

PHÁT HIỆN LOÃNG XƯƠNG DỰA TRÊN ẢNH X-RAY CỘT SỐNG SỬ DỤNG HỌC SÂU KẾT HỢP THÔNG TIN LÂM SÀNG

Vũ Xuân Hạnh¹, Vũ Thiện Vinh², Đỗ Duy Trinh¹, Ngô Văn Sơn¹

Email: hanhvx@hou.edu.vn, ORCID: 0000-0002-6739-1494

Ngày tòa soạn nhận được bài báo: 05/12/2025. Ngày phản biện đánh giá: 05/06/2026.

Ngày bài báo được duyệt đăng: 23/06/2026.

DOI: 10.59266/houjs.2026.1333

Tóm tắt: Loãng xương làm tăng nguy cơ gãy xương và thường được chẩn đoán bằng DXA (Dual-energy X-ray Absorptiometry), tuy nhiên thiết bị này có chi phí cao và chưa phổ biến tại nhiều cơ sở y tế tuyến dưới. Nghiên cứu này đề xuất quy trình sàng lọc loãng xương hai bước, không sử dụng các biến DXA ở đầu vào, dựa trên ảnh X-ray cột sống thắt lưng kết hợp thông tin lâm sàng. Quy trình được đánh giá trên tập X-ray của bộ dữ liệu LUMOS gồm 803 bệnh nhân và 1.620 ảnh. Ở bước đầu, OsteoNet và FCoTNet được dùng để trích xuất vector đặc trưng 128 chiều từ ảnh X-ray. Bước tiếp theo, vector này được kết hợp với các nhóm đặc trưng lâm sàng và đưa vào XGBoost để phân loại ba nhóm: Bình thường, Giảm mật độ xương và Loãng xương. Kết quả cho thấy các tổ hợp chứa biến DXA như mật độ loãng xương (Bone Mineral Density - BMD) hoặc T-score đều đạt AUC = 1,000 do nhận được định nghĩa trực tiếp từ T-score, phản ánh rõ ràng và không có ý nghĩa lâm sàng. Sau khi loại bỏ biến DXA, baseline chỉ dùng thông tin lâm sàng đạt AUC = 0,709; trong khi OsteoNet + Demo+Morph đạt AUC = 0,761. Kiểm định DeLong cho thấy cải thiện có ý nghĩa so với chỉ dùng embedding ảnh ($p = 0,0002$) và ở mức cận biên so với baseline lâm sàng thuần ($p = 0,050$).

Từ khóa: loãng xương, học sâu, X-Ray cột sống, vector embedding, XGBoost, LUMOS, DeLong test

I. Đặt vấn đề

Loãng xương là bệnh lý đặc trưng bởi sự suy giảm mật độ khoáng chất trong xương, khiến xương trở nên giòn và dễ gãy. Theo Johnell và Kanis (2006), bệnh ảnh hưởng đến hơn 200 triệu người trên thế giới và gây ra hơn 8,9 triệu ca gãy xương mỗi năm. Tổ chức Y tế Thế giới (World Health Organization, 2003) định nghĩa loãng xương được xác định khi chỉ số T-score đo bằng máy DXA

(Dual-energy X-ray Absorptiometry, phương pháp đo hấp thụ tia X năng lượng kép) ≤ -2.5 độ lệch chuẩn, tức mức độ chênh lệch mật độ xương của bệnh nhân so với người trẻ khỏe mạnh cùng chủng tộc và giới tính.

Máy DXA thường được sử dụng để chẩn đoán, nhưng có hai vấn đề lớn: (i) chi phí đầu tư máy rất cao (50,000-200,000 USD/máy), (ii) không phải bệnh viện nào cũng có (Kanis và cộng sự, 2019).

¹ Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

² Bệnh viện đa khoa Y học cổ truyền Hà Nội, Hà Nội, Việt Nam

Ngược lại, ảnh X-ray cột sống được chụp thường quy tại hầu hết các bệnh viện với chi phí thấp hơn đáng kể. Tuy nhiên, việc đọc X-ray bằng mắt thường để chẩn đoán loãng xương có độ chính xác hạn chế và phụ thuộc nhiều vào kinh nghiệm của bác sĩ (Genant & cộng sự, 1993).

