

Hiệu quả ứng dụng trí tuệ nhân tạo trong hỗ trợ chẩn đoán X-quang ngực

Effectiveness of artificial intelligence in chest radiograph interpretation

Bùi Trọng Dương^{1*}, Lâm Khánh²,
Phùng Anh Tuấn¹ và Vũ Đình Triển²

¹Học viện Quân Y,
²Bệnh viện Trung ương Quân đội 108

Tóm tắt

Mục tiêu: Đánh giá hiệu quả ứng dụng phần mềm trí tuệ nhân tạo DrAid trong phân tích X-quang ngực thẳng. **Đối tượng và phương pháp:** Nghiên cứu mô tả cắt ngang trên 637 phim chụp X-quang ngực thẳng tại Bệnh viện Trung ương Quân đội 108 từ tháng 10/2024 đến tháng 03/2025. Đánh giá kết quả phân tích của AI DrAid về độ nhạy, độ đặc hiệu và diện tích dưới đường cong ROC (AUC), có so sánh với kết quả của bác sĩ chẩn đoán hình ảnh và tiêu chuẩn đối chiếu dựa trên cắt lớp vi tính. **Kết quả:** Với các loại tổn thương khác nhau, mô hình AI đạt độ nhạy dao động từ 64,58% đến 85,91%, độ đặc hiệu từ 83% đến 98,52%, và AUC từ 0,792 đến 0,885. AI có độ nhạy, độ đặc hiệu và AUC cao hơn so bác sĩ chẩn đoán hình ảnh cả khi đọc độc lập hoặc có sự hỗ trợ của AI đối với hầu hết trong số các tổn thương được khảo sát. Việc sử dụng AI giúp tiết kiệm trung bình 17,03 giây (tương đương 21%) thời gian phân tích phim. **Kết luận:** AI là công cụ hiệu quả trong việc hỗ trợ nâng cao chất lượng và hiệu suất chẩn đoán X-quang ngực thẳng. Tuy nhiên, cần hiểu rõ về đặc điểm, năng lực của công cụ AI để đạt được hiệu quả tốt nhất trong quá trình sử dụng.

Từ khóa: Trí tuệ nhân tạo, X-quang ngực thẳng.

Summary

Objective: To evaluate the effectiveness of the DrAid artificial intelligence (AI) software in interpreting posteroanterior chest radiographs. **Subjects and Methods:** A cross-sectional descriptive study was conducted on 637 posteroanterior chest X-rays performed at 108 Military Central Hospital from October 2024 to March 2025. The sensitivity, specificity, and area under the curve (AUC) of the DrAid AI system were assessed and compared with interpretations by radiologists, using computed tomography (CT) findings as the reference standard. **Results:** Across different lesion types, the AI model achieved a sensitivity ranging from 64.58% to 85.91%, a specificity between 83% and 98.52%, and an area under the ROC curve (AUC) from 0.792 to 0.885. AI model demonstrated higher sensitivity, specificity, and AUC than radiologists in detecting most types of pulmonary abnormalities assessed in the study, both in independent readings and when AI assistance was provided. The use of AI reduced the average image interpretation time by 17.03 seconds, equivalent to a 21% improvement in reading efficiency. **Conclusion:** Artificial intelligence proves to be an effective tool for enhancing the quality and efficiency of chest X-ray interpretation. However, optimal effectiveness requires a clear understanding of the AI system's capabilities and limitations.

Keywords: Artificial intelligence, chest xray, chest radiograph.

Ngày nhận bài: 5/6/2025, ngày chấp nhận đăng: 7/7/2025

* Tác giả liên hệ: trongduongcdha@gmail.com - Học viện Quân Y

I. ĐẶT VẤN ĐỀ

Cho tới hiện tại, X-quang ngực vẫn đang là phương tiện chẩn đoán hình ảnh đầu tay trong tầm soát, chẩn đoán các bệnh lý vùng lồng ngực, đặc biệt là các bệnh lý mang tính toàn cầu như ung thư phổi, lao, Covid-19... Tuy nhiên, phương pháp này trong nhiều trường hợp còn nhiều hạn chế so với chụp cắt lớp vi tính (CT), chụp phát xạ positron...¹. Với sự phát triển của khoa học công nghệ hiện đại, trong những năm trở lại đây, trí tuệ nhân tạo (Artificial intelligence-AI) đang là hướng đi được quan tâm nhiều nhất nhằm nâng cao hiệu suất và chất lượng của báo cáo kết quả phân tích X-quang ngực.

