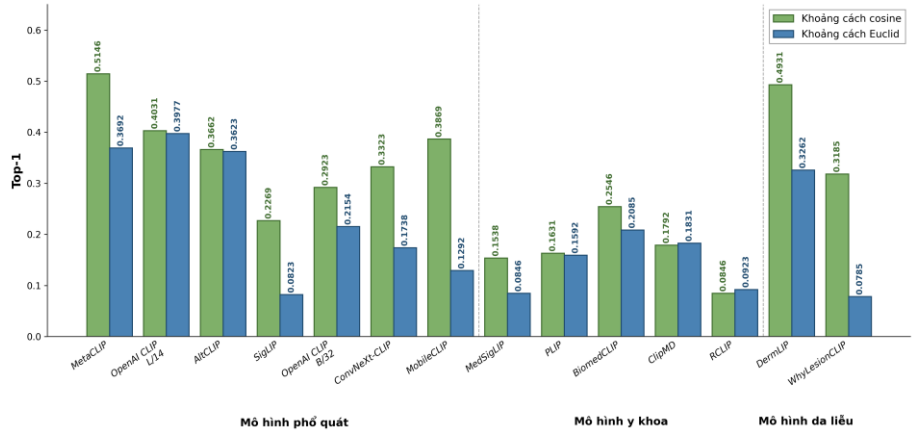
Trong đó, tử số đo mức độ cùng hướng của hai véc-tơ, mẫu số chuẩn hóa triệt tiêu ảnh hưởng của độ lớn; giá trị Cosine dao động trong [−1, 1] và càng gần 1 khi hai véc-tơ càng song song cùng chiều. Khi các véc-tơ được chuẩn hóa L2, khoảng cách Euclid và độ tương đồng cosine có quan hệ $d_{Euclid}(u, v) = \sqrt{2 - 2 \cdot cos(u, v)}$, cho thứ hạng truy xuất hoàn toàn đồng nhất [42]; Qian et al. [43] đã phân tích chi tiết quan hệ này trong bối cảnh truy vấn láng giềng gần nhất (nearest neighbor). Trong trường hợp không chuẩn hóa, đặc biệt với biểu diễn đa phương thức, cosine cho kết quả ổn định hơn nhờ triệt tiêu chênh lệch về chuẩn véc-tơ [42].

Để minh chứng, nghiên cứu thực hiện một thực nghiệm trên 1.300 ảnh với cấu hình truy xuất ảnh bằng văn bản. Top-1 ở phần lớn mô hình suy giảm rõ khi chuyển từ cosine sang Euclid: WhyLesionCLIP [12] giảm từ 0,32 xuống 0,08 (~75%), MobileCLIP từ 0,39 xuống 0,13 (~67%), SigLIP từ 0,23 xuống 0,08 (~64%); ngay cả các mô hình

mạnh như MetaCLIP [31] và DermLIP [4] cũng giảm từ 0,51 xuống 0,37 (~28%) và từ 0,49 xuống 0,33 (~34%). OpenAI-CLIP ViT-Base/32 [8] và ViT-Large/14 [8] suy giảm thấp hơn (lần lượt 0,29 xuống 0,22 và giữ ~0,40), cho thấy mức ảnh hưởng phụ thuộc vào phân bố embedding của từng mô hình. Sự suy giảm này phản ánh khoảng cách phương thức giữa vision encoder và text encoder trong các mô hình CLIP, vốn được huấn luyện bằng contrastive loss để căn chỉnh theo hướng véc-tơ. Steck et al. [44] chỉ ra hạn chế của cosine với embedding học từ mô hình tuyến tính có chính quy hóa, nhưng các mô hình CLIP trong nghiên cứu này đều được tối ưu trên mặt cầu đơn vị và chuẩn hóa L2 nên không gặp vấn đề đó. Dựa trên cả cơ sở lý thuyết và thực nghiệm, cosine được chọn làm độ đo chuẩn cho quá trình đánh giá.



Hình 3. *So sánh Top-1 truy xuất của 14 mô hình embedding qua khoảng cách cosine vs Euclid (truy xuất ảnh theo ảnh)*



Hình 4. *So sánh Top-1 truy xuất của 14 mô hình embedding qua khoảng cách cosine vs Euclid (truy xuất ảnh bằng văn bản)*

Hình 3 và Hình 4 trực quan hóa khác biệt giữa hai kịch bản. Hai kịch bản này được đối chiếu để kiểm chứng độ ổn định của cosine và Euclid khi điều kiện đầu vào thay đổi: truy xuất ảnh theo ảnh dùng cùng một bộ mã hóa cho cả truy vấn và tra cứu (cùng phương thức ảnh), trong khi truy xuất ảnh bằng văn bản dùng hai bộ mã hóa khác phương thức (vision và text) – vốn được biết là chịu ảnh hưởng của khoảng cách phương thức trong các mô hình CLIP. Nếu cosine và Euclid cho kết quả tương đương ở cả hai kịch bản, có thể chọn tùy ý; nếu khác biệt, kết quả cho biết độ đo nào phù hợp hơn cho truy xuất đa phương thức. Ở truy xuất ảnh theo ảnh (Hình 3), cosine và Euclid gần như trùng nhau trên cả 14 mô hình – phù hợp với đẳng thức $d_{Euclid}(u, v) = \sqrt{2 - 2 \cdot cos(u, v)}$, khi embedding đã chuẩn hóa L2. Ở truy xuất ảnh bằng văn bản (Hình 4), Top-1 thấp hơn hẳn (0,08–0,51 so với 0,77–0,94 ở Bảng 4–6) do tác vụ chéo phương thức khó hơn, và mức suy giảm khi chuyển sang Euclid tỷ lệ thuận với khoảng cách phương thức của mỗi mô hình – kết quả khẳng định cosine là lựa chọn phù hợp cho truy xuất đa phương thức.

E. QUY TRÌNH ĐÁNH GIÁ

Quy trình đánh giá truy xuất ảnh được thực hiện theo giao thức leave-one-out trên toàn bộ tập dữ liệu 1.300 ảnh: mỗi ảnh lần lượt đóng vai trò ảnh truy vấn, 1.299 ảnh còn lại làm thư viện tìm kiếm. Việc tách hoàn toàn ảnh truy vấn khỏi thư viện trong mỗi lượt đảm bảo kết quả phản ánh đúng khả năng truy xuất ảnh tương đồng, không bị nhiễu bởi đối sánh ảnh với chính nó. Trước khi tính toán, toàn bộ véc-tơ embedding được chuẩn hóa L2. Với mỗi truy vấn, hệ thống tính độ tương đồng cosine giữa ảnh truy vấn và từng ảnh trong thư viện, rồi xếp hạng theo thứ tự giảm dần. Quy trình lặp cho tất cả 1.300 ảnh, tạo thành 1.300 lượt truy vấn độc lập. Giao thức leave-one-out giúp tận dụng tối đa lượng dữ liệu có hạn (100 ảnh mỗi lớp) cho cả vai trò truy vấn lẫn tìm kiếm, đồng thời tránh

rò rỉ dữ liệu giữa hai tập. Song song với cosine, khoảng cách Euclid cũng được tính trên cùng giao thức để đối chiếu hai độ đo.

Chỉ số đánh giá hiệu năng truy xuất trong nghiên cứu này là Top-K, định nghĩa là tỷ lệ phần trăm các lượt truy vấn có ít nhất một ảnh cùng lớp bệnh với ảnh truy vấn xuất hiện trong K kết quả được xếp hạng cao nhất. Chỉ số này còn được gọi là Recall@K hoặc Hit@K trong tài liệu về truy xuất ảnh; bài báo dùng thống nhất tên gọi Top-K trong toàn bộ phần trình bày và các bảng kết quả. Cần phân biệt Top-K trong truy xuất ảnh với độ chính xác Top-K trong phân loại: chỉ số sau yêu cầu nhãn đúng nằm trong K lớp có xác suất dự đoán cao nhất, còn Top-K trong truy xuất yêu cầu ít nhất một ảnh cùng lớp xuất hiện trong K ảnh được truy xuất gần nhất. Nghiên cứu báo cáo Top-1, Top-5, Top-10 và Top-20 để mô tả cấu trúc không gian embedding ở các mức độ chi tiết khác nhau. Top-1 phản ánh khả năng phân tách các đặc trưng vi mô ở vùng lân cận gần nhất, còn Top-5 đến Top-20 phản ánh cấu trúc gom cụm ngữ nghĩa cục bộ. Trường hợp mô hình nhầm ở Top-1 do tương đồng hình thái cao giữa các bệnh nhiễm trùng nhưng vẫn đưa kết quả đúng vào Top-5 hay Top-10 cho thấy không gian embedding vẫn xếp các tổn thương cùng bản chất ở vùng lân cận. Cách báo cáo này phù hợp với kịch bản hỗ trợ chẩn đoán dựa trên truy xuất, nơi bác sĩ thường tham khảo một nhóm các ca bệnh tương tự để đối chiếu trước khi đưa ra chẩn đoán cuối cùng.

IV. KẾT QUẢ THỰC NGHIỆM VÀ PHÂN TÍCH

Toàn bộ thí nghiệm được thực hiện trên Google Colaboratory với GPU NVIDIA Tesla T4 (15 GB VRAM), CPU Intel Xeon @ 2.00GHz, RAM 12,7 GB và hệ điều hành Ubuntu 22.04. Các framework và thư viện sử dụng gồm PyTorch 2.10.0+cu128, transformers 5.0.0 và OpenCLIP 3.3.0. Trước khi đo thời gian chính thức, mỗi mô hình được khởi động trước 10 lượt suy luận. Tất cả 14 mô hình đều được đo trong cùng một phiên chạy với cấu hình đồng nhất gồm cùng GPU, precision fp32 và batch size nhằm đảm bảo tính nhất quán giữa các phép đo. Cột "Thời gian (s/ảnh)" trong Bảng 4–6 chỉ ghi nhận thời gian trích xuất embedding từ ảnh, không bao gồm thời gian similarity search trong gallery. Thời gian này được đo trên toàn bộ 1.300 ảnh và báo cáo dưới dạng giá trị trung bình. Toàn bộ checkpoint được tải trực tiếp từ HuggingFace Hub thông qua thư viện transformers; phiên bản và identifier của 14 mô hình đã được liệt kê ở Bảng 3 để đảm bảo khả năng tái lập nghiên cứu.

Về tiền xử lý ảnh, mỗi mô hình sử dụng processor đi kèm do tác giả công bố trên HuggingFace Hub, bao gồm AutoProcessor của transformers hoặc preprocess_val của OpenCLIP, nhằm đảm bảo dữ liệu đầu vào phù hợp với giai đoạn huấn luyện ban đầu. Kích thước ảnh đầu vào được điều chỉnh theo yêu cầu của từng mô hình. Quy trình chuẩn hóa như mean và standard deviation cho từng kênh màu cũng được giữ nguyên theo cấu hình gốc của từng mô hình, không áp dụng thêm các biến đổi khác như data augmentation hay center crop tùy biến. Cuối cùng, toàn bộ embedding được chuẩn hóa L2 trước khi tính độ tương đồng cosine nhằm loại bỏ ảnh hưởng của độ lớn véc-tơ và đảm bảo tính nhất quán giữa các mô hình.

A. NHÓM MÔ HÌNH PHỔ QUÁT

Nhóm này gồm các biến thể của CLIP được huấn luyện trên những bộ dữ liệu Internet quy mô lớn và không được thiết kế riêng cho y tế hay da liễu. Vì vậy, hiệu năng của chúng khi áp dụng vào ảnh bệnh lý chủ yếu phản ánh khả năng khái quát hóa thị giác của mô hình hơn là mức độ phù hợp với miền dữ liệu da liễu.