Gần đây, học sâu đã chứng minh khả năng trích xuất thông tin về mật độ xương từ ảnh X-Ray (Yamamoto & cộng sự, 2020); (Hsieh & cộng sự, 2021). Tuy nhiên, một vấn đề quan trọng thường bị bỏ qua trong các nghiên cứu này là việc đưa trực tiếp kết quả đo DXA (BMD, T-score) vào đặc trưng đầu vào, trong khi nhãn phân loại lại được xác định chính từ những giá trị đó, dẫn đến hiện tượng rò rỉ nhãn (label leakage).

Nghiên cứu này có bốn đóng góp chính: (i) đề xuất quy trình hai bước gồm: mạng học sâu trích xuất đặc trưng và XGBoost phân loại sử dụng thông tin lâm sàng; (ii) so sánh hai kiến trúc học sâu OsteoNet và FCoTNet để xác định mô hình phù hợp nhất cho bài toán; (iii) phân tích và định lượng mức độ ảnh hưởng của từng nhóm đặc trưng DXA đến kết quả phân loại; (iv) xác định 17 đặc trưng nhân trắc và hình thái đốt sống là tổ hợp thông tin lâm sàng tối ưu, không yêu cầu đo DXA.

Bảng 1. Thống kê phân chia dữ liệu bộ dữ liệu LUMOS

Tập dữ liệu	Tổng	Bình thường	Giảm mật độ	Loãng xương
Huấn luyện	560	202	204	154
Kiểm định	120	43	44	33
Kiểm tra	123	44	43	36

2.2. Tiền xử lý ảnh

Ảnh X-Ray được lưu dưới dạng tệp DICOM (Digital Imaging Communications in Medicine). Quy trình tiền xử lý gồm 4 bước: (1) Đọc giá trị pixel

II. Phương pháp nghiên cứu

2.1. Bộ dữ liệu LUMOS

Nghiên cứu sử dụng bộ dữ liệu LUMOS được Shi và cộng sự (2025) công bố. Bộ dữ liệu gồm 803 bệnh nhân người Châu Á với: ảnh X-Ray cột sống thắt lưng chụp từ hai góc (nhìn thẳng AP và nhìn ngang LAT), tổng cộng 1,620 ảnh; ảnh CT cho 280 bệnh nhân; và bảng thông tin lâm sàng với 49 chỉ số mỗi người.

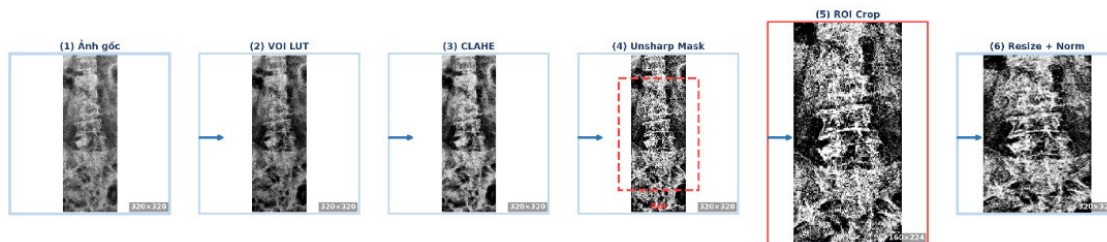
Mỗi bệnh nhân được phân vào một trong ba nhóm dựa trên chỉ số T-score đo từ DXA theo tiêu chuẩn WHO: Bình thường ($T > -1.0$; 289 người), Giảm mật độ xương ($-2.5 < T \leq -1.0$; 291 người), và Loãng xương ($T \leq -2.5$; 223 người). Dữ liệu được chia theo tỷ lệ 70/15/15 cho các tập huấn luyện, kiểm định và kiểm tra, đảm bảo mỗi bệnh nhân chỉ xuất hiện trong đúng một tập.