Rất nhiều nghiên cứu trên thế giới đã chỉ ra rằng AI cung cấp sự hỗ trợ hiệu quả trong việc nâng cao độ nhạy, độ đặc hiệu của chẩn đoán, đồng thời giảm thời gian đọc phim giúp nâng cao hiệu suất làm việc và giảm tình trạng quá tải, mệt mỏi cho bác sĩ chẩn đoán hình ảnh [2, 3]. Tuy nhiên, hiệu quả của AI trong chẩn đoán X-quang ngực thẳng cho tới nay vẫn chưa thực sự đi được đến kết luận cuối cùng, một phần là do tính “khó giải thích” trong các gợi ý của AI, phần khác là do sự chưa thống nhất giữa các nghiên cứu được tiến hành. Trong khi nhiều nghiên cứu cho thấy AI phát huy tốt vai trò của mình trong việc phát hiện các loại tổn thương khác nhau [3], một số khác lại đánh giá AI không làm tốt bằng các bác sĩ chẩn đoán hình ảnh trong điều kiện thực tế lâm sàng [4-6]. Ngoài ra, các nghiên cứu còn chỉ ra rằng đối với một số loại tổn thương cụ thể như tràn dịch màng phổi, các tổn thương mạch máu, trung thất và xương thì AI tỏ ra kém hiệu quả hơn so với các tác vụ khác cũng như so với bác sĩ chẩn đoán hình ảnh^{7,8}. Ngoài ra, các nghiên cứu đã được tiến hành đa phần đều lấy tiêu chuẩn đối chứng (ground truth) là đồng thuận của hai hay nhiều chuyên gia, bác sĩ chẩn đoán hình ảnh khác nhau. Tuy nhiên, do hạn chế về bản chất hình ảnh của X-quang và giới hạn thị lực của con người nên sự đồng thuận của các chuyên gia, bác sĩ chẩn đoán hình ảnh cũng không mang lại được sự chính xác hoàn toàn.

Vì lý do đó, nghiên cứu này được tiến hành để đánh giá một cách khách quan về tính chính xác và hiệu quả hỗ trợ của AI trong phân tích một số tổn

thương cơ bản thường gặp trên phim chụp X-quang ngực thẳng

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

2.1. Đối tượng

637 phim chụp X-quang ngực thẳng được chụp tại Bệnh viện Trung ương Quân đội 108 từ tháng 10/2024 đến tháng 03/2025 kèm theo phim chụp cắt lớp vi tính tương ứng dưới định dạng DICOM

2.2. Phương pháp

Thiết kế nghiên cứu: mô tả cắt ngang

Cỡ mẫu và phương pháp chọn mẫu: lấy mẫu thuận tiện với cỡ mẫu tối thiểu được tính theo công thức:

$$N_{se/sp} \geq \frac{Z_a^2 \times p_{se/sp} \times (1 - p_{se/sp})}{w^2 \times Pre}$$

Với $Z_a^2 = 1.96$, sai số mong muốn tối đa $w = 5\%$, độ nhạy và độ đặc hiệu mong muốn là 0.8, tỉ lệ lưu hành tổn thương tham khảo trong kết quả nghiên cứu tương tự là $Pre = 48,2\%^2$, từ đó ta tính được cỡ mẫu tối thiểu 511 phim chụp.

Phương tiện sử dụng: AI được sử dụng trong nghiên cứu là phần mềm DrAid do VinBrain tiến hành đã được FDA chấp thuận, đã thương mại hóa tại một số cơ sở y tế tại Việt Nam. Dữ liệu X-quang ngực thẳng được chụp trên máy Brivo DR-F của GE Medical Systems sản xuất năm 2021.



Hình 1. Các tổn thương được AI xác định và đánh dấu vị trí

Phương pháp tiến hành

Dữ liệu phim chụp X-quang ngực thẳng được đưa vào phần mềm DrAid dưới định dạng chuẩn DICOM. Các phim chụp được phân tích độc lập bởi AI và 6 bác sĩ chẩn đoán hình ảnh với mức độ kinh

nghiệm khác nhau (4 bác sĩ mới dưới 5 năm kinh nghiệm và 2 bác sĩ trên 5 năm kinh nghiệm). Các phim chụp được đọc lại bởi cùng bác sĩ với sự hỗ trợ của AI sau đó 1 tháng để tránh nhiễu do trí nhớ tạm thời. Thời gian phân tích phim được đo bằng plug-in liên kết trực tiếp với phần mềm, tính từ thời điểm mở phim cho tới lúc lập báo cáo, tính theo đơn vị giây.

Tiêu chí đánh giá là có hay không có các dạng tổn thương sau trên phim chụp: Dạng nốt/khối, đồng đặc, xẹp phổi, tràn dịch màng phổi, tràn khí

màng phổi và gãy xương. Tiêu chuẩn đối chiếu được xây dựng dựa trên CT phổi. Một bác sĩ chẩn đoán hình ảnh có kinh nghiệm (không nằm trong nhóm 6 bác sĩ đọc phim nghiên cứu) đối chiếu các tổn thương trên CT và vị trí tương ứng trên X-quang để đưa ra kết luận có thể hay không thể phát hiện tổn thương đó trên X-quang. Đôi khi do tư thế bệnh nhân hoặc do các tổn thương kích thước nhỏ, bị che lấp... nên dù xác định được trên CT nhưng không thể phát hiện được trên X-quang (Hình 2).



Hình 2. Tràn dịch màng phổi phải trên CT nhưng không rõ ràng trên X-quang do bệnh nhân chụp tư thế nằm

2.3. Xử lý số liệu

Số liệu được xử lý bằng phần mềm SPSS 22.0 và Medcalc 23.2.1. Các kết quả phân tích của AI, bác sĩ và bác sĩ với hỗ trợ của AI được đánh giá độ nhạy, độ đặc hiệu và diện tích dưới đường cong ROC (AUC). Độ tương đồng giữa các mô hình chẩn đoán được đánh giá bằng kiểm định Mc Nemar và Cochran's Q test cho độ nhạy, độ đặc hiệu và Delong method cho AUC. Thời gian phân tích phim của các phương

thức được so sánh bằng kiểm định Wilcoxon-sign rank test.