Bảng 4. *Hiệu năng truy xuất của các mô hình phổ quát*

Mô hình	Dim	Top-1	Top-5	Top-10	Top-20	Thời gian (s/ảnh)
MetaCLIP (H14)	1024	94,08	99,46	99,77	99,85	0,0228
SigLIP (SO400M)	1152	93,31	98,62	99,62	99,77	0,0515
OpenAI-CLIP (Large-P14)	768	93,46	99,00	99,77	99,92	0,0114
ConvNeXt-CLIP (Base)	640	91,62	97,77	98,85	99,46	0,0093
AltCLIP (Base)	768	93,46	99,00	99,77	99,92	0,0103
OpenAI-CLIP (Base-P32)	512	91,69	97,62	98,92	99,38	0,0011
MobileCLIP (S2)	512	90,69	97,77	98,62	99,62	0,0058

Kết quả cho thấy một xu hướng khá rõ: với các mô hình không được tinh chỉnh cho y khoa, mô hình càng lớn và được huấn luyện trên dữ liệu càng nhiều thì hiệu năng truy xuất càng cao. Điều này cho thấy các mô hình phổ quát vẫn có thể học được những đặc trưng thị giác đủ tốt để chuyển sang bài toán da liễu. Trong nhóm này, MetaCLIP (H14-FullCC2.5B) [31] đạt Top-1 cao nhất với 94,08%, tiếp theo là AltCLIP và OpenAI-CLIP ViT-Large/14 [8] cùng đạt 93,46%, còn SigLIP SO400M [32] đạt 93,31%. Ở Top-5, MetaCLIP đạt gần 99,5%, trong khi AltCLIP, OpenAI-CLIP ViT-Large/14 và SigLIP đều vượt 98,5%. Được huấn luyện trên dữ liệu Internet quy mô lớn, các mô hình này học được nhiều đặc trưng thị giác cơ bản như màu sắc, kết cấu, hoa văn và hình dạng tổng thể nên vẫn hoạt động tốt trên ảnh da liễu, kể cả khi các bệnh có biểu hiện khá giống nhau. AltCLIP [33] đạt kết quả tương đương OpenAI-CLIP ViT-Large/14 dù chỉ dùng kiến trúc ViT-Base nhỏ hơn. Điều này cho thấy huấn luyện đa ngôn ngữ không làm giảm chất lượng biểu diễn thị giác, thậm chí có thể giúp mô hình học được các đặc trưng ổn định hơn.

Ngược lại, các mô hình nhỏ hơn hoặc ưu tiên tốc độ suy luận cho kết quả thấp hơn. OpenAI-CLIP ViT-Base/32 [8] đạt 91,69%, ConvNeXt-CLIP đạt 91,62% và MobileCLIP S2 đạt 90,69%. Mức chênh lệch khoảng 2–3% so với MetaCLIP cho thấy quy mô mô hình và dữ liệu huấn luyện vẫn ảnh hưởng rõ đến khả năng tách các lớp bệnh có hình thái phức tạp. Tuy vậy, ngay cả những mô hình phổ quát nhỏ nhất vẫn vượt 90% Top-1, cho thấy khả năng chuyển giao khá tốt của các đặc trưng thị giác học từ dữ liệu Internet sang ảnh da liễu lâm sàng. Mức này cũng cao hơn đáng kể so với các mô hình y khoa tổng quát lệch miền như RCLIP với Top-1 chỉ đạt 76,92%. Nhìn chung, ở nhóm mô hình phổ quát, quy mô mô hình và dữ liệu huấn luyện ảnh hưởng khá rõ đến chất lượng embedding, ngay cả khi áp dụng cho miền chuyên biệt như da liễu. Toàn bộ mô hình trong nhóm đều đạt trên 90% Top-1 ở bài toán truy xuất ảnh theo ảnh, cho thấy khả năng chuyển giao tốt của các đặc trưng thị giác sang ảnh da liễu lâm sàng.

B. NHÓM MÔ HÌNH Y KHOA TỔNG QUÁT

Bảng 5. *Hiệu năng truy xuất của các mô hình y khoa tổng quát*

Mô hình	Dim	Top-1	Top-5	Top-10	Top-20	Thời gian (s/ảnh)
BiomedCLIP PubMed	512	89,38	96,46	98,00	99,08	0,0024
MedSigLIP	1152	93,15	98,85	99,54	99,92	0,0174
PLIP Base	512	90,92	97,46	98,92	99,69	0,0010
ClipMD Base	512	87,92	95,85	98,46	99,23	0,0010
RCLIP Base	512	76,92	88,92	93,46	96,38	0,0117

Kết quả thực nghiệm cho thấy mô hình có chứa nhiều ảnh da lâm sàng trong quá trình huấn luyện thường cho embedding tốt hơn so với những mô hình chỉ mang nhãn 'y khoa' nhưng lệch miền dữ liệu về phương thức và phổ bệnh. RCLIP [40] cho kết quả thấp nhất với Top-1 chỉ đạt 76,92%. Nguyên nhân chủ yếu là vì mô hình này được huấn luyện chủ yếu trên ảnh X-quang và CT, vốn rất khác với ảnh màu chụp bề mặt da nên khả năng chuyển giao bị hạn chế. ClipMD đạt 87,92%, cao hơn RCLIP nhờ dữ liệu huấn luyện đa dạng hơn nhưng vẫn còn hạn chế do quy mô nhỏ và số lượng ảnh da chưa nhiều. BiomedCLIP đạt 89,38%, cải thiện rõ rệt nhờ được huấn luyện trên hơn 15 triệu cặp ảnh–văn bản từ PubMed Central với nhiều loại ảnh lâm sàng khác nhau. Điều này cho thấy khi dữ liệu huấn luyện gần với ảnh da lâm sàng hơn thì chất lượng embedding cũng được cải thiện đáng kể. PLIP [38] dù được phát triển cho ảnh giải phẫu bệnh vẫn đạt Top-1 = 90,92%, cao hơn BiomedCLIP. Kết quả này cho thấy các đặc trưng hình thái học từ ảnh mô học vẫn có thể chuyển giao phần nào sang ảnh da liễu, nhất là ở các yếu tố như kết cấu và phân bố tổn thương.

Trong nhóm y khoa, MedSigLIP [37] đạt kết quả cao nhất với Top-1 = 93,15%, gần tương đương các mô hình phổ quát mạnh nhất. Kết quả này nhiều khả năng đến từ kiến trúc SigLIP SO400M quy mô lớn và tập huấn luyện có chứa ảnh da liễu lâm sàng. Đổi lại, MedSigLIP có thời gian xử lý chậm hơn các mô hình nhỏ như BiomedCLIP hay PLIP do số lượng tham số và token đầu vào lớn hơn. Ngược lại, PLIP và ClipMD có tốc độ suy luận nhanh nhất nhóm với 0,0010 giây/ảnh, còn BiomedCLIP đạt 0,0024 giây/ảnh nhờ sử dụng kiến trúc ViT-Base gọn nhẹ. Nhìn chung, hiệu năng của nhóm y khoa tăng dần theo mức độ xuất hiện của ảnh da liễu trong tập huấn luyện. Những mô hình lệch miền mạnh như RCLIP cho kết quả thấp, trong khi các mô hình có dữ liệu gần với ảnh da lâm sàng hơn đạt hiệu năng cao hơn rõ rệt. Điều này cho thấy yếu tố quyết định nằm ở mức độ phù hợp của dữ liệu huấn luyện, chứ không chỉ ở nhãn "y khoa".

C. NHÓM MÔ HÌNH CHUYÊN BIỆT DA LIỄU

Bảng 6. *Hiệu năng truy xuất của các mô hình chuyên biệt da liễu*

Mô hình	Dim	Top-1	Top-5	Top-10	Top-20	Thời gian (s/ảnh)
DermLIP	512	94,15	98,85	99,62	99,92	0,0026
WhyLesionCLIP ViT-L/14	768	91,77	97,69	99,00	99,85	0,0095

Nhóm mô hình chuyên biệt cho thấy khá rõ hiệu quả của việc tối ưu theo đúng miền dữ liệu. DermLIP (ViT-B/16) [4] đạt Top-1 = 94,15%, gần như tương đương MetaCLIP H14 [31] với 94,08% dù hai mô hình có chênh lệch rất lớn về quy mô. Để kiểm tra sự khác biệt này, nghiên cứu thực hiện bootstrap 95% CI với 10.000 lần lấy mẫu, McNemar exact test và paired permutation test với 10.000 lần hoán vị trên toàn bộ 1.300 truy vấn. Kết quả cho thấy hiệu số Top-1 chỉ +0,08 điểm phần trăm, bootstrap 95% CI = [−0,77%, +0,92%], còn McNemar exact test và paired permutation test đều cho $p = 1,0$ ở mức ý nghĩa $\alpha = 0,05$. Điều này không có nghĩa hai mô hình hoàn toàn giống nhau, mà chỉ cho thấy chưa có đủ bằng chứng thống kê để kết luận có khác biệt đáng kể trong điều kiện thí nghiệm hiện tại. Kết quả kiểm định McNemar cho thấy trong 1.300 truy vấn, hai mô hình cùng dự đoán đúng ở 1.209 trường hợp và cùng sai ở 62 trường hợp. Chỉ có 29 trường hợp cho kết quả khác nhau, trong đó DermLIP đúng còn MetaCLIP sai ở 15 trường hợp, và ngược lại là 14 trường hợp. Sự chênh lệch gần như cân bằng này khá đáng chú ý khi MetaCLIP H14 có khoảng 1 tỷ tham số với embedding 1.024 chiều, còn DermLIP chỉ dùng kiến trúc ViT-Base/16 khoảng 86 triệu tham số với embedding 512 chiều. Điều này cho thấy huấn luyện trên dữ liệu chuyên biệt có thể giúp một mô hình nhỏ đạt hiệu năng tương đương mô hình phổ quát lớn hơn rất nhiều.

Về tốc độ xử lý, DermLIP đạt 0,0026 giây/ảnh, nhanh hơn MetaCLIP H14 (0,0228 giây/ảnh) khoảng 9 lần và nhanh hơn MedSigLIP (0,0174 giây/ảnh) khoảng 6,7 lần. Trong nhóm ViT-Base/16, tốc độ của DermLIP gần với BiomedCLIP (0,0024 giây/ảnh) và chậm hơn một chút so với PLIP hay ClipMD (0,0010 giây/ảnh). Khác biệt nhỏ này chủ yếu đến từ cách nạp mô hình qua thư viện open_clip và transformers hơn là chi phí tính toán lõi. Nhìn chung, DermLIP cho thấy khả năng cân bằng khá tốt giữa hiệu năng truy xuất và chi phí tính toán. Ở các mức Top-K cao hơn, DermLIP đạt 98,85% ở Top-5, 99,62% ở Top-10 và 99,92% ở Top-20, tức là gần như toàn bộ truy vấn đều tìm được ít nhất một ảnh cùng lớp bệnh trong 20 kết quả gần nhất. Tuy vậy, DermLIP không phải mô hình duy nhất đạt mức này. AltCLIP, OpenAI-CLIP ViT-Large/14 và MedSigLIP cũng đạt 99,92%, còn MetaCLIP đạt 99,85%. Điều này cho thấy khác biệt giữa các mô hình thể hiện rõ hơn ở Top-1 và Top-5, nơi khả năng phân biệt các đặc trưng hình thái nhỏ đóng vai trò quan trọng hơn.

WhyLesionCLIP [12] dù được phát triển cho da liễu nhưng chỉ đạt Top-1 = 91,77%, thấp hơn DermLIP khá rõ. Nguyên nhân chủ yếu nằm ở cách huấn luyện. WhyLesionCLIP sử dụng kiến trúc Concept Bottleneck kết hợp tri thức y khoa từ tài liệu chuyên ngành và PubMed, với embedding tập trung nhiều vào các khái niệm lâm sàng như bất đối xứng, viền không đều hay mạng sắc tố bất thường. Cách tiếp cận này giúp tăng khả năng diễn giải nhưng lại không tối ưu trực tiếp cho độ tương đồng thị giác giữa các ảnh. Ngoài ra, dữ liệu huấn luyện của mô hình chủ yếu là ảnh nội soi da từ các bộ ISIC như HAM10000 [45], trong khi nghiên cứu hiện tại sử dụng ảnh lâm sàng bề mặt da của nhóm bệnh nhiễm trùng. Sự khác biệt về cả loại ảnh lẫn phổ bệnh lý làm giảm khả năng chuyển giao của mô hình sang dữ liệu đích. Dù WhyLesionCLIP sử dụng embedding 768 chiều, lớn hơn DermLIP với 512 chiều, điều này không mang lại lợi thế rõ rệt cho bài toán truy xuất ảnh. Trường hợp này cho thấy hai yếu tố quyết định chất lượng embedding cho bài toán truy xuất ảnh quan trọng hơn kích thước véc-tơ: (1) mục tiêu huấn luyện — tối ưu cho diễn giải (như Concept Bottleneck) khác với tối ưu cho độ tương đồng thị giác trực tiếp; (2) mức độ phù hợp của dữ liệu huấn luyện với dữ liệu đích về phương thức (nội soi da vs lâm sàng bề mặt) và phổ bệnh (ung thư da vs nhiễm trùng). Hệ quả thực tiễn: khi chọn backbone cho pipeline da liễu, cần ưu tiên mô hình huấn luyện trực tiếp trên độ tương đồng ảnh và có dữ liệu cùng phương thức/phổ bệnh với ứng dụng đích, thay vì chỉ dựa vào nhãn "chuyên biệt da liễu" hay kích thước embedding lớn hơn.