Tuy nhiên, nghiên cứu này chỉ sử dụng ảnh X-Ray vì: (i) tập con CT chỉ chiếm 35% tổng số bệnh nhân, dẫn đến tập kiểm tra CT chỉ có 34 mẫu, quá nhỏ để đưa ra kết luận thống kê đáng tin cậy; (ii) mục tiêu của nghiên cứu là đề xuất quy trình sàng lọc thực tế tại các cơ sở y tế tuyến dưới, nơi X-Ray sẵn có phổ biến hơn nhiều so với CT.

từ DICOM và áp dụng hệ số hiệu chỉnh theo thông số thiết bị chụp; (2) Điều chỉnh độ tương phản theo thông số cửa sổ sáng có sẵn trong file DICOM; nếu thông số này không có thì dùng phương pháp kéo căng

biểu đồ mức xám từ phần trăm thứ 1 đến 99; (3) Đệm ảnh về hình vuông bằng cách thêm viền màu xám trung tính, giúp giữ

tỉ lệ ảnh gốc khi thay đổi kích thước; (4) Thay đổi kích thước về 320×320 pixel bằng thuật toán Lanczos4.



Hình 1. Minh họa quy trình xử lý X-Ray

Tăng cường dữ liệu: Để hạn chế hiện tượng quá khớp (overfitting), mỗi ảnh đưa vào huấn luyện được biến đổi ngẫu nhiên thông qua các phép: lật ngang, xoay $\pm 5^\circ$, dịch chuyển vị trí $\pm 5\%$, thay đổi nhẹ độ sáng/tương phản, thêm nhiễu Gaussian nhẹ, và che ngẫu nhiên 1-4 vùng nhỏ.

2.3. Phân nhóm thông tin lâm sàng

Bộ dữ liệu LUMOS cung cấp 49 chỉ số lâm sàng mỗi bệnh nhân, được chia thành 4 nhóm như Bảng 2. Cần lưu ý rằng nhãn phân loại (Bình thường/

Giảm mật độ/Loãng xương) được xác định trực tiếp từ chỉ số T-score DXA ($T \leq -2.5 \rightarrow$ Loãng xương). Do đó, nếu đưa T-score hoặc BMD vào đặc trưng đầu vào, mô hình không thực sự học cách nhận biết loãng xương từ ảnh, mà chỉ đọc lại giá trị T-score để suy ra nhãn, dẫn đến kết quả đánh giá rất cao nhưng không có giá trị thực tiễn. Nhóm đặc trưng hình thái đốt sống là kết quả đo thủ công của chuyên gia X-ray, công bố cùng bộ dữ liệu LUMOS, không phải output của thuật toán tự động.

Bảng 2. Phân nhóm 49 đặc trưng lâm sàng theo nguồn gốc

Nhóm	Số đặc trưng	Các đặc trưng và ý nghĩa	Nguồn gốc
Thông tin cá nhân (Demo)	5	Tuổi (năm), Giới tính (M/F), Chiều cao (cm), Cân nặng (kg), BMI (kg/m^2)	Hỏi bệnh/đo nhân trắc
Hình thái đốt sống (Morph)	12	avgh_Lx-Ly: chiều cao trung bình đốt sống (mm) avgw_Lx-Ly: chiều rộng trung bình đốt sống (mm)	Đo từ ảnh X-Ray
Mật độ xương DXA (BMD/BMC)	13	bmd: mật độ khoáng xương (g/cm^2) bmd_Lx-Ly: BMD tại từng cặp L1-L4 bmc_Lx-Ly: lượng khoáng chất xương (g)	Đo bằng máy DXA
Điểm số DXA (T/Z/Area)	19	T-score / T_Lx-Ly: độ lệch chuẩn so với người trẻ (nhãn được định nghĩa từ đây) Z-score: độ lệch chuẩn so với người cùng tuổi Diện tích chiếu xương (cm^2)	Tính từ kết quả DXA