2.4. Đạo đức nghiên cứu

Nghiên cứu được tiến hành hoàn toàn độc lập với quá trình chẩn đoán và điều trị lâm sàng, các kết quả phân tích phim được sử dụng giới hạn trong phạm vi nghiên cứu và không được áp dụng cho quá trình thăm khám và điều trị thực tế của bệnh nhân.

III. KẾT QUẢ

3.1. Đặc điểm chung của đối tượng nghiên cứu

Bảng 1. Đặc điểm chung của đối tượng nghiên cứu

Đặc điểm		Số lượng (n = 637)	Tỉ lệ %
Giới tính	Nam	439	68,9
	Nữ	198	31,1
Tuổi	Trung bình	60.49 ± 15.6	

Tuổi trung bình của bệnh nhân là 60,49 ± 15,6 tuổi, với các bệnh nhân chủ yếu là nam (68,9%).

3.2. Tính chính xác và hiệu quả hỗ trợ phát hiện tổn thương của AI

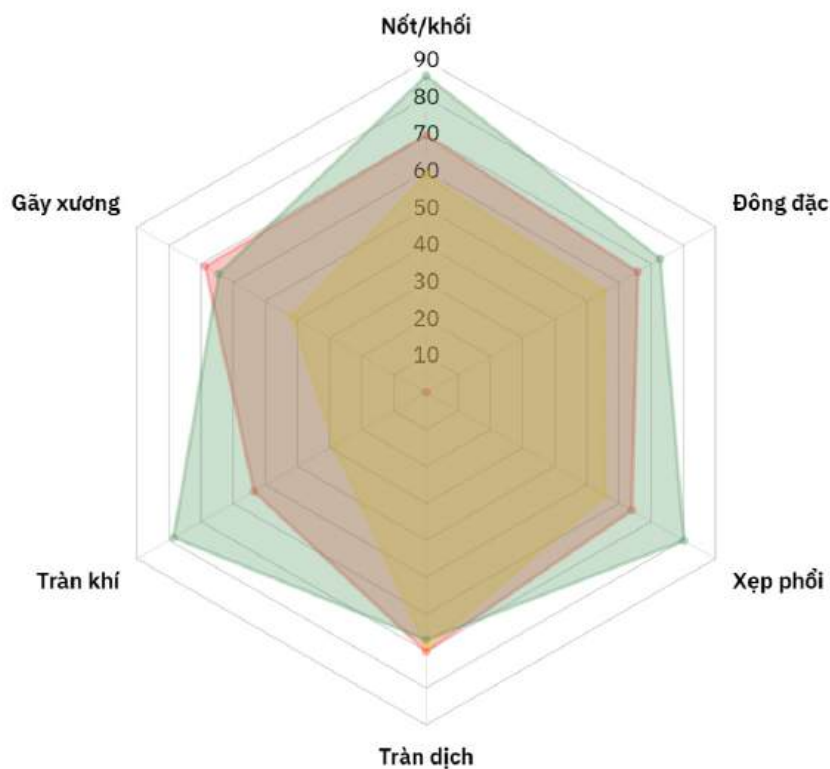
Bảng 2. Độ nhạy của các phương pháp đọc X-quang (Se-95%CI)

Tổn thương	Bác sĩ	AI	Bác sĩ + AI	p
Dạng nốt/khối	59,06* (55,24-62,88)	85,91* (83,21-88,61)	69,80* (66,23-73,37)	0,000
Đông đặc phổi	55,26* (51,40-59,12)	72,81 (69,35-76,27)	65,79 (62,11-69,47)	0,002
Xẹp phổi	55,81 (51,95-59,67)	80,23* (77,14-83,32)	63,95 (60,22-67,68)	0,000
Tràn dịch màng phổi	69,15 (65,56-72,74)	67,02 (63,37-70,67)	70,21 (66,66-73,76)	0,814
Tràn khí màng phổi	28,57* (25,06-32,08)	78,57 (75,38-81,76)	53,57 (49,70-57,44)	0,000
Gãy xương	41,67* (37,84-45,50)	64,58 (60,87-68,29)	68,75 (65,15-72,35)	0,004

P-value được tính chung cho sự khác biệt về độ nhạy của các phương pháp chẩn đoán (Cochran's Q test).

** - Độ nhạy của mô hình có sự khác biệt với các mô hình còn lại (Mc Nemar test với $p < 0,05$).*

Độ nhạy của AI cao hơn so với bác sĩ đọc độc lập và bác sĩ đọc với hỗ trợ của AI ở hầu hết các dạng tổn thương với $p < 0,05$ (trừ tràn dịch màng phổi), đặc biệt là với tổn thương tràn khí màng phổi. Tương tự, độ nhạy của bác sĩ khi có sự hỗ trợ của AI cao hơn so với đọc độc lập, với $p < 0,05$.