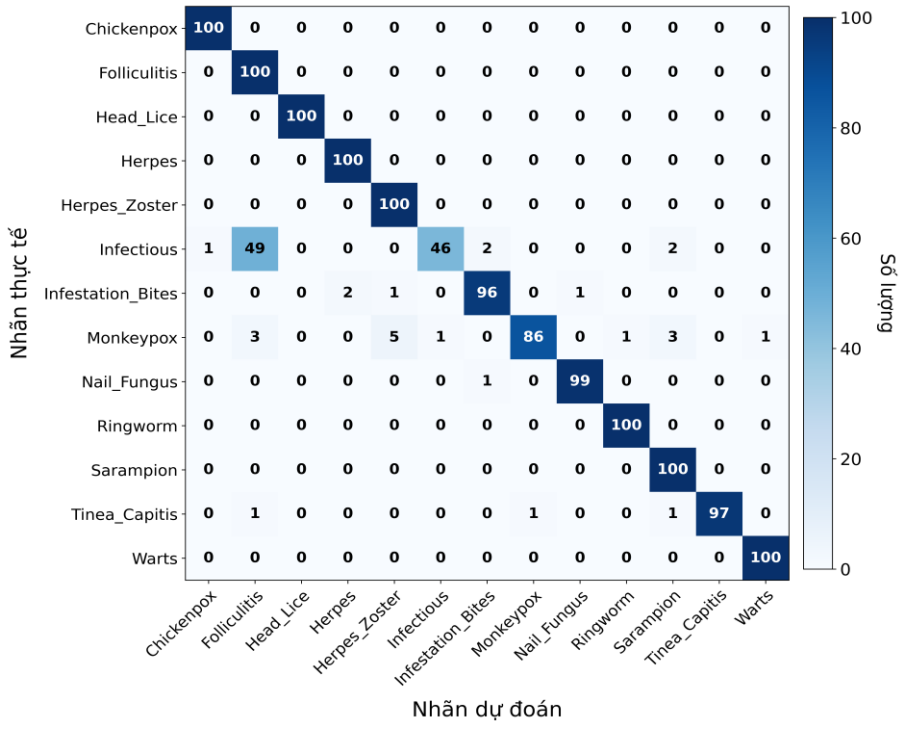
D. PHÂN TÍCH NHẦM LẪN THEO LỚP BỆNH

Bảng 7 trình bày Top-1 theo từng lớp của DermLIP [4] và MetaCLIP [31], còn Hình 5 và Hình 6 minh họa ma trận nhầm lẫn của hai mô hình trên 1.300 lượt truy xuất ảnh theo ảnh. Ở mức tổng thể, DermLIP đạt Top-1 trung bình 94,15% và MetaCLIP đạt 94,08% (gần như bằng nhau), giống nhau ở phần lớn các lớp: cùng đạt 100% trên Folliculitis, Head Lice, Herpes, Herpes Zoster, Ringworm, Sarampion và Warts; cùng đạt 99% ở Nail Fungus và 97% ở Tinea Capitis. Khác biệt chỉ xuất hiện ở các lớp có hình thái phức tạp hơn: DermLIP cao hơn ở Chickenpox

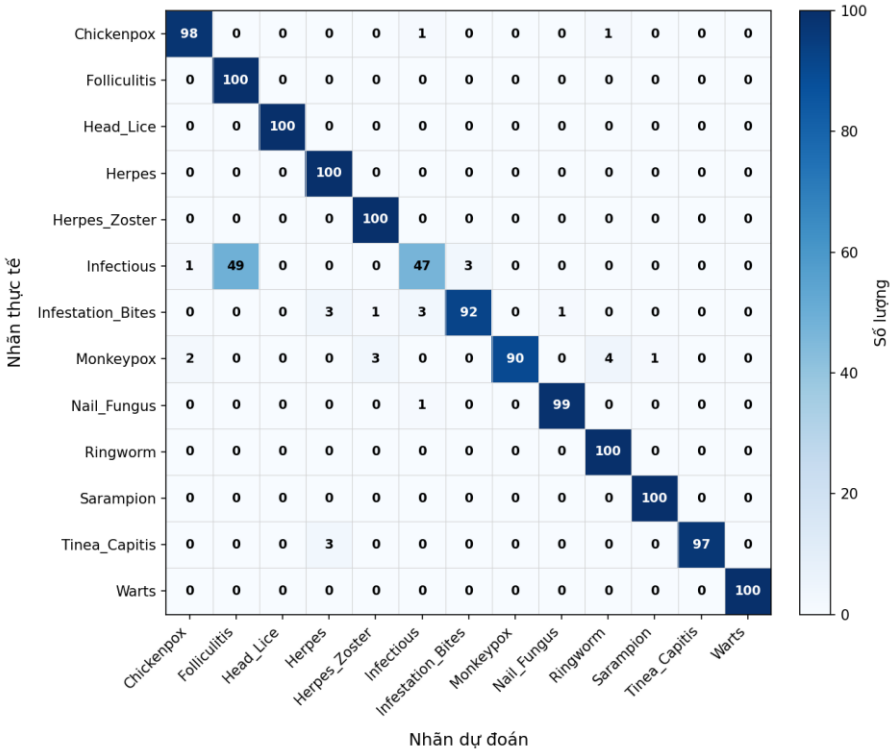
(100% so với 98%) và Infestation/Bites (96% so với 92%), còn MetaCLIP cao hơn ở Monkeypox (90% so với 86%) và Infectious (47% so với 46%).

Bảng 7. *So sánh Top-1 theo từng lớp bệnh giữa DermLIP (ViT-B/16) và MetaCLIP (H14). Cột "Chênh lệch" = (số đoán đúng của DermLIP) − (số đoán đúng của MetaCLIP); ở dòng "Tổng / TB", chênh lệch +1 ca tương ứng hiệu số Top-1 +0,08 điểm phần trăm.*

Bệnh	Số mẫu	DermLIP Đoán đúng	DermLIP Top-1 (%)	MetaCLIP Đoán đúng	MetaCLIP Top-1 (%)	Chênh lệch
Chickenpox	100	100	100	98	98	+2
Folliculitis	100	100	100	100	100	0
Head Lice	100	100	100	100	100	0
Herpes	100	100	100	100	100	0
Herpes Zoster	100	100	100	100	100	0
Infectious	100	46	46	47	47	−1
Infestation/Bites	100	96	96	92	92	+4
Monkeypox	100	86	86	90	90	−4
Nail Fungus	100	99	99	99	99	0
Ringworm	100	100	100	100	100	0
Sarampion	100	100	100	100	100	0
Tinea Capitis	100	97	97	97	97	0
Warts	100	100	100	100	100	0
Tổng / TB	**1.300**	**1.224**	**94,15**	**1.223**	**94,08**	**+1**



Hình 5. *Ma trận nhầm lẫn – Mô hình DermLIP (ViT-B/16)*



Hình 6. *Ma trận nhầm lẫn – Mô hình MetaCLIP (H14)*

Xét ma trận nhầm lẫn của DermLIP, mô hình đạt 100% Top-1 trên 8/13 lớp. Ở Monkeypox, 14 trường hợp sai được phân bố như sau: 5 ca sang Herpes Zoster, 3 ca sang Folliculitis, 3 ca sang Sarampion, và 3 ca lẻ sang Warts, Infectious và Ringworm (mỗi bệnh 1 ca) — phần lớn (9/14 ca) rơi vào các bệnh có biểu hiện mụn nước tương tự nhau trên ảnh lâm sàng. Lớp Infectious có kết quả thấp nhất với 46%, phần lớn (49/54 ca) ảnh bị nhầm sang Folliculitis; 5 ca còn lại phân tán sang Chickenpox (1 ca), Infestation/Bites (2 ca) và Sarampion (2 ca). Rà soát thủ công cho thấy nhiều ảnh trong nhóm Infectious quả thực có hình thái gần với Folliculitis – một phần sai số do vậy có thể bắt nguồn từ tính tổng quát và thiếu nhất quán của nhãn nguồn, chứ không hẳn là giới hạn của mô hình. Khi loại lớp Infectious khỏi đánh giá, Top-1 của DermLIP tăng từ 94,15% lên 98,17%, còn MetaCLIP tăng từ 94,08% lên 98,00%. Xu hướng tương tự lặp lại trên toàn bộ các mô hình mà không làm xáo trộn thứ hạng. Như vậy chất lượng nhãn có tác động đáng kể tới kết quả truy xuất, đặc biệt ở những lớp mang tính tổng quát; bước xác nhận nhãn bởi chuyên gia da liễu được kiến nghị cho các nghiên cứu mở rộng (xem mục Giới hạn nghiên cứu).

Đối với MetaCLIP, mô hình đạt 100% trên 7 lớp bệnh, ít hơn DermLIP một lớp. Khác biệt rõ nhất nằm ở Chickenpox và Infestation/Bites, nơi DermLIP cho kết quả cao hơn – bằng chứng cho ưu thế của mô hình chuyên biệt khi học các đặc trưng hình thái đặc trưng của da liễu. Ngược lại, MetaCLIP cao hơn ở Monkeypox, gợi ý rằng dữ liệu Internet quy mô lớn đem lại lợi thế ở những bệnh có biểu hiện không điển hình nhờ kho đặc trưng thị giác tổng quát phong phú. Nói chung, dù không có khác biệt thống kê (như đã trình bày ở mục IV.C), phân tích theo từng lớp cho thấy hai mô hình vận hành theo hai cơ chế biểu diễn khác nhau: DermLIP nắm bắt tốt đặc trưng hình thái đặc hiệu của da liễu, còn MetaCLIP khái quát hóa mạnh hơn ở những lớp có biểu hiện thị giác đa dạng hoặc ít đặc trưng.

E. KIỂM TRA CONTAMINATION

Để định lượng khả năng các mô hình chuyên biệt da liễu – nhất là DermLIP – đã "nhìn thấy" ảnh trong tập đánh giá ở giai đoạn huấn luyện, nghiên cứu áp dụng quy trình kiểm tra trùng lặp ở cấp độ ảnh bằng perceptual hashing (pHash). Quy trình gồm năm bước: (i) tính pHash cho toàn bộ 1.300 ảnh trong tập đánh giá; (ii) tính pHash cho subset của tập huấn luyện Derm1M [4] có liên quan đến phổ bệnh nhiễm trùng; (iii) tìm mọi cặp ảnh có Hamming distance ≤ 10 (ngưỡng chặt, đủ bắt các biến thể nhỏ về crop, độ sáng hay nén ảnh); (iv) hai thành viên trong nhóm rà soát thủ công bằng mắt từng cặp được pHash đánh dấu, phân loại vào bốn mức loại trừ nhau: YES (trùng hoàn toàn, cùng ảnh nguồn), SAME_LESION (cùng tổn thương trên cùng bệnh nhân, chụp khác góc hoặc khác thời điểm), SAME_DISEASE (hai ảnh của hai bệnh nhân khác nhau nhưng cùng bệnh lý) và NO (cặp có pHash khớp nhưng nhãn bệnh không khớp giữa hai tập); (v) tổng hợp kết quả theo từng lớp bệnh. Hình 7, Hình 8 và Hình 9 minh họa lần lượt các trường hợp YES, SAME_DISEASE và contamination (cặp pHash khớp nhưng khác nhãn) thu được từ quy trình rà soát thủ công.