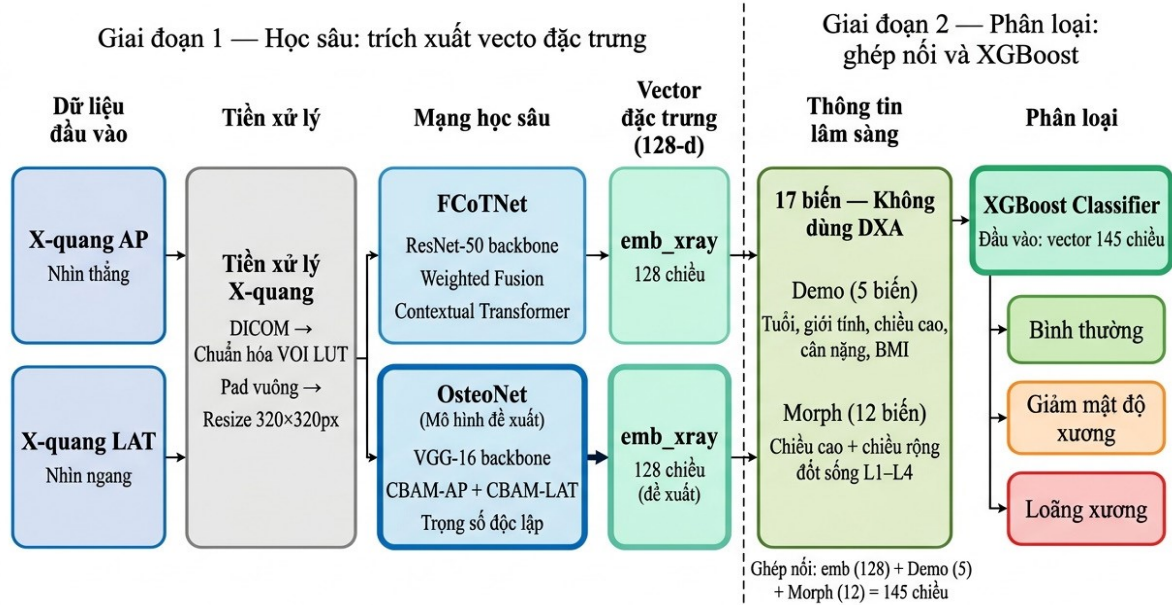
2.4. Trích xuất đặc trưng hình ảnh sử dụng học sâu

Thay vì để mạng học sâu vừa trích xuất đặc trưng vừa đưa ra quyết định phân loại cuối cùng trong một mô hình

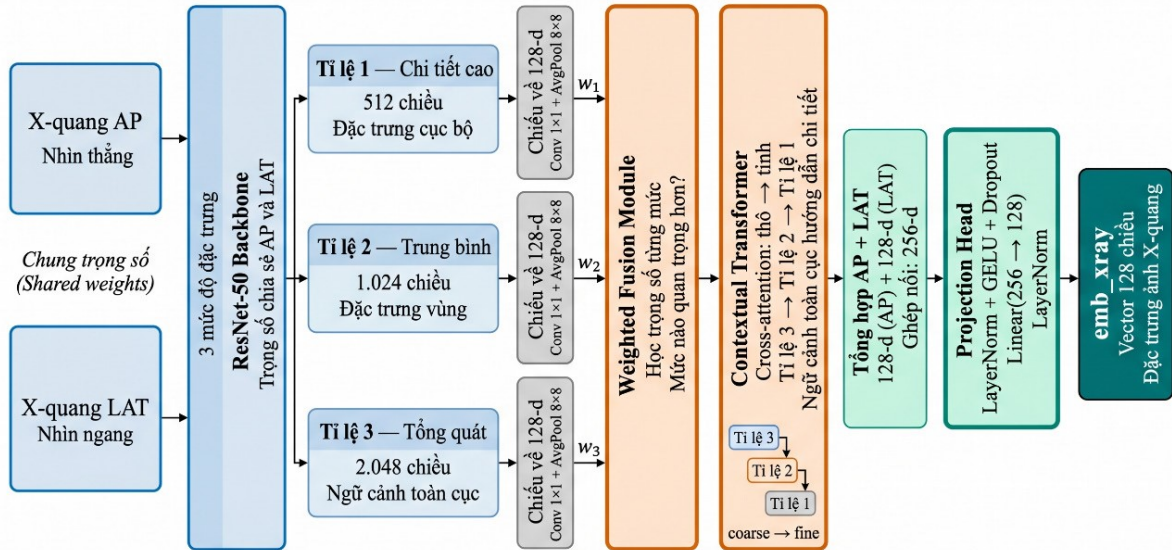
end-to-end, nghiên cứu này tách bài toán thành hai bước độc lập (Hình 2). Ở bước thứ nhất, mạng học sâu được huấn luyện nhằm chuyển đổi ảnh X-ray thành một vector đặc trưng số có kích thước cố định