Hình 3. So sánh độ nhạy của bác sĩ (vàng), AI (xanh lá) và bác sĩ có hỗ trợ AI (đỏ)

Bảng 3. Độ đặc hiệu của các phương pháp đọc X-quang (Sp-95%CI)

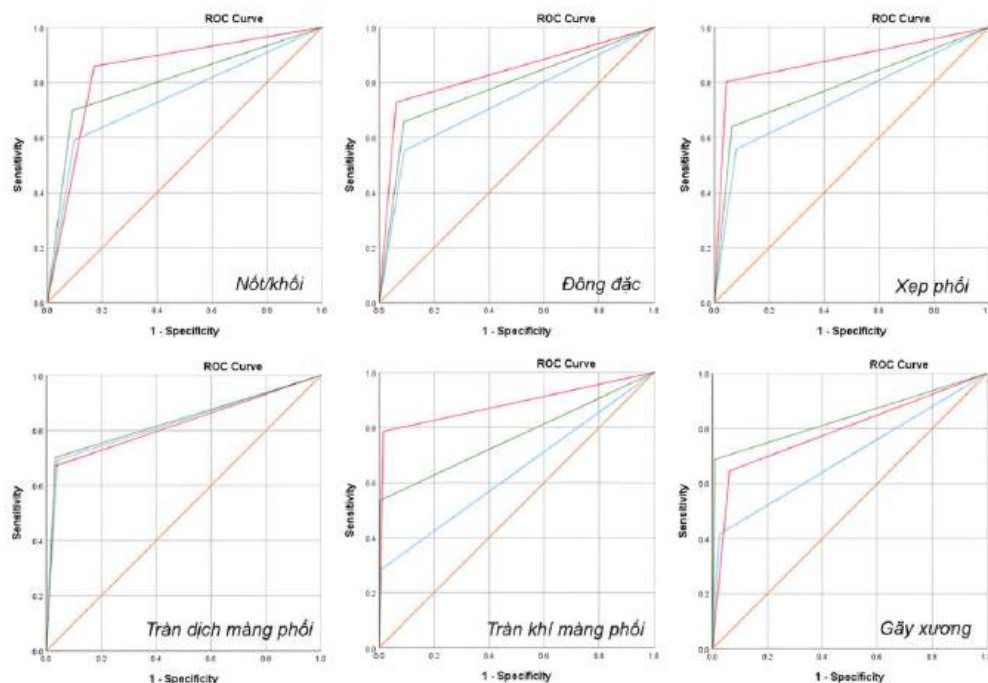
Tổn thương	Bác sĩ	AI	Bác sĩ + AI	p
Dạng nốt/khối	90,16 (87,85-92,47)	83* (80,08-85,92)	91 (88,78-93,22)	0,000
Đông đặc phổi	90,82 (88,58-93,06)	93,88* (92,02-95,74)	91,01 (88,79-93,23)	0,042
Xẹp phổi	92,01** (89,90-94,12)	95,64** (94,05-97,23)	93,65 (91,76-95,54)	0,002
Tràn dịch màng phổi	95,95 (94,42-97,48)	96,87 (95,52-98,22)	96,67 (95,28-98,06)	0,497
Tràn khí màng phổi	99,18 (98,48-99,88)	98,52** (97,58-99,46)	99,67** (99,22-100)	0,016
Gãy xương	97,45* (96,23-98,67)	93,72* (91,84-95,6)	99,15* (98,44-99,86)	0,000

P-value được tính chung cho sự khác biệt về độ đặc hiệu của các phương pháp chẩn đoán (Cochran's Q test)

** - Độ đặc hiệu của mô hình có sự khác biệt với các mô hình còn lại (Mc Nemar test với $p < 0.05$)*

*** - Độ đặc hiệu của 2 mô hình chẩn đoán có sự khác biệt*

Độ đặc hiệu của AI cao hơn của bác sĩ đọc độc lập với $p < 0,05$ đối với các tổn thương đông đặc phổi, xẹp phổi, tuy nhiên lại thấp hơn đối với dạng nốt/khối và gãy xương. Độ đặc hiệu của chẩn đoán tăng lên khi có sự hỗ trợ của AI so với bác sĩ đọc độc lập, tuy nhiên sự khác biệt chỉ có ý nghĩa thống kê đối với tổn thương gãy xương.



Hình 4. Biểu đồ đường cong ROC của các phương pháp phân tích X-quang của bác sĩ (xanh dương), AI (đỏ) và bác sĩ với sự hỗ trợ của AI (xanh lá)

Bảng 4. Diện tích dưới đường cong ROC của các phương pháp đọc X-quang (AUC-95%CI)

Tổn thương	Bác sĩ	AI	Bác sĩ + AI	p
Dạng nốt/khối	0,746* (0,695-0,797)	0,884* (0,807-0,882)	0,804* (0,757-0,850)	0,008
Đông đặc phổi	0,730* (0,672-0,789)	0,833* (0,783-0,884)	0,784* (0,730-0,838)	0,04
Xẹp phổi	0,739 (0,673-0,806)	0,879* (0,829-0,930)	0,788 (0,725-0,851)	0,08
Tràn dịch màng phổi	0,825 (0,768-0,883)	0,819 (0,761-0,878)	0,834 (0,778-0,891)	0,803
Tràn khí màng phổi	0,639* (0,515-0,763)	0,885* (0,795-0,976)	0,766* (0,648-0,884)	0,03
Gãy xương	0,696* (0,602-0,789)	0,792 (0,709-0,874)	0,840 (0,759-0,920)	0,22

P-value là giá trị lớn nhất khi so sánh từng cặp test chẩn đoán (Delong method)

**-AUC của mô hình khác với các mô hình còn lại với $p < 0,05$*

Diện tích dưới đường cong ROC của chẩn đoán tạo ra bởi AI cao hơn so với bác sĩ đọc độc lập hoặc có hỗ trợ của AI, ngoại trừ đối với tổn thương tràn dịch màng phổi. AUC của bác sĩ với hỗ trợ của AI cao hơn so với đọc độc lập ở tất cả các dạng tổn thương