Hình 7. *Ví dụ hạng mục YES – cặp ảnh trùng hoàn toàn giữa tập đánh giá và Derm1M (cùng ảnh nguồn, cùng nhãn Herpes).*

Hình 8. *Ví dụ hạng mục SAME_DISEASE – hai bệnh nhân khác nhau cùng mắc Chickenpox; cùng nhãn nên không tính là contamination theo định nghĩa nghiêm ngặt.*

Hình 9. *Ví dụ contamination (nhãn "NO" theo quy ước rà soát thủ công) – cặp ảnh có pHash khớp (Hamming ≤ 10) nhưng khác nhãn giữa tập đánh giá (Warts) và Derm1M (Chickenpox); đây là trường hợp được tính vào 51 cặp NO (trong tổng 63 ảnh contamination) ở Bảng 8. (Lưu ý: hạng mục SAME_LESION không được minh họa do trên 1.300 ảnh đánh giá không phát hiện cặp nào thuộc loại này – xem cột SAME_LESION trong Bảng 8.)*

Một điểm về định nghĩa cần làm rõ: nghiên cứu này tách hai khái niệm. (1) Ở cấp độ cặp ảnh, mỗi cặp pHash khớp được phân loại vào bốn nhóm loại trừ nhau là YES, SAME_LESION, SAME_DISEASE và NO — trong đó NO là cặp có pHash khớp nhưng nhãn ứng với hai bệnh khác nhau giữa tập đánh giá và Derm1M (loại trùng lặp nhãn nghiêm trọng nhất vì làm sai lệch thông tin nhãn). Tổng 158 = 102 YES + 0 SAME_LESION + 5 SAME_DISEASE + 51 NO. (2) Ở cấp độ ảnh, cột "Số ảnh contamination xác nhận" đếm số ảnh đánh giá duy nhất có ít nhất một cặp pHash khớp với Derm1M (bất kể thuộc loại nào sau rà soát thủ công), vì một ảnh có thể khớp với nhiều ảnh trong Derm1M và sinh ra nhiều cặp ở nhiều loại khác nhau. Cách tách này cho phép báo cáo đồng thời (a) mức trùng lặp nhãn thực sự thông qua NO — chỉ 51/158 cặp (~32,3%) — và (b) phạm vi rộng hơn của khả năng dữ liệu trùng lặp qua 63 ảnh eval (~4,85% của 1.300) có ít nhất một cặp pHash khớp. Hai con số này cho cùng một kết luận: mức ảnh hưởng của contamination đến kết quả truy xuất là nhỏ.

Bảng 8. *Kết quả kiểm tra contamination giữa tập đánh giá 1.300 ảnh và Derm1M bằng perceptual hashing (pHash, Hamming ≤ 10)*

Lớp bệnh	Số ảnh eval	Cặp pHash khớp (raw)	YES (cặp trùng hoàn toàn)	SAME_LESION (cùng tổn thương)	SAME_DISEASE (cùng bệnh, khác BN)	NO (cặp khác nhãn)	Số ảnh contamination xác nhận	Tỷ lệ (%)
Chickenpox	100	10	5	0	4	1	3	3,0
Folliculitis	100	0	0	0	0	0	0	0,0
Head Lice	100	0	0	0	0	0	0	0,0
Herpes	100	39	21	0	0	18	18	18,0
Herpes Zoster	100	5	3	0	0	2	3	3,0
Infectious	100	21	16	0	1	4	7	7,0
Infestation /Bites	100	6	4	0	0	2	4	4,0
Monkeypox	100	45	32	0	0	13	11	11,0
Nail Fungus	100	0	0	0	0	0	0	0,0
Ringworm	100	8	5	0	0	3	4	4,0
Sarampion	100	15	10	0	0	5	7	7,0
Tinea Capitis	100	3	2	0	0	1	2	2,0
Warts	100	6	4	0	0	2	4	4,0
Tổng	**1.300**	**158**	**102**	**0**	**5**	**51**	**63**	**4,85**

Kết quả kiểm tra trên 13 lớp bệnh được tổng hợp trong Bảng 8. Tổng số 158 cặp pHash khớp được phân thành 102 YES, 0 SAME_LESION, 5 SAME_DISEASE và 51 NO; như vậy ngay ở cấp độ cặp, chỉ ~32,3% cặp pHash khớp thực sự dẫn đến trùng lặp nhãn nghiêm trọng. Theo định nghĩa đã nêu, có 63/1.300 ảnh (4,85%) được xác nhận là contamination. Tỷ lệ này thay đổi rất lớn giữa các lớp, từ 0% ở Folliculitis, Head Lice và Nail Fungus đến 18% ở Herpes. Hai lớp đáng lưu ý nhất là Herpes (18% contamination, 18/100 ảnh) và Monkeypox (11%, 11/100 ảnh) — đây là các lớp có nhiều ảnh trùng với Derm1M nhất do nguồn ảnh y khoa của hai bệnh này được chia sẻ rộng rãi trên Internet. Cần phân tích kỹ ảnh hưởng đến kết quả truy xuất của các lớp này, đặc biệt ở Herpes nơi 18 trên 100 ảnh đánh giá có cặp pHash khớp với tập huấn luyện Derm1M. Riêng ba lớp Folliculitis, Head Lice và Nail Fungus không có cặp pHash nào khớp ngay từ bước sàng lọc đầu, phù hợp với giả thuyết rằng các tập nguồn của ba lớp này (Dermatology_hair và Skin-Disease-Dataset) không thuộc phạm vi của Derm1M.

Bảng 9. *Phân tích độ nhạy: Top-1 của 14 mô hình trước và sau khi loại bỏ 63 ảnh contamination*

Mô hình	Top-1 toàn bộ (%)	Số ảnh (toàn bộ)	Top-1 sau loại contamination (%)	Số ảnh (sạch)	Chênh lệch (điểm %)
DermLIP	94,15	1.300	94,18	1.237	+0,03
MetaCLIP H14	94,08	1.300	94,10	1.237	+0,02
AltCLIP Base	93,46	1.300	93,37	1.237	−0,09
OpenAI-CLIP Large P14	93,46	1.300	93,37	1.237	−0,09
SigLIP SO400M	93,31	1.300	93,29	1.237	−0,02
MedSigLIP	93,15	1.300	93,05	1.237	−0,10
WhyLesionCLIP ViT-L/14	91,77	1.300	91,67	1.237	−0,10
OpenAI-CLIP Base P32	91,69	1.300	91,75	1.237	+0,06
ConvNeXt-CLIP Base	91,62	1.300	91,59	1.237	−0,03
MobileCLIP S2	90,69	1.300	90,70	1.237	+0,01
PLIP Base	90,92	1.300	91,11	1.237	+0,19
BiomedCLIP PubMed	89,38	1.300	89,33	1.237	−0,05
ClipMD Base	87,92	1.300	88,20	1.237	+0,28
RCLIP Base	76,92	1.300	76,96	1.237	+0,04

Bảng 9 tổng hợp kết quả phân tích độ nhạy sau khi loại 63 ảnh contamination, làm tập đánh giá giảm từ 1.300 xuống 1.237 ảnh. Có thể rút ra ba nhận xét. Trước hết, Top-1 của 14 mô hình chỉ thay đổi nhẹ, từ −0,10 điểm phần trăm (WhyLesionCLIP và MedSigLIP) đến +0,28 điểm phần trăm (ClipMD); 7 mô hình tăng nhẹ và 7 mô hình giảm nhẹ với biên độ hầu hết dưới 0,10 pp. Việc phân bố hai chiều và biên độ nhỏ này cho thấy 63 contamination không có hiệu ứng hệ thống lên hiệu năng — ảnh hưởng nằm trong ngưỡng nhiễu thống kê. Tiếp theo, thứ hạng giữa 14 mô hình không thay đổi: DermLIP và MetaCLIP H14 vẫn ở vị trí dẫn đầu, RCLIP vẫn xếp cuối, và phân bố ba nhóm phổ quát – y khoa – chuyên biệt vẫn giữ nguyên. Cụ thể, Top-1 của DermLIP tăng nhẹ +0,03 điểm sau khi loại contamination (94,15% → 94,18%), tương đương MetaCLIP H14 (+0,02). Nếu DermLIP hưởng lợi bất cân xứng từ contamination do được huấn luyện trên Derm1M, thì khi loại contamination Top-1 của DermLIP phải giảm mạnh hơn các mô hình khác — nhưng thực tế DermLIP còn tăng nhẹ và mức biến động cũng không khác biệt rõ với các mô hình ngoài nhóm chuyên biệt. Quan sát này cho phép bác bỏ giả thuyết rằng hiệu năng cao của DermLIP chủ yếu đến từ việc đã thấy dữ liệu đánh giá, đồng thời củng cố luận điểm trung tâm của bài báo: tri thức miền chuyên biệt mới là yếu tố quyết định, độc lập với mức contamination quan sát được ở đây.

Sau khi loại contamination, hiệu số Top-1 giữa DermLIP và MetaCLIP H14 vẫn ở mức +0,08 điểm phần trăm (94,15% so với 94,08% trước khi loại; 94,18% so với 94,10% sau khi loại), nằm trọn trong khoảng tin cậy bootstrap 95% [−0,77%, +0,92%] đã báo cáo ở mục IV.C. Do đó, kết luận về tính tương đương thống kê giữa hai mô hình không bị thay đổi bởi việc loại bỏ contamination. Tổng kết lại, phân tích contamination cho thấy mức contamination thực tế tương đối thấp (4,85%), phân bố tương đối đều giữa các nhóm mô hình, và không làm thay đổi các kết luận chính của nghiên cứu. Dù vậy, một tập đánh giá hold-out hoàn toàn độc lập, lấy từ các nguồn không nằm trong Derm1M, vẫn là bổ sung cần cân nhắc cho các nghiên cứu mở rộng để loại bỏ triệt để ảnh hưởng của contamination.

V. THẢO LUẬN

Kết quả thực nghiệm phức tạp hơn giả định ban đầu rằng mô hình chuyên biệt luôn vượt trội trên miền dữ liệu tương ứng: DermLIP đạt hiệu năng ngang MetaCLIP H14 dù nhỏ hơn khoảng 10 lần về số tham số, trong khi WhyLesionCLIP cũng chuyên biệt da liễu lại thấp hơn nhiều mô hình phổ quát. Nhóm y khoa tổng quát không có ưu thế như giả định ban đầu; thậm chí RCLIP còn xếp cuối toàn bộ thí nghiệm.

A. ẢNH HƯỞNG CỦA MIỀN DỮ LIỆU HUẤN LUYỆN ĐẾN CHẤT LƯỢNG EMBEDDING

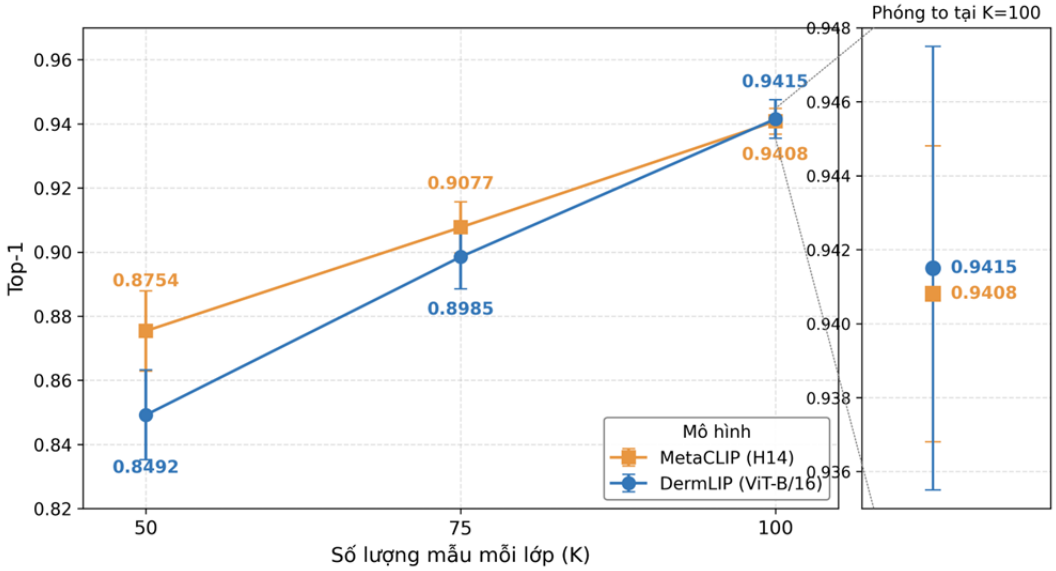
Khác biệt giữa ba nhóm mô hình – phổ quát, y khoa tổng quát, chuyên biệt da liễu – thể hiện rõ qua kết quả. Mô hình huấn luyện trên dữ liệu càng gần với miền da liễu thì không gian embedding tạo ra càng ổn định và phù hợp với bài toán truy xuất ảnh da liễu. Điều này nhất quán với kết quả của Khun Jush et al. [5]: embedding tiền huấn luyện vẫn có thể cho hiệu năng truy xuất ảnh y tế có ý nghĩa mà không cần huấn luyện lại, miễn là không gian embedding đủ tốt để phản ánh các đặc trưng quan trọng. Nghiên cứu này đi xa hơn một bước: mức độ "đủ tốt" phụ thuộc nhiều vào độ phù hợp giữa miền dữ liệu huấn luyện và miền ứng dụng, không chỉ vào quy mô kiến trúc.