(embedding), đóng vai trò biểu diễn thông tin hình ảnh ở mức trừu tượng. Ở bước thứ hai, các vector này được kết hợp với thông tin lâm sàng của bệnh nhân và đưa vào mô hình XGBoost để thực hiện phân loại.

Thiết kế này giúp tách riêng quá trình học biểu diễn ảnh và quá trình đánh giá đóng góp của từng nhóm đặc trưng, đồng thời cho phép thử nghiệm nhiều tổ hợp đầu vào mà không cần huấn luyện lại mạng học sâu.



Hình 2. Sơ đồ quy trình hai bước: (1) mạng học sâu trích xuất vector đặc trưng từ ảnh; (2) XGBoost kết hợp vector với thông tin cá nhân để phân loại



Hình 3. Kiến trúc FCoTNet

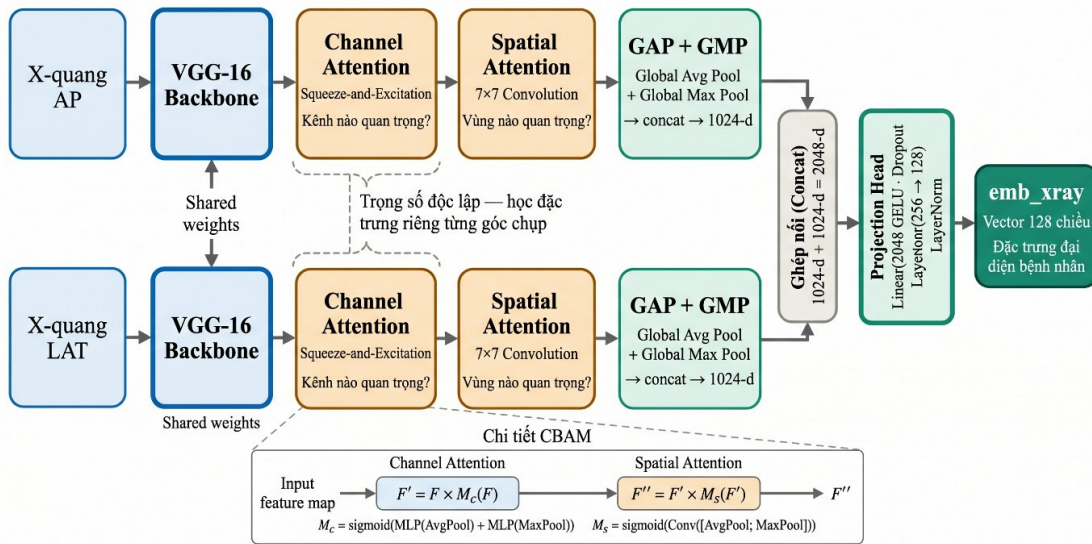
Mạng FCoTNet: FCoTNet được đề xuất bởi Xue và cộng sự (2023). Mạng sử dụng ResNet-50 làm nền tảng, trích xuất đặc trưng ở ba mức độ chi tiết khác nhau: mức chi tiết nhỏ (512 chiều), mức

vừa (1.024 chiều), và mức tổng thể (2.048 chiều). Thay vì kết hợp cố định ba mức này, mạng học trọng số động cho từng mức thông qua cơ chế kết hợp có trọng số (Weighted Fusion, sử dụng hàm Softmax).

Tiếp đó, mô-đun Contextual Transformer áp dụng cơ chế cross-attention từ mức tổng thể xuống mức chi tiết, giúp ngữ cảnh toàn cục định hướng việc phân tích chi tiết cục bộ. Cả hai góc chụp AP và LAT được xử lý qua cùng một bộ trọng số, cho kết quả đầu ra là vector đặc trưng 128 chiều cho mỗi bệnh nhân.