Bảng 5. Lựa chọn của bác sĩ khi không đồng thuận với AI

Tổn thương	n	Giữ nguyên	Theo kết quả AI	Giữ nguyên đúng	AI đúng
Dạng nốt/khối	127	78 (61,4%)	49 (38,6%)	61 (48%)	66 (52%)
Đông đặc phổi	96	46 (47,9%)	50 (52,1%)	30 (31,3%)	66 (68,7%)
Xẹp phổi	69	32 (46,4%)	37 (53,6%)	14 (20,3%)	55 (79,7%)
Tràn dịch màng phổi	43	19 (44,2%)	24 (55,8%)	20 (46,5%)	23 (53,5%)
Tràn khí màng phổi	22	14 (63,6%)	8 (36,4%)	6 (27,3%)	16 (72,6%)
Gãy xương	59	37 (62,7%)	22 (37,3%)	35 (59,3%)	24 (40,7%)
Tổng	416	226 (54,3%)	190 (45,6%)	166 (39,9%)	250 (60,1%)

Có tổng cộng 416 trường hợp kết quả của bác sĩ đọc độc lập không đồng thuận với gợi ý của AI về sự hiện diện của một loại tổn thương nào đó. Sau khi tham khảo gợi ý của AI, đa phần các trường hợp bác sĩ sẽ giữ nguyên ý kiến của mình (54,3%), nhất là đối với tổn thương dạng nốt/khối, tràn khí màng phổi và gãy xương. Khi không đồng thuận, AI đúng trong 60,1% số ca, còn bác sĩ đúng 39,9%.

Bảng 6. Thời gian phân tích phim trung bình của bác sĩ

Mức kinh nghiệm	Đọc độc lập	Đọc với AI hỗ trợ	p	Tăng/giảm tuyệt đối	Tăng/giảm tương đối
Bác sĩ mới	85,39 ± 63,24s	64,85 ± 33,74s	0,000	-20,54s	24,05%
Bác sĩ có kinh nghiệm	74,27 ± 42,61s	60,78 ± 34,35s	0,000	-13,49s	18,16%
Chung	79,86 ± 54,21s	62,83 ± 34,08s	0,000	-17,03s	21,32%

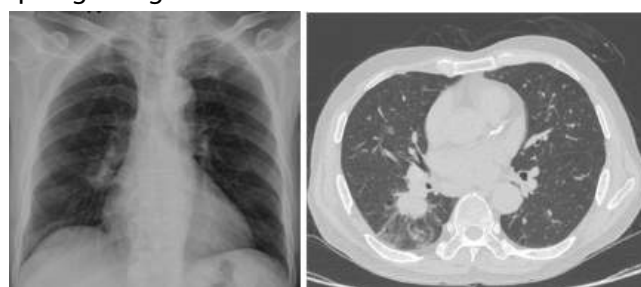
Thời gian phân tích phim của bác sĩ khi có AI hỗ trợ ngắn hơn so với khi đọc độc lập, sự khác biệt có ý nghĩa thống kê với $p < 0,05$. Thời gian đọc phim khi có sự hỗ trợ của AI được rút ngắn 21,32%, và ảnh hưởng của AI đến hiệu suất đọc phim lớn hơn ở các bác sĩ mới so với các bác sĩ có kinh nghiệm.

IV. BÀN LUẬN

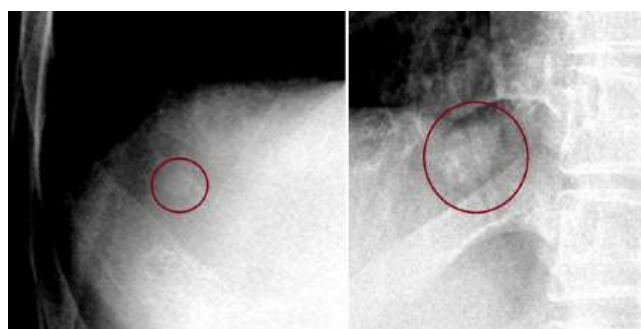
Tuổi trung bình của đối tượng nghiên cứu là 60.49 $\pm$ 15.6 tuổi, với giới tính chủ yếu là nam (Bảng 1), phù hợp với phân bố tuổi và giới trong các nghiên cứu tương tự, nguyên nhân chủ yếu là do tỉ lệ bệnh phổi cần tầm soát và thăm khám cao hơn ở nhóm bệnh nhân nam và cao tuổi^{4,7,8}.