Ở nhóm phổ quát, các kiến trúc quy mô lớn như MetaCLIP [31] và Google-SigLIP [32] đạt kết quả khá cạnh tranh: MetaCLIP đạt Top-1 = 94,08%, ngang với DermLIP chuyên biệt. Điều này cho thấy quy mô dữ liệu huấn luyện và độ đa dạng thị giác có thể bù lại được phần nào sự thiếu hụt về tri thức miền – tương tự nhận định của Raghu et al. [6] rằng đặc trưng từ mô hình lớn có tính chuyển giao đáng kể sang ảnh y tế. Tuy nhiên, phân tích nhầm lẫn theo lớp cho thấy mô hình phổ quát vẫn dựa nhiều vào đặc trưng thị giác tổng quát: MetaCLIP đạt 90% trên Monkeypox (nhờ vốn đặc trưng thị giác đa dạng) nhưng chỉ 98% trên Chickenpox – lớp mà DermLIP đạt 100% nhờ học được các đặc trưng hình thái da liễu điển hình. Kết quả ở nhóm y khoa tổng quát cũng phức tạp tương tự. Khác với các benchmark y sinh hiện có, nơi BiomedCLIP [10] vẫn được ghi nhận vượt trội ở nhiều bài toán y tế, các mô hình y khoa trong nghiên cứu này lại không thể hiện được ưu thế khi áp dụng cho ảnh da liễu lâm sàng bề mặt. RCLIP (76,92%) thấp hơn mọi mô hình phổ quát; BiomedCLIP (89,38%) cũng không vượt được ngưỡng 90% mà ngay cả MobileCLIP S2 (90,69%) – một mô hình phổ quát quy mô nhỏ (~35M tham số) – đã chạm tới. Cần lưu ý rằng MobileCLIP S2 được huấn luyện trên DataCompDR – tập dữ liệu được tăng cường (reinforced) bằng caption tổng hợp từ mô hình lớn, do đó việc nó vượt BiomedCLIP một phần nhờ chiến lược dữ liệu chứ không hoàn toàn nhờ kiến trúc đơn thuần. Tuy nhiên, ngay cả khi tính đến yếu tố này, kết luận chính vẫn đứng vững: trong miền da liễu lâm sàng bề mặt, các mô hình y khoa được huấn luyện trên dữ liệu lệch miền (X-quang, giải phẫu bệnh) không tự động kế thừa được lợi thế từ nhãn "y khoa". Một góc nhìn cần được bổ sung vào tài liệu hiện có: BiomedCLIP tuy mạnh trên các bài toán X-quang, giải phẫu bệnh và nhãn khoa [10], nhưng hiệu năng đó không tự động chuyển giao sang ảnh da liễu lâm sàng bề mặt – loại ảnh có đặc thù riêng về màu sắc, bề mặt, nền và bối cảnh chụp, khác hẳn các loại ảnh y tế thường thấy trong benchmark. Như vậy, các benchmark y sinh cần đưa ảnh da liễu lâm sàng vào như một miền kiểm tra bắt buộc, thay vì mặc nhiên cho rằng kết quả ở các miền y tế khác sẽ chuyển giao được sang đây.

MedSigLIP [37] đạt kết quả cao nhất trong nhóm y khoa: đạt Top-1 = 93,15%, chỉ kém DermLIP và MetaCLIP H14 khoảng 0,93–1,00 điểm phần trăm, đồng thời cao hơn các mô hình y khoa lệch miền (RCLIP 76,92%, ClipMD 87,92%, BiomedCLIP 89,38%, PLIP 90,92%). Vị trí của MedSigLIP tạo nên một phép so sánh ba chiều có giá trị: DermLIP đại diện cho mô hình chuyên biệt da liễu quy mô nhỏ (~86M tham số), MetaCLIP H14 đại diện cho mô hình phổ quát quy mô rất lớn (~1B tham số), và MedSigLIP đại diện cho mô hình y khoa đa chuyên khoa quy mô trung bình (vision encoder sử dụng ~400M tham số) có bao phủ da liễu trong tập huấn luyện. Việc cả ba đạt trong dải 93–94% Top-1, trong khi các mô hình y khoa lệch miền thấp hơn 4–17 điểm phần trăm, củng cố luận điểm trung tâm: chất lượng embedding trên ảnh da liễu được quyết định bởi mức độ bao phủ ảnh da trong tập huấn luyện, không phải bởi nhãn miền chung "y khoa" hay đơn thuần quy mô kiến trúc. Khoảng cách khoảng 1 điểm phần trăm giữa MedSigLIP và DermLIP tuy nhỏ vẫn cho thấy huấn luyện chuyên biệt cho da liễu (như ở DermLIP) tạo ra cấu trúc gom cụm chặt hơn so với huấn luyện trên hỗn hợp đa chuyên khoa có bao phủ da liễu (như MedSigLIP). Vision encoder của DermLIP (~86 triệu tham số) nhỏ hơn của MedSigLIP (~400 triệu tham số) khoảng 4–5 lần, và thời gian trích xuất embedding thực tế cũng tỷ lệ tương ứng (0,0026 giây/ảnh so với 0,0174 giây/ảnh, tức nhanh hơn khoảng 6,7 lần) — DermLIP do đó có ưu thế kép về cả tốc độ lẫn hiệu năng so với MedSigLIP. Dù vậy, MedSigLIP vẫn là mô hình nền tảng cần cân nhắc khi ứng dụng đòi hỏi tính tổng quát đa miền y tế thay vì chỉ riêng da liễu.

Nhóm chuyên biệt da liễu cho kết quả phân hóa hơn dự kiến. DermLIP [4] dẫn đầu nhóm với hiệu năng cạnh tranh và ổn định trên hầu hết các chỉ số, ngang với MetaCLIP H14 – mô hình lớn hơn khoảng 10 lần – mà không có khác biệt thống kê trong các kiểm định đã thực hiện. Điều này phù hợp với phát hiện của P. Tschandl et al. [15] rằng truy xuất ảnh dựa trên nội dung dùng đặc trưng từ mạng nơ-ron chuyên biệt da liễu có thể đạt độ chính xác chẩn đoán ngang với dự đoán softmax truyền thống, và mở rộng phát hiện đó sang bối cảnh các mô hình nền tảng thị giác–ngôn ngữ hiện đại. Nhưng cùng nhóm chuyên biệt, WhyLesionCLIP chỉ đạt 91,77%, thấp hơn DermLIP và thậm chí thấp hơn một số mô hình phổ quát. Sự phân hóa nội bộ này cho thấy nhãn "chuyên biệt" không đảm bảo hiệu quả. Các yếu tố quyết định thật sự là chiến lược huấn luyện (học tương phản trực tiếp trên ảnh da lâm sàng ở DermLIP, đối lập với kiến trúc Concept Bottleneck từ textbook ở WhyLesionCLIP), loại dữ liệu (ảnh lâm sàng macro so với nội soi da ISIC) và mục tiêu tối ưu hóa. Điều này bổ sung cho quan điểm của Gassner et al. [16]: hiệu năng truy xuất dựa trên nội dung trong da liễu phụ thuộc nhiều vào độ phù hợp giữa kiến trúc trích xuất đặc trưng và loại ảnh đầu vào.

Hiệu năng cao của DermLIP có thể quy về hai yếu tố cốt lõi. Trước hết là dữ liệu huấn luyện: DermLIP được học tương phản trực tiếp trên Derm1M – tập gồm khoảng 1 triệu cặp ảnh–văn bản da liễu lâm sàng, với các thuật ngữ y khoa cụ thể như "vesicular eruption" hay "dermatomal distribution" thay vì những mô tả tổng quát kiểu "skin disease". Cách này giúp vision encoder học được các đặc trưng phân biệt được các bệnh có hình thái gần nhau và phù hợp tự nhiên với mục tiêu truy xuất ảnh. Yếu tố thứ hai là sự căn chỉnh ngữ nghĩa thông qua text encoder huấn luyện trên thuật ngữ chuyên biệt da liễu – điểm khác biệt so với các mô hình phổ quát trên dữ liệu Internet, nơi tín hiệu chuyên ngành dễ bị pha loãng. Sự kết hợp hai yếu tố này tạo ra không gian embedding gom cụm tốt theo từng lớp bệnh; DermLIP đạt Top-1 = 100% trên 8/13 lớp – mức MetaCLIP H14 dù lớn hơn khoảng 10 lần về số tham số vẫn không đạt được. Như vậy, với các bài toán da liễu chuyên biệt, huấn luyện trực tiếp trên dữ liệu cùng miền có thể hiệu quả hơn chỉ mở rộng quy mô kiến trúc hay dữ liệu tổng quát.



Hình 10. *Ảnh hưởng của số lượng ảnh mỗi lớp (K) đến Top-1 của DermLIP và MetaCLIP*

Để kiểm tra tính ổn định của thứ hạng giữa DermLIP [4] và MetaCLIP [31] – hai mô hình dẫn đầu, nghiên cứu thực hiện một phân tích độ nhạy bằng cách thay đổi số ảnh mỗi lớp trong thư viện (K = 50, 75, 100). Hình 10 cho thấy xu hướng rõ rệt: tại K = 50, MetaCLIP dẫn trước DermLIP khoảng 2,6 điểm phần trăm (0,8754 so với 0,8492); tại K = 75 khoảng cách thu hẹp còn 0,9 điểm (0,9077 so với 0,8985); và tại K = 100, hai mô hình đạt xấp xỉ cùng giá trị (DermLIP 0,9415; MetaCLIP 0,9408). DermLIP cải thiện nhanh hơn (+9,2 điểm từ K=50 đến K=100) so với MetaCLIP (+6,5 điểm), tức là hưởng lợi nhiều hơn khi thư viện có thêm ảnh cùng lớp. Hiện tượng này phản ánh cấu trúc gom cụm trong không gian embedding: DermLIP, nhờ được huấn luyện trực tiếp trên dữ liệu da liễu, tổ chức các ảnh cùng bệnh lý trong vùng lân cận nhỏ và chặt hơn. Khi thư viện còn nhỏ, số ứng viên đúng trong vùng lân cận bị giới hạn nên lợi thế này chưa phản ánh đầy đủ qua Top-1. Thư viện càng mở rộng, cấu trúc gom cụm chặt của DermLIP càng thể hiện rõ, trong khi MetaCLIP với không gian embedding phổ quát hơn nên vùng gom cụm lỏng, mức cải thiện cận biên thấp hơn. Phát hiện này gợi ý rằng việc hai mô hình đạt xấp xỉ cùng Top-1 (DermLIP 94,15%, MetaCLIP 94,08%) tại K = 100 (báo cáo ở Bảng 4 và Bảng 6) không nên được coi là kết quả ổn định ở mọi quy mô thư viện, mà phụ thuộc vào quy mô đánh giá: ở K nhỏ hơn, MetaCLIP còn lợi thế rõ rệt nhờ vốn đặc trưng thị giác đa dạng, trong khi DermLIP cần thư viện đủ lớn để cấu trúc gom cụm chặt thể hiện rõ. Tuy vậy, phân tích chỉ thực hiện tới K = 100 do giới hạn dữ liệu hiện có; xác nhận xu hướng thu hẹp hay đảo ngược thứ hạng ở quy mô lớn hơn (K = 200–500) vẫn là việc cần làm trong các nghiên cứu tiếp theo.

B. TÁC ĐỘNG CỦA VIỆC CHỈ SỬ DỤNG NHÓM BỆNH NHIỄM TRÙNG

Trong ngữ cảnh khi đọc kết quả: tập dữ liệu nghiên cứu chỉ gồm các bệnh nhiễm trùng, vốn là nhóm khó phân biệt nhất trong da liễu. Các benchmark da liễu thường gộp nhiều nhóm lớn khác nhau như u ác tính, u lành, bệnh viêm và da bình thường — ranh giới giữa những nhóm này tương đối rõ và mô hình dễ phân tách. Ở đây thì khác: 13 lớp đều là bệnh nhiễm trùng, dùng chung gần như toàn bộ dấu hiệu thị giác cơ bản (đỏ da, mụn nước, bong vảy, sẩn nhỏ, tổn thương dạng mảng), nên mô hình phải dựa vào các khác biệt rất tinh tế. Hệ quả của độ khó này thể hiện rõ ở phân tích nhầm lẫn. Lớp Infectious là ví dụ điển hình: trong 54 ảnh mà DermLIP truy xuất sai, phần lớn (49 ảnh) bị nhầm sang Folliculitis vì hai bệnh có hình thái gần như không phân biệt được trên ảnh tĩnh; 5 ảnh còn lại phân tán sang Chickenpox (1 ca), Infestation/Bites (2 ca) và Sarampion (2 ca). Tương tự, 14/100 ảnh Monkeypox bị truy xuất sai sang các bệnh khác, chủ yếu vào các bệnh có biểu hiện mụn nước ở giai đoạn đầu: 5 ca sang Herpes Zoster, 3 ca sang Folliculitis, 3 ca sang Sarampion; 3 ca còn lại phân tán sang Warts, Infectious và

Ringworm (mỗi bệnh 1 ca). Những nhầm lẫn này không phản ánh giới hạn của mô hình mà phản ánh đúng độ khó của chính bài toán. Trong thực tế, bác sĩ da liễu phân biệt các bệnh trên cũng không chỉ dựa vào hình thái tổn thương; họ kết hợp thêm bệnh sử của bệnh nhân, vùng cơ thể nơi tổn thương xuất hiện và diễn tiến theo thời gian. Mô hình chỉ có một ảnh tĩnh duy nhất, nên việc đạt 86–94% trên những lớp này đã là kết quả đáng kể. Ở ngữ cảnh đó, các giá trị Top-1 ở mức 90–94% không nên đọc như "hiệu năng trung bình", mà cần đặt vào đúng độ khó của bài toán. Việc hầu hết mô hình đạt trên 97% với Top-5 và trên 99% với Top-20 cho thấy không gian embedding vẫn giữ được cấu trúc gom cụm ngữ nghĩa ổn định: ảnh đúng bệnh lý luôn nằm trong vùng lân cận gần, chỉ bị "đẩy" khỏi Top-1 bởi ảnh thuộc bệnh có hình thái gần giống hệt. Tình huống này phù hợp với cách bác sĩ dùng hệ thống hỗ trợ chẩn đoán trên thực tế – tham khảo một nhóm ca bệnh tương tự – và là kịch bản mà Top-5, Top-10 phản ánh đúng hơn Top-1. Như vậy, Top-K với K > 1 nên được ưu tiên làm chỉ số đánh giá chính khi thiết kế hệ thống truy xuất ảnh dựa trên nội dung cho nhóm bệnh có hình thái chồng lấn cao.

C. VAI TRÒ CỦA ĐỘ ĐO TƯƠNG ĐỒNG TRONG ĐÁNH GIÁ EMBEDDING

Kết quả ở cả hai kịch bản truy xuất ảnh theo ảnh và truy xuất ảnh bằng văn bản đều xác nhận lựa chọn cosine đã được phân tích ở mục III.D, đồng thời cho thấy hai điểm đáng chú ý trong triển khai thực tế. Thứ nhất, khi embedding đã được chuẩn hóa L2, cosine và Euclid cho kết quả gần như tương đương trên cả 14 mô hình. Điều này cho thấy khác biệt giữa hai độ đo gần như không còn nhiều ý nghĩa thực tiễn, và việc cải thiện chất lượng embedding sẽ quan trọng hơn việc tối ưu độ đo khoảng cách. Thứ hai, ở bài toán truy xuất chéo phương thức, Euclid suy giảm rất mạnh do khoảng cách phương thức (modality gap), với Top-1 giảm tới 64–75% ở các mô hình như SigLIP, MobileCLIP và WhyLesionCLIP, trong khi cosine vẫn giữ được độ ổn định. Vì vậy, với các hệ thống truy xuất da liễu có tích hợp truy vấn bằng văn bản, cosine gần như là lựa chọn bắt buộc.

D. HÀM Ý ĐỐI VỚI THIẾT KẾ PIPELINE FINE-TUNING

Từ góc độ ứng dụng, kết quả nghiên cứu gợi ra một số điểm đáng chú ý cho việc thiết kế pipeline fine-tuning trong da liễu. Trước hết, việc chọn backbone không nên chỉ dựa vào độ chính xác phân loại như trong nhiều benchmark hiện nay, mà cần xem thêm chất lượng embedding thông qua bài toán truy xuất ảnh. Trong nghiên cứu này, DermLIP với kiến trúc ViT-Base tương đối gọn nhẹ (~86 triệu tham số, 0,0026 giây/ảnh – nhanh hơn khoảng 9 lần so với MetaCLIP H14) vẫn tạo được không gian embedding có cấu trúc gom cụm bệnh lý tốt hơn nhiều mô hình y khoa lớn hơn. Điều này đặc biệt quan trọng trong các bài toán dữ liệu nhỏ, vì mô hình có embedding tốt ngay từ đầu thường cần ít dữ liệu hơn để đạt hiệu năng mong muốn sau fine-tuning.

Sự chênh lệch hiệu năng trong nhóm mô hình y khoa: từ RCLIP (76,92%), BiomedCLIP (89,38%) đến MedSigLIP (93,15%) – cho thấy các mô hình y khoa lệch miền nên được loại sớm khỏi pipeline da liễu. Những mô hình như RCLIP hay ClipMD không mang lại lợi thế rõ rệt so với mô hình phổ quát; việc chọn chúng chỉ vì nhãn "y tế" có thể khiến quá trình fine-tuning bắt đầu từ một nền tảng kém hơn. Ngược lại, MedSigLIP với kiến trúc lớn và dữ liệu huấn luyện đa chuyên khoa có bao phủ da liễu lại phù hợp hơn cho các bài toán đánh giá offline hoặc chắt lọc tri thức, nơi tốc độ xử lý không phải ưu tiên chính. Cuối cùng, bài toán truy xuất ảnh theo ảnh giúp phát hiện những khác biệt mà cách đánh giá phân loại truyền thống thường bỏ sót. Ví dụ, DermLIP (94,15%) và MetaCLIP (94,08%) có Top-1 gần như tương đương nhưng lại có kiểu nhầm lẫn khác nhau ở từng lớp bệnh. Nếu chỉ nhìn vào độ chính xác tổng thể thì những khác biệt này sẽ không được thấy. Vì vậy, truy xuất ảnh kèm phân tích theo từng lớp bệnh nên được xem là một bước đánh giá quan trọng khi lựa chọn backbone cho các hệ thống AI da liễu, bên cạnh các chỉ số phân loại truyền thống.

E. GIỚI HẠN NGHIÊN CỨU

Nghiên cứu có một số giới hạn cần được thừa nhận.

Đầu tiên, tập dữ liệu chỉ gồm 1.300 ảnh với 100 ảnh mỗi lớp. Quy mô này phù hợp cho đánh giá ban đầu trong thiết lập truy xuất k-NN không tinh chỉnh nhưng chưa đủ lớn để khái quát hóa kết luận cho toàn bộ phổ bệnh da liễu. Phân tích độ nhạy cho thấy thứ hạng giữa DermLIP và MetaCLIP thu hẹp khi tăng số ảnh mỗi lớp từ 50 lên 100, tức là kết quả có thể thay đổi ở quy mô lớn hơn.

Thứ hai, phạm vi nghiên cứu giới hạn ở nhóm bệnh nhiễm trùng – nhóm có khá nhiều đặc điểm hình thái chia sẻ chung như mụn nước, bong vảy và đỏ da lan tỏa. Các nhóm bệnh khác có thể đặt ra thách thức rất khác cho các mô hình embedding và làm thay đổi thứ hạng hiệu năng: ung thư da (melanoma, carcinoma tế bào đáy, carcinoma tế bào vảy) có hệ đặc trưng hình thái riêng (bất đối xứng, bờ không đều, phân bố sắc tố phức tạp); các bệnh tự miễn (lupus ban đỏ, vẩy nến, pemphigus) có biểu hiện đa hình thái. Do đó, kết quả của nghiên cứu này không thể khái quát ngay sang các nhóm bệnh ung thư da hay tự miễn – mỗi nhóm cần được đánh giá riêng với giao thức tương đương trước khi rút ra kết luận về mô hình nền tảng phù hợp. Ưu thế của DermLIP ghi nhận ở đây cũng chưa thể khẳng định sẽ duy trì khi áp dụng cho các nhóm bệnh nói trên, đặc biệt khi DermLIP được huấn luyện chủ yếu trên ảnh lâm sàng đa nhóm bệnh nhưng có thể có tỷ lệ ảnh nhiễm trùng đáng kể trong tập huấn luyện.

Thứ ba, nghiên cứu chỉ dùng ảnh lâm sàng không bao gồm ảnh nội soi da hay giải phẫu bệnh. Lựa chọn này giúp dữ liệu đồng nhất và phản ánh bối cảnh khám sơ bộ ở tuyến cơ sở, nhưng kết quả thu được chưa đủ để nói về năng lực của mô hình trong thực hành da liễu tổng quát.

Thứ tư, dữ liệu được tổng hợp từ nhiều nguồn công khai trên Internet, với nhãn kế thừa trực tiếp từ tập gốc mà chưa qua bước xác nhận lại bởi bác sĩ da liễu độc lập. Ảnh hưởng của giới hạn này hiện rõ ở phân tích nhầm lẫn, đặc biệt ở lớp Infectious – lớp có tính tổng quát cao, được tổng hợp từ nhiều nguồn khác nhau và chỉ đạt Top-1 = 46%, với nhiều ảnh bị nhầm sang Folliculitis. Rà soát thủ công cho thấy một số ảnh trong lớp này có hình thái gần với Folliculitis; sai số do vậy không chỉ đến từ mô hình mà còn liên quan đến chất lượng nhãn. Để định lượng tác động, nghiên cứu thực hiện phân tích loại trừ lớp Infectious khỏi tập đánh giá: Top-1 của mọi mô hình đều tăng khoảng 3,7–4,8 điểm phần trăm – DermLIP tăng từ 94,15% lên 98,17%, MetaCLIP tăng từ 94,08% lên 98,00%. Xu hướng này lặp lại đồng đều trên toàn bộ mô hình mà không xáo trộn thứ hạng. Như vậy chất lượng nhãn là một yếu tố đáng kể chi phối kết quả truy xuất. Một số biện pháp giảm thiểu đã được áp dụng trong nghiên cứu hiện tại: loại ảnh trùng lặp giữa các nguồn, loại ảnh nội soi và giải phẫu bệnh để đảm bảo đồng nhất phương thức, báo cáo kết quả theo từng lớp để người đọc có cái nhìn rõ hơn về độ tin cậy của mô hình. Tuy vậy, những biện pháp này chưa thể thay thế bước xác nhận nhãn bởi chuyên gia da liễu. Các nghiên cứu tiếp theo cần xây dựng quy trình kiểm tra và xác nhận nhãn với sự tham gia của nhiều bác sĩ da liễu độc lập, đồng thời chia nhỏ các lớp tổng quát như Infectious thành những nhóm bệnh cụ thể hơn. Bên cạnh đó, đánh giá giá trị thực tiễn của truy xuất qua reader study với bác sĩ lâm sàng cũng là việc cần triển khai để xem các ảnh truy xuất thực sự hữu ích đến đâu trong hỗ trợ chẩn đoán.

Thứ năm, kiểm tra contamination định lượng bằng perceptual hashing (pHash, Hamming ≤ 10) giữa tập đánh giá 1.300 ảnh và Derm1M [4] cho thấy tỷ lệ contamination thực tế tương đối thấp (4,85%, tương đương 63 ảnh) và không làm thay đổi thứ hạng tương đối giữa 14 mô hình (kết quả chi tiết ở mục IV.E, Bảng 8–9). Tuy nhiên, phạm vi kiểm tra còn một số giới hạn. Phép đối chiếu pHash chỉ chạy trên subset của Derm1M liên quan tới phổ bệnh nhiễm trùng, chưa bao quát toàn bộ tập huấn luyện gốc, cũng như tập tiền huấn luyện của các mô hình khác như MetaCLIP H14 (FullCC2.5B). Ngoài ra, ngưỡng Hamming ≤ 10 có thể bỏ sót những ảnh đã trải qua biến đổi mạnh như xoay góc lớn, đổi tỷ lệ khung hình hay tái mã hóa nhiều lần. Một tập đánh giá hold-out hoàn toàn độc lập, lấy từ những nguồn không nằm trong tập huấn luyện của bất kỳ mô hình nào được đánh giá (ví dụ ảnh thu trực tiếp tại bệnh viện hợp tác, không công khai trên Internet) vẫn là bước bổ sung cần thiết cho các nghiên cứu mở rộng nếu muốn loại trừ tuyệt đối ảnh hưởng của contamination đến kết luận về hiệu năng tương đối giữa các nhóm mô hình.