Vì sự khác biệt trong tiêu chuẩn đối chiếu, độ nhạy, độ đặc hiệu của các báo cáo chẩn đoán hình ảnh trong nghiên cứu của chúng tôi đều tương đối thấp so với các nghiên cứu AI tương tự. Cụ thể, độ nhạy của bác sĩ đọc độc lập cao nhất đối với tổn thương dạng nốt/khối là 59,06% (95%CI 55,24-62,88) và thấp nhất chỉ 28,7% (95%CI 25,06-32,08) đối với tràn khí màng phổi (Bảng 2), thấp hơn so với các nghiên cứu tương tự sử dụng tiêu chuẩn đối chiếu là ý kiến chuyên gia^{6,9}, tuy nhiên lại tương đồng với các nghiên cứu so sánh hiệu quả của X-quang và CT trong chẩn đoán bệnh lý phổi trước đây^{10,11}. Trong nghiên cứu tương tự của Niehoff và cộng sự, trong số 499 phim X-quang có 237 trường hợp có CT đối chiếu, còn lại 262 trường hợp dựa vào sự đồng thuận của các chuyên gia và các phim X-quang khác của bệnh nhân, cho thấy độ nhạy của bác sĩ đọc từ 43%-78%, cao hơn so với nghiên cứu này nhưng thấp hơn so với các nghiên cứu không có CT đối chiếu⁷. Điều này cho thấy tầm quan trọng của việc có một tiêu chuẩn đối chiếu chính xác và minh bạch trong việc đánh giá độ chính xác của một mô hình đọc phim (bác sĩ, AI hoặc bác sĩ có sự hỗ trợ của AI). Đặc biệt đối với các tổn thương như nốt/khối nhỏ, đồng đặc phổi giai đoạn sớm, gây xương ít di lệch thì việc phát hiện trên X-quang càng trở nên khó khăn, với độ nhạy và độ đặc hiệu khá thấp^{10,11}. Có những tổn thương dù lớn, rõ ràng trên CT nhưng do sự chồng lấp hình ảnh nên cũng khó có thể phát hiện được trên phim X-quang ngực, dù cho được đọc bởi chuyên gia chẩn đoán hình ảnh

hiều năm kinh nghiệm (Hình 5). Trong khi đó, AI, với khả năng lượng hóa tổn thương và tổng hợp thông tin từ nguồn dữ liệu lớn, đang được hy vọng sẽ có khả năng phát hiện tổn thương vượt xa giới hạn thị giác của con người (Ví dụ các trường hợp trong hình 6). Vì vậy, việc sử dụng tiêu chuẩn đối chiếu dựa trên ý kiến chủ quan của các chuyên gia X-quang sẽ hạn chế tính chính xác của kết quả. Trong nghiên cứu này, các tổn thương được định vị trên CT sẽ được đối chiếu trên X-quang về khả năng hiển thị, để tạo ra tiêu chuẩn đối chiếu khách quan, chính xác nhất, với hy vọng đưa độ chính xác của X-quang tiến gần hơn tới CT.



Hình 5. Tổn thương dạng khối lớn trên CT (phải) nhưng do chồng lấp với rốn phổi nên khó phát hiện được trên X-quang (trái)



Hình 6. Các tổn thương sau cơ hoành, sát bờ tim phải điều chỉnh tương phản đến mức gần tối đa mới nhận biết được, dễ bị bỏ sót khi đọc bằng mắt nhưng được nhận diện bởi AI

Độ nhạy của AI DrAid trong phân tích phim X-quang ngực thẳng từ 64,58% (95%CI 60,87-68,29) đối với tổn thương gây xương tới 85,91% (95% CI 83,21-88,61) đối với dạng nốt/khối (Bảng 2), tương tự với nghiên cứu của Niehoff và cộng sự [7]. Độ nhạy này được cho là ở mức chấp nhận được đối với một test chẩn đoán, tuy nhiên trong trường hợp X-

quang ngực, với sự chồng lấp của tổn thương và các cấu trúc xung quanh, khi đối chiếu với CT, đây là một độ nhạy tương đối cao. Khi so sánh với bác sĩ đọc độc lập, sử dụng kiểm định Mc Nemar, độ nhạy của AI cao hơn nhiều so với bác sĩ (85,91% - 59,06% đối với tổn thương dạng nốt khối, 78,57% - 28,57% đối với tràn khí màng phổi), sự khác biệt có ý nghĩa thống kê ở hầu hết các loại tổn thương, trừ tràn dịch màng phổi. Điều này cho thấy AI vượt trội hơn bác sĩ chẩn đoán hình ảnh trong việc phát hiện tổn thương trên X-quang ngực. Tương tự như vậy, hiệu quả cải thiện chất lượng chẩn đoán cũng được thể hiện khi độ nhạy của bác sĩ có sự hỗ trợ của AI cao hơn so với bác sĩ đọc độc lập ở tất cả các dạng tổn thương, sự khác biệt có ý nghĩa thống kê đối với dạng nốt/khối, đông đặc, tràn khí màng phổi và gãy xương.

Độ đặc hiệu của AI cao nhất đối với tổn thương tràn khí màng phổi (98,52%; 95%CI 97,58-99,46) và thấp nhất ở dạng nốt/khối (83%; 95%CI 80,08-85,92) (Bảng 3). Một cách tương đối, khi so sánh giữa các nghiên cứu AI trong X-quang ngực khác nhau, độ nhạy càng cao tương ứng với việc phải chấp nhận với độ đặc hiệu thấp hơn [4, 8, 9]. Tuy nhiên, độ đặc hiệu của AI cũng cao hơn so với bác sĩ đọc độc lập, sự khác biệt có ý nghĩa thống kê với $p < 0,05$ ở hầu hết các tổn thương, trừ tràn dịch màng phổi. Khi so sánh kết quả đọc của bác sĩ có hoặc không có AI hỗ trợ, độ đặc hiệu cũng tăng lên đáng kể (từ 25,57%-69,15% lên 53,57%-70,21%) (Bảng 3). Về cơ bản, kết quả của nghiên cứu này đồng thuận với các tác giả khác về việc sự hỗ trợ của AI giúp nâng cao độ nhạy, độ đặc hiệu của chẩn đoán, cho thấy vai trò hỗ trợ đắc lực của AI trong việc phân tích phim X-quang ngực [2, 12, 13].