Thứ sáu, ngoài lớp Infectious đã được phân tích chi tiết bằng phân tích loại trừ lớp, nghiên cứu chưa thực hiện được phân tích định lượng về tính nhất quán nhãn liên nguồn cho các lớp đa nguồn còn lại, gồm Herpes (2 nguồn), Monkeypox (2 nguồn) và Sarampion (2 nguồn). Ba phân tích chuyên sâu sau đây cần được thực hiện nhưng chưa kịp triển khai do giới hạn thời gian và nguồn lực: (i) kiểm tra trùng lặp ảnh giữa các nguồn bằng perceptual hashing, (ii) báo cáo độ chính xác truy xuất theo từng lớp cho từng nguồn trong cùng một lớp, (iii) phân tích hướng nhầm lẫn liên nguồn (ảnh từ nguồn A bị truy xuất sang ảnh cùng lớp từ nguồn B hay sang lớp khác). Tuy nhiên, các bằng chứng gián tiếp từ kết quả theo từng lớp ở Bảng 7 cho thấy vấn đề nhãn không đồng nhất liên nguồn có lẽ không nghiêm trọng ở các lớp đa nguồn ngoài Infectious: Herpes đạt 100% Top-1 trên cả 80 ảnh từ Skin-Disease-detection lẫn 20 ảnh từ skin disease, Sarampion đạt 100% trên cả 66 ảnh từ Roboflow lẫn 34 ảnh từ MSID, Monkeypox đạt 86–90% trên cả 87 ảnh Roboflow lẫn 13 ảnh MSID. Nếu nhãn giữa các nguồn không nhất quán nghiêm trọng, độ chính xác tổng hợp khó có thể đạt mức 100% hoặc gần 100% như đã quan sát. Dù vậy, đây cũng không phải bằng chứng đầy đủ vì nó chỉ phản ánh tính nhất quán nội bộ của từng nguồn, không khẳng định được rằng định nghĩa nhãn giữa các nguồn là tương đương. Trong các nghiên cứu mở rộng, ba phân tích nêu trên cùng quá trình xác nhận nhãn độc lập bởi bác sĩ da liễu là cần thiết để loại trừ tác động của nhiễu nhãn liên nguồn lên kết quả đánh giá. Cuối cùng, toàn bộ đánh giá được thực hiện theo phương thức k-NN retrieval không tinh chỉnh – các mô hình không được tinh chỉnh trên tập dữ liệu đích. Cách làm này đúng với kịch bản triển khai nhanh, chi phí thấp, nhưng chưa cho biết thứ hạng có thay đổi sau khi tinh chỉnh hay không. Bước tiếp theo của nhóm nghiên cứu sẽ là tinh chỉnh các mô hình tiềm năng nhất trên dữ liệu đích và đánh giá lại hiệu năng sau huấn luyện.

VI. KẾT LUẬN

Bài báo xây dựng khung đánh giá thống nhất so sánh 14 mô hình embedding thuộc ba nhóm (phổ quát, y khoa tổng quát, chuyên biệt da liễu) trên cùng giao thức truy xuất ảnh theo ảnh với 1.300 ảnh lâm sàng của 13 bệnh nhiễm trùng (leave-one-out), kết hợp ba kiểm định thống kê và phân tích contamination với Derm1M. Lựa chọn truy xuất ảnh theo ảnh ở đây phản ánh trực tiếp cấu trúc gom cụm trong không gian embedding, sát với cách các pipeline thực tế khai thác mô hình nền tảng.

Phát hiện đáng chú ý nhất là miền dữ liệu huấn luyện, chứ không phải quy mô kiến trúc, mới chính là yếu tố quyết định chất lượng embedding cho bài toán này. DermLIP (ViT-B/16, ~86 triệu tham số) đạt Top-1 = 94,15%, gần như ngang ngửa MetaCLIP H14 (~1 tỷ tham số, 94,08%) dù nhỏ hơn khoảng 10 lần về quy mô. Hiệu số chỉ +0,08 điểm phần trăm với bootstrap 95% CI [−0,77%, +0,92%], cả ba kiểm định đều cho p = 1,0; thêm vào đó, mức độ ổn định ở từng lớp bệnh (hai mô hình ngang nhau trên 9/13 lớp) càng củng cố nhận định này. Nói cách khác, tri thức miền chuyên biệt thu được từ huấn luyện trực tiếp trên ảnh da liễu có thể bù lại được khoảng cách rất lớn về kiến trúc, dù để khẳng định tính tương đương một cách chính thức vẫn cần các kiểm định equivalence chuyên biệt ở những nghiên cứu sau. Xu hướng này còn rõ hơn trong nhóm mô hình y khoa: hiệu năng tăng dần theo mức độ xuất hiện của ảnh da lâm sàng trong tập huấn luyện. Cụ thể, RCLIP (huấn luyện X-quang/CT) chỉ đạt 76,92%, rồi đến ClipMD 87,92%, BiomedCLIP 89,38%, PLIP 90,92%, và cao nhất là MedSigLIP có bao phủ da liễu với 93,15%. Khoảng cách 16 điểm phần trăm giữa đầu trên và đầu dưới của xu hướng này cho thấy yếu tố quyết định không phải là nhãn "y tế" gắn lên mô hình, mà là tính tương đồng về phương thức ảnh và phổ bệnh giữa dữ liệu huấn luyện và dữ liệu đích. Một quan sát thú vị khác là PLIP, dù được phát triển cho ảnh giải phẫu bệnh, vẫn vượt các mô hình y khoa lệch miền khác và còn cao hơn cả BiomedCLIP, gợi ý rằng những đặc trưng hình thái học từ ảnh mô học, đặc biệt là kết cấu và phân bố tổn thương, có khả năng chuyển giao phần nào sang ảnh da liễu lâm sàng. Phân tích contamination với tập Derm1M cũng cho thấy mức trùng lặp thực tế khoảng 4,85% (63 ảnh trên 1.300) và không xáo trộn thứ hạng giữa các mô hình, củng cố thêm độ tin cậy cho các kết luận trên.

Ngoài những phát hiện về mô hình, kết quả thực nghiệm còn xác nhận cosine là độ đo phù hợp hơn Euclid cho truy xuất đa phương thức. Khi embedding đã chuẩn hóa L2, hai độ đo cho kết quả gần như trùng nhau ở truy xuất ảnh theo ảnh; nhưng ở truy xuất ảnh bằng văn bản, Top-1 sụt từ 28% đến 75% khi đổi sang Euclid do khoảng cách phương thức trong các mô hình CLIP, vấn đề mà cosine xử lý được nhờ chỉ đo góc giữa các véc-tơ.

Trên cơ sở đánh giá đồng thời ba khía cạnh là hiệu năng truy xuất, chi phí tính toán và cấu trúc gom cụm theo lớp bệnh, bài báo đề xuất bốn khuyến nghị thực tiễn cho pipeline fine-tuning ảnh da liễu nhiễm trùng. Trước hết, DermLIP là điểm khởi đầu mặc định hợp lý nhờ cân bằng tốt giữa hiệu năng cao (94,15% Top-1, 8/13 lớp đạt 100% Top-1) và chi phí thấp (~86 triệu tham số, 0,0026 giây/ảnh, nhanh hơn MetaCLIP H14 khoảng 9 lần). Tiếp theo, các mô hình y khoa lệch miền như RCLIP và ClipMD nên được loại sớm khỏi danh sách ứng viên để tránh khởi đầu fine-tuning từ một nền tảng yếu. MetaCLIP H14 vẫn là lựa chọn tốt khi ứng dụng đòi hỏi khả năng khái quát đa miền, dù phải chấp nhận chi phí tính toán cao gấp gần 10 lần. Cuối cùng, MedSigLIP là một phương án trung gian phù hợp khi ứng dụng cần bao phủ thêm những miền y tế khác ngoài da liễu.

Tóm lại, bài báo có bốn đóng góp chính. Thứ nhất là một khung đánh giá thống nhất để so sánh ba nhóm mô hình embedding trên cùng giao thức truy xuất ảnh theo ảnh; khung này có thể tái sử dụng cho các bài toán da liễu khác hoặc những miền y tế có đặc thù tương tự. Thứ hai là bằng chứng thực nghiệm về vai trò quyết định của miền dữ liệu huấn luyện đối với chất lượng embedding, đặt trong so sánh trực tiếp với ảnh hưởng của quy mô kiến trúc; đây là phát hiện thiết thực với những nhóm nghiên cứu có tài nguyên hạn chế khi phải lựa chọn giữa đầu tư vào mô hình lớn và đầu tư vào dữ liệu chuyên biệt. Thứ ba là một quy trình phân tích contamination định lượng bằng perceptual hashing, kèm theo phân loại ba mức và phân tích độ nhạy sau khi loại trừ, đã xác nhận tỷ lệ contamination thực tế thấp và không ảnh hưởng đến thứ hạng giữa các mô hình. Thứ tư là cơ sở định lượng cho việc chọn backbone trong pipeline fine-tuning, với DermLIP được xác định là điểm khởi đầu phù hợp cho hệ thống truy xuất ảnh da liễu nhiễm trùng.

Tất nhiên, các kết luận hiện tại mới được xác lập trên tập 1.300 ảnh lâm sàng bề mặt thuộc nhóm bệnh nhiễm trùng, nên việc kiểm chứng khả năng tổng quát hóa của chúng cần thêm những nghiên cứu mở rộng. Hướng đầu tiên là mở rộng đánh giá sang các nhóm bệnh khác như ung thư da và bệnh tự miễn, vốn có đặc trưng hình thái khác hẳn nhóm nhiễm trùng, để xem thứ hạng giữa các mô hình có ổn định khi đổi miền bệnh lý hay không. Mở rộng quy mô dữ liệu cũng là một bước cần thiết để củng cố độ tin cậy thống kê. Bên cạnh đánh giá embedding không tinh chỉnh, các bước tiếp theo nên tinh chỉnh những mô hình tiềm năng trên dữ liệu đích và đánh giá lại trên các tác vụ hạ nguồn như phân loại, phân đoạn và truy xuất, để xem ưu thế ở giai đoạn không tinh chỉnh có còn được duy trì sau huấn luyện chuyên biệt hay không. Một hướng nữa là bổ sung ảnh nội soi da và ảnh giải phẫu bệnh nhằm xây dựng thiết lập đánh giá đa phương thức, gần hơn với thực hành lâm sàng. Cuối cùng, các đánh giá có sự tham gia trực tiếp của bác sĩ da liễu là cần thiết để xác nhận giá trị ứng dụng thực tế của hệ thống truy xuất, đặc biệt trong các tình huống hỗ trợ chẩn đoán tại tuyến cơ sở hay y tế từ xa.

VII. LỜI CẢM ƠN

Bài báo này được hoàn thành trong khuôn khổ đề tài nghiên cứu khoa học cấp trường mã số HSV2025-27 với tên đề tài "Xây dựng ứng dụng AI đa phương thức chẩn đoán bệnh da liễu nhiễm trùng", do Trường Đại học Ngoại ngữ – Tin học Thành phố Hồ Chí Minh (HUFLIT) tài trợ, dưới sự hướng dẫn khoa học của ThS. Bùi Thị Thanh Tú (Khoa Công nghệ thông tin, HUFLIT). Nhóm tác giả trân trọng cảm ơn sự định hướng tận tình của giảng viên hướng dẫn

cùng sự hỗ trợ từ Nhà trường trong suốt quá trình thực hiện đề tài. Nhóm tác giả cũng xin chân thành cảm ơn Ban biên tập Tạp chí Khoa học Trường Đại học Ngoại ngữ – Tin học Thành phố Hồ Chí Minh cùng các nhà phản biện ẩn danh đã có những góp ý sâu sắc, mang tính xây dựng, góp phần quan trọng nâng cao chất lượng bài báo.

VIII. TÀI LIỆU THAM KHẢO

[1] A. Mehta et al. (2025). Deep learning image processing models in dermatopathology, Diagnostics, Vol. 15, No. 19, p. 2517. DOI: https://doi.org/10.3390/diagnostics15192517.

[2] R. Daneshjou et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set, Science Advances, Vol. 8, No. 32, p. eabq6147. DOI: https://doi.org/10.1126/sciadv.abq6147.

[3] D. Wen, A. Soltan, E. Trucco, and R. N. Matin (2024). From data to diagnosis: Skin cancer image datasets for artificial intelligence, Clinical and Experimental Dermatology, Vol. 49, No. 7, pp. 675-685. DOI: https://doi.org/10.1093/ced/llae112.

[4] S. Yan, M. Hu, Y. Jiang, et al. (2025). Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12681-12690.

[5] F. Khun Jush, T. Truong, S. Vogler, and M. Lenga (2024). Medical image retrieval using pretrained embeddings, Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1-5. DOI: https://doi.org/10.1109/ISBI56570.2024.10635170.

[6] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio (2019). Transfusion: Understanding transfer learning for medical imaging, Advances in Neural Information Processing Systems (NeurIPS), Vol. 32, pp. 3347-3357.

[7] S. Azizi, B. Mustafa, F. Ryan, et al. (2021). Big self-supervised models advance medical image classification, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3478-3488.

[8] A. Radford, J. W. Kim, C. Hallacy, et al. (2021). Learning transferable visual models from natural language supervision, Proceedings of the 38th International Conference on Machine Learning (ICML), Vol. 139, pp. 8748-8763.

[9] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung (2021). GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3942-3951. DOI: https://doi.org/10.1109/ICCV48922.2021.00391.

[10] S. Zhang, Y. Xu, N. Usuyama, et al. (2024). BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, NEJM AI, Vol. 2, No. 1. DOI: https://doi.org/10.1056/AIoa2400640.

[11] Z. Zhao, Y. Liu, H. Wu, et al. (2025). CLIP in medical imaging: A survey, Medical Image Analysis, Vol. 102, p. 103551. DOI: https://doi.org/10.1016/j.media.2025.103551.

[12] Y. Yang, M. Gandhi, Y. Wang, et al. (2024). A textbook remedy for domain shifts: Knowledge priors for medical image analysis, Advances in Neural Information Processing Systems (NeurIPS).

[13] S. Yan, Z. Yu, C. Primiero, et al. (2025). A multimodal vision foundation model for clinical dermatology, Nature Medicine, Vol. 31, No. 8, pp. 2691-2702. DOI: https://doi.org/10.1038/s41591-025-03747-y.

[14] J. Zhou, X. He, L. Sun, et al. (2024). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4, Nature Communications, Vol. 15, p. 5649. DOI: https://doi.org/10.1038/s41467-024-50043-3.

[15] P. Tschandl, G. Argenziano, M. Razmara, and J. Yap (2019). Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features, British Journal of Dermatology, Vol. 181, No. 1, pp. 155-165. DOI: https://doi.org/10.1111/bjd.17189.

[16] M. Gassner, J. B. Garcia, S. Tanadini-Lang, et al. (2023). Saliency-enhanced content-based image retrieval for diagnosis support in dermatology consultation: Reader study, JMIR Dermatology, Vol. 6, p. e42129. DOI: https://doi.org/10.2196/42129.

[17] W. Barhoumi and A. Khelifa (2021). Skin lesion image retrieval using transfer learning-based approach for query-driven distance recommendation, Computers in Biology and Medicine, Vol. 137, p. 104825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104825.

[18] K. Huang, K. Sun, J. Li, et al. (2025). Intelligent strategy for severity scoring of skin diseases based on clinical decision-making thinking with lesion-aware transformer, Artificial Intelligence Review, Vol. 58, p. 95. DOI: https://doi.org/10.1007/s10462-024-11083-9.

[19] M. Rashad, I. Afifi, and M. Abdelfatah (2023). RbQE: An efficient method for content-based medical image retrieval based on query expansion, Journal of Digital Imaging, Vol. 36, pp. 1248-1261. DOI: https://doi.org/10.1007/s10278-023-00782-9.

[20] Y. Liu, A. Jain, C. Eng, et al. (2020). A deep learning system for differential diagnosis of skin diseases, Nature Medicine, Vol. 26, No. 6, pp. 900-908. DOI: https://doi.org/10.1038/s41591-020-0842-3.

[21] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt (2022). Transfer learning for medical image classification: A literature review, BMC Medical Imaging, Vol. 22, No. 1, p. 69. DOI: https://doi.org/10.1186/s12880-022-00793-7.
[22] S. Biswas (2023). Skin Disease Dataset, https://www.kaggle.com/datasets/subirbiswas19/skin-disease-dataset, Accessed date: 20/05/2025.
[23] B. Jyothula (2023). Dermatology Hair Dataset, https://www.kaggle.com/datasets/bhaskarjyothula/dermatology-hair, Accessed date: 20/05/2025.
[24] Ravi (2022). Human Monkeypox Dataset, https://universe.roboflow.com/ravi-zsd7r/human-monkeypox, Accessed date: 20/05/2025.
[25] F. Gk (2023). Skin Disease Dataset, https://universe.roboflow.com/fredrick-gk-wpenz/skin-disease-hwlrn, Accessed date: 20/05/2025.
[26] I. Bos (2023). Skin Disease Detection Dataset, https://universe.roboflow.com/icik-bos/skin-disease-detection-inlte, Accessed date: 20/05/2025.
[27] N. H. Ali et al. (2022). Monkeypox Skin Images Dataset (MSID), Mendeley Data, Vol. 3. DOI: https://doi.org/10.17632/r9bfpnvyxr.3.
[28] Yolo (2023). Skin Dataset, https://universe.roboflow.com/yolo-reobj/skin-ua2dl, Accessed date: 20/05/2025.
[29] Skin (2023). Skin Dataset, https://universe.roboflow.com/skin-vmibf/skin-1qn4y, Accessed date: 20/05/2025.
[30] Gaurav (2023). Candida Intertrigo Dataset, https://universe.roboflow.com/gaurav-6cgzc/candida-intertrigo-hj2jx, Accessed date: 20/05/2025.
[31] H. Xu, S. Xie, X. E. Tan, et al. (2024). Demystifying CLIP data, Proceedings of the International Conference on Learning Representations (ICLR).
[32] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer (2023). Sigmoid loss for language image pre-training, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11975-11986.
[33] Z. Chen, G. Liu, B.-W. Zhang, et al. (2023). AltCLIP: Altering the language encoder in CLIP for extended language capabilities, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8666-8682. DOI: https://doi.org/10.48550/arXiv.2211.06679.
[34] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie (2022). A ConvNet for the 2020s, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976-11986.
[35] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev (2023). Reproducible scaling laws for contrastive language-image learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2829.
[36] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel (2024). MobileCLIP: Fast image-text models through multi-modal reinforced training, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15963-15974.
[37] A. Sellergren, S. Kazemzadeh, T. Jaroensri, et al. (2025). MedGemma technical report, arXiv:2507.05201.
[38] Z. Huang, F. Bianchi, M. Yuksekgonul, et al. (2023). A visual-language foundation model for pathology image analysis using medical Twitter, Nature Medicine, Vol. 29, No. 9, pp. 2307-2316. DOI: https://doi.org/10.1038/s41591-023-02504-3.
[39] I. Glassberg and T. Hope (2023). Increasing textual context size boosts medical image-text matching, arXiv:2303.13340.
[40] K. Shahhosseini (2023). RCLIP: CLIP model fine-tuned on radiology images and their captions, https://huggingface.co/kaveh/rclip, Accessed date: 20/05/2025.
[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton (2020). A simple framework for contrastive learning of visual representations, Proceedings of the 37th International Conference on Machine Learning (ICML), Vol. 119, pp. 1597-1607.
[42] C. M. Bishop (2006). Pattern Recognition and Machine Learning, Springer, New York, NY, USA.
[43] G. Qian, S. Sural, Y. Gu, and S. Pramanik (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries, Proceedings of the 2004 ACM Symposium on Applied Computing (SAC), pp. 1232-1237. DOI: https://doi.org/10.1145/967900.968151.
[44] H. Steck, C. Ekanadham, and N. Kallus (2024). Is cosine-similarity of embeddings really about similarity?, Companion Proceedings of the ACM Web Conference (WWW), pp. 887-890. DOI: https://doi.org/10.1145/3589335.3651526.
[45] P. Tschandl, C. Rosendahl, and H. Kittler (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific Data, Vol. 5, p. 180161. DOI: https://doi.org/10.1038/sdata.2018.161.

EVALUATION OF EMBEDDING SPACES OF FOUNDATION MODELS FOR INFECTIOUS DERMATOLOGY IMAGES TO GUIDE BACKBONE SELECTION FOR FINE-TUNING PIPELINES

Pham Thi Anh Thu, Nguyen Gia Khoi, Tran Trung Hieu, Tran Anh Duc, Bui Thi Thanh Tu

ABSTRACT— In the context of limited dermatological labels, a key question is which embedding model should serve as the backbone for image retrieval and analysis applications. To answer this, we compare 14 models across three groups (universal, general medical, and dermatology-specialized) on an image-to-image retrieval task with leave-one-out protocol, using 1,300 clinical surface skin images of 13 infectious diseases. The evaluation combines statistical testing (bootstrap 95% CI, McNemar's exact test, paired permutation test) with class-wise confusion analysis. Key findings: (i) DermLIP (ViT-B/16, ~86M parameters) achieves Top-1 = 94.15%, statistically indistinguishable from MetaCLIP H14 (~1B parameters, 94.08%): difference +0.08 pp, bootstrap 95% CI [−0.77%, +0.92%], McNemar p = 1.0, permutation p = 1.0; (ii) domain-specific knowledge can compensate for not scaling up architecture (~10× parameter gap); (iii) general medical models trained on domain-shifted data perform worse than universal models, indicating that the "medical" label by itself confers no advantage when the training-domain modality and disease coverage do not match the target domain; (iv) a pHash-based contamination analysis with Derm1M shows only ~4.85% overlap (63/1,300 images), which does not affect the relative ranking of models. These findings provide quantitative grounds for backbone selection in dermatology image fine-tuning pipelines.

Keywords— Dermatology image embedding, embedding space evaluation, foundation model for infectious skin disease, cosine similarity in image retrieval, backbone selection for fine-tuning pipeline.

Phạm Thị Anh Thư là sinh viên Khoa Công nghệ thông tin, Trường đại học Ngoại ngữ – Tin học TP.HCM (HUFLIT), hiện là Data Engineer tại FintechAI và chủ nhiệm đề tài nghiên cứu ứng dụng AI trong y tế. Cô có kinh nghiệm phát triển các dự án AI, Blockchain và Fintech, tập trung vào giải pháp công nghệ cho lĩnh vực tài chính và y khoa.

Nguyễn Gia Khôi Cử nhân Công nghệ Thông tin, Trường Đại học Ngoại ngữ – Tin học TP.HCM (HUFLIT) năm 2026. Hiện anh đang đảm nhiệm vị trí Backend Engineer tại Công ty TNHH Thương mại dịch vụ Phương Quân và tham gia nghiên cứu, phát triển các ứng dụng AI trong lĩnh vực phần mềm.

Trần Trung Hiếu là sinh viên ngành Công nghệ Thông tin, Trường Đại học Ngoại ngữ – Tin học TP.HCM (HUFLIT), hiện đang đảm nhiệm vị trí Software Engineer tại TGL Solutions và tham gia nghiên cứu, phát triển các dự án công nghệ trong lĩnh vực phần mềm và trí tuệ nhân tạo.

Trần Anh Đức là sinh viên Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ – Tin học Thành phố Hồ Chí Minh (HUFLIT).

Bùi Thị Thanh Tú là Thạc sĩ Công nghệ Thông tin tại Học viện tin học Pháp ngữ (IFI) – Đại học Bách Khoa Hà Nội năm 2003. Hiện cô là giảng viên Khoa Công nghệ Thông tin, Trường Đại học Ngoại ngữ – Tin học TP.HCM (HUFLIT). Cô có hơn 15 năm kinh nghiệm trong lĩnh vực công nghệ phần mềm và nghiên cứu ứng dụng AI trong y tế, giáo dục.