Tương tự như vậy, diện tích dưới đường cong ROC của các phương pháp chẩn đoán của AI cũng cao hơn của bác sĩ đọc độc lập. Sự hỗ trợ của AI giúp tăng AUC của kết quả chẩn đoán đối với tất cả các dạng tổn thương, từ 0,639-0,825 khi đọc độc lập lên 0,766-0,840 khi có sự hỗ trợ của AI (Bảng 4). Kết quả này đồng thuận với các nghiên cứu tương tự cho thấy chẩn đoán của AI có AUC cao hơn của bác sĩ và chẩn đoán của bác sĩ có hỗ trợ của AI sẽ giúp tăng AUC so với chẩn đoán độc lập [3, 6, 9]. Hình 4 cho

thấy biểu đồ đường cong ROC của các mô hình chẩn đoán, trong đó AUC của AI là cao nhất trong hầu hết các loại tổn thương.

Ngoài hiệu quả nâng cao tính chính xác của chẩn đoán X-quang ngực, AI còn giảm thời gian chẩn đoán, trung bình bác sĩ tiết kiệm được khoảng 17 giây tương đương với 21% thời gian phân tích phim khi có sự hỗ trợ của AI, sự khác biệt có ý nghĩa thống kê với $p < 0,05$. Hiệu quả hỗ trợ gia tăng hiệu suất phân tích phim cao hơn ở nhóm bác sĩ mới (24,05%) so với nhóm bác sĩ có kinh nghiệm (18,16%). Trong các nghiên cứu tương tự, AI giúp giảm 10-41% thời gian chẩn đoán ^{2, 3, 9}, và tác dụng này rõ ràng hơn ở các nhóm bác sĩ có số năm kinh nghiệm ít hơn hoặc không chuyên về chẩn đoán hình ảnh lồng ngực ^{2, 3}. Điều này cho phép giảm áp lực công việc cho bác sĩ chẩn đoán hình ảnh và nâng cao hiệu suất làm việc. Đặc biệt là trong bối cảnh của các đơn vị cấp cứu, việc nâng cao hiệu suất đọc phim giúp tranh thủ được nhiều thời gian quý giá hơn cho bệnh nhân cũng như chú trọng thêm được vào các công tác chăm sóc và điều trị khác.

Tuy nhiên, điều thú vị là trong nghiên cứu này, kết quả cho thấy độ nhạy, độ đặc hiệu và AUC của bản thân AI cao hơn so với bác sĩ có sự hỗ trợ của AI ở đa số các loại tổn thương được khảo sát, trừ tràn dịch màng phổi và gãy xương. Đối với tổn thương tràn dịch màng phổi, khi bệnh nhân chụp ở tư thế nằm, việc xác định tràn dịch thông qua sự chênh lệch mức độ cản quang lan tỏa của hai trường phổi, tuy nhiên điều này phụ thuộc rất nhiều vào kinh nghiệm của bác sĩ chẩn đoán hình ảnh và AI sẽ gặp khó khăn trong việc nhận diện sự có mặt của dịch trong trường hợp này. Còn đối với tổn thương gãy xương, độ nhạy của AI cao hơn nhiều so với bác sĩ (64,58%; 95%CI 60,87-68,29 so với 41,67%; 95%CI 37,84-45,50) nên khi các tổn thương được đánh dấu nhờ AI, các bác sĩ có thể dễ dàng nhận diện, từ đó tăng độ nhạy, độ đặc hiệu và AUC. Tuy nhiên, để lý giải được vì sao trong đa phần các tổn thương khác thì độ nhạy, độ đặc hiệu và AUC của AI lại cao hơn so với bác sĩ dù đã tham khảo gợi ý của AI, các trường hợp không đồng thuận giữa AI và bác sĩ đọc độc lập đã được lọc ra và tiến hành phân tích (Bảng 5). Kết

quả cho thấy, trong hầu hết các trường hợp không đồng thuận, khi có sự hỗ trợ của AI, các bác sĩ có xu hướng giữ nguyên lựa chọn ban đầu của mình (54,3%) hơn là nghe theo AI. Trong khi đó, khoảng 60% trong số đó thì gợi ý của AI là chính xác. Chính vì vậy, điều này đặt ra một vấn đề về việc hiệu quả hỗ trợ nâng cao độ nhạy, độ đặc hiệu và tính chính xác của AI sẽ bị ảnh hưởng bởi niềm tin vào AI của bác sĩ chẩn đoán hình ảnh. Tuy nhiên, cho tới thời điểm hiện tại, tính minh bạch và tính khó giải thích của AI vẫn đang là một vấn đề ảnh hưởng tới niềm tin của các bác sĩ sử dụng. Trong một nghiên cứu bao gồm 113 bác sĩ và 15 kỹ thuật viên chẩn đoán hình ảnh về niềm tin vào chẩn đoán của AI, chỉ có 29,7% tin tưởng, 42,2% không chắc chắn và tới 28,1% không tin vào các báo cáo của AI¹⁴. Điều này đòi hỏi các công cụ AI cần có những nghiên cứu toàn diện, khách quan để chứng minh được hiệu quả của mình, đồng thời cũng cần có những thuật toán để giải thích, minh họa và tăng tính thuyết phục cho các gợi ý mà AI đưa ra (ví dụ như thông qua việc đánh dấu chính xác vị trí tổn thương, cung cấp hình ảnh đã phân tách, điều chỉnh và các thông số liên quan), từ đó cung cấp đầy đủ dữ kiện hơn để bác sĩ chẩn đoán hình ảnh đưa ra quyết định cuối cùng một cách chính xác nhất.

Tuy nhiên, việc tin tưởng vào AI cũng cần cân bằng với sự tỉnh táo khi sử dụng. Bởi vì mỗi công cụ AI có những điểm mạnh và điểm yếu riêng, chúng có thể nhạy với tổn thương này nhưng lại kém nhạy với một số tổn thương khác^{8, 15}. Đồng thời việc đào tạo AI mà không có các thuật toán phù hợp với y học dẫn tới AI đang tạo ra các con đường tắt trong chẩn đoán, dễ dẫn tới sai lầm ví dụ như sử dụng các marker của ICU để định hướng đến viêm phổi hay dựa vào sự hiện diện của ống dẫn lưu để kết luận tràn khí màng phổi [16, 17]. Chính vì vậy, để phát huy tối đa hiệu quả hỗ trợ của AI, bác sĩ chẩn đoán hình ảnh phải cân bằng giữa niềm tin và sự hiểu biết về AI để đưa ra lựa chọn cuối cùng chính xác nhất.

V. KẾT LUẬN

AI cho thấy độ nhạy, độ đặc hiệu và AUC cao trong chẩn đoán X-quang ngực thẳng. Đồng thời

việc sử dụng AI không chỉ giúp nâng cao độ chính xác mà còn giúp tiết kiệm thời gian chẩn đoán, giảm gánh nặng công việc và nâng cao hiệu suất phân tích hình ảnh X-quang ngực. Để tối đa hóa hiệu quả hỗ trợ của AI trong quá trình phân tích X-quang nói riêng và chẩn đoán hình ảnh nói chung, việc hiểu rõ về mô hình đang được sử dụng để đưa ra quyết định đúng đắn cuối cùng là điều hết sức cần thiết.

TÀI LIỆU THAM KHẢO

1. Guillermo Avendaño Cervantes (2016), *Radiographic quality, Technical Fundamentals of Radiology and CT/OP Publishing*. p. 2-1-2-11.
2. Souhail Bennani, Nor-Eddine Regnard et al (2023) *Using AI to Improve Radiologist Performance in Detection of Abnormalities on Chest Radiographs* 309(3): 230860.
3. Lili Guo, Changsheng Zhou et al (2024) *Deep Learning for Chest X-ray Diagnosis: Competition Between Radiologists with or Without Artificial Intelligence Assistance*, Journal of Imaging Informatics in Medicine 37(3): 922-934.
4. Lind Plesner L, Müller FC et al (2023) *Commercially Available Chest Radiograph AI Tools for Detecting Airspace Disease, Pneumothorax, and Pleural Effusion*, Radiology, 308(3), p. e231236.
5. Y. Vasilev, A. Vladzymirskyy et al. (2023), *AI-Based CXR First Reading: Current Limitations to Ensure Practical Value*, Diagnostics (Basel), 13(8).
6. J. Hofmeister N Garin et al (2024) *Validating the accuracy of deep learning for the diagnosis of pneumonia on chest x-ray against a robust multimodal reference diagnosis: a post hoc analysis of two prospective studies*, Eur Radiol Exp 8(1): 20.
7. Julius Henning Niehoff, Jana Kalaitzidis et al (2023) *Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays*, Scientific Reports 13(1): 3680.
8. Queralt Miró Catalina, Josep Vidal-Alaball et al (2024), *Real-world testing of an artificial intelligence algorithm for the analysis of chest X-rays in primary care settings*, Scientific Reports 14(1): 5199.
9. Ahn JS, Ebrahimian S et al (2022) *Association of Artificial Intelligence-Aided Chest Radiograph*

- Interpretation With Reader Performance and Efficiency*, JAMA Netw Open 5(8): 2229289.
10. Shelmerdine SC, Langan D et al (2018) *Chest radiographs versus CT for the detection of rib fractures in children (DRIFT): A diagnostic accuracy observational study*. Lancet Child Adolesc Health 2(11): 802-811.
 11. Self WH, Courtney DM et al (2013) *High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ED patients: Implications for diagnosing pneumonia*, Am J Emerg Med 31(2): 401-405.
 12. Pamela G Anderson, Hannah Tarder-Stoll et al (2024) *Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays*, Scientific Reports 14(1): 25151.
 13. Lee HW, Jin KN et al (2023) *Artificial Intelligence Solution for Chest Radiographs in Respiratory Outpatient Clinics: Multicenter Prospective Randomized Clinical Trial*. Ann Am Thorac Soc 20(5): 660-667.
 14. Hussein Soleimantabar, Ali Mahdavi et al (2025) *Artificial Intelligence in Radiology: Perceptions, Adoption Barriers, and Trust Among Iranian Radiologists in a Global Context & InfoScience Trends* 2(3): 1-12.
 15. Niehoff JH, Kalaitzidis J et al (2023) *Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays*. Sci Rep 13(1): 3680.
 16. Banerjee I, Bhattacharjee K et al (2023) *"Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation*. J Am Coll Radiol 20(9): 842-851.
 17. Brandon G Hill, Frances L Koback et al (2024) *The risk of shortcutting in deep learning algorithms for medical imaging research*, Scientific Reports 14(1): 29224.