Mạng OsteoNet: OsteoNet được đề xuất bởi Muzaffar và cộng sự (2025) là kiến trúc cho kết quả tốt nhất trong nghiên cứu này, sử dụng VGG-16 làm nền tảng.

Điểm nổi bật của mạng là cơ chế CBAM (Convolutional Block Attention Module), kết hợp hai loại chú ý: theo kênh (Channel Attention) và theo không gian (Spatial Attention). Đặc biệt, mỗi góc chụp AP và LAT được trang bị một bộ CBAM riêng với trọng số hoàn toàn độc lập, cho phép mạng học các đặc trưng bổ sung từ từng góc chụp: AP thể hiện chiều rộng đốt sống, trong khi LAT thể hiện chiều cao đốt sống. Kết quả từ hai góc sau đó được ghép lại và chiếu xuống vector đặc trưng 128 chiều.



Hình 4. Kiến trúc OsteoNet

2.5. Ghép nối vector đặc trưng với thông tin lâm sàng

Sau khi có vector đặc trưng từ ảnh (128 chiều), bước tiếp theo là ghép nối

với thông tin lâm sàng của bệnh nhân để tạo vector đầu vào cho XGBoost. Bảng 3 liệt kê tất cả các tổ hợp được thử nghiệm.

Bảng 3. Các tổ hợp vector đặc trưng được thử nghiệm

STT	Tên tổ hợp	Thành phần	Số chiều
1	Chỉ embedding	vector ảnh X-Ray (128 chiều)	128
2	Emb + Thông tin cá nhân	vector + tuổi/giới/chiều cao/cân/BMI (5)	133
3	Emb + Hình thái đốt sống	vector + avgh+avgw (12)	140
4	Emb + Demo+Morph	vector + thông tin cá nhân (5) + hình thái (12)	145
5	Emb + BMD	vector + mật độ xương DXA (7)	135
6	Emb + T-score	vector + điểm T-score DXA (7)	135
7	Emb + Toàn bộ DXA	vector + tất cả 32 đặc trưng	160

2.6. Phương pháp đánh giá

AUC (Area Under the ROC Curve) - diện tích dưới đường cong ROC (Receiver Operating Characteristic) được xây dựng từ mối quan hệ giữa TPR (True Positive Rate) và FPR (False Positive Rate):

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} \quad (1)$$

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \quad (2)$$

Trong nghiên cứu này, chúng tôi sử dụng AUC macro theo phương pháp one-vs-rest vì nó đánh giá đồng đều cả 3 nhóm bệnh nhân, kể cả nhóm thiểu số.

F1 macro và F1 theo từng lớp: F1 là trung bình điều hòa giữa Recall (khả năng phát hiện đầy đủ các ca bệnh) và Precision (khả năng hạn chế chẩn đoán nhầm).

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{Var(AUC_1) + Var(AUC_2) - 2 \cdot Cov(AUC_1, AUC_2)}} \quad (6)$$

2.7. Giao thức huấn luyện

Tất cả mô hình học sâu được huấn luyện trên GPU NVIDIA RTX 3050, sử dụng bộ tối ưu hóa AdamW với chiến lược tốc độ học phân biệt: lớp cuối cùng có $LR = 1 \times 10^{-4}$, trong khi lớp backbone có $LR = 1 \times 10^{-5}$ nhằm bảo toàn tri thức đã học từ tiền huấn luyện. Số epoch được đặt là 50, với tốc độ học giảm dần theo lịch trình cosine. Để xử lý mất cân bằng lớp, mỗi lớp được gán trọng số tỷ lệ nghịch với số mẫu: [2.78; 2.76; 3.60] tương ứng với các nhóm Bình thường, Giảm mật độ và Loãng xương.

Với XGBoost (Chen và Guestrin, 2016), chúng tôi sử dụng tối ưu hóa Bayes

$$F1_{macro} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (4)$$

Accuracy (Độ chính xác): Tỷ lệ phân loại đúng trên tổng số mẫu.

$$Acc = \frac{\text{số mẫu dự đoán đúng}}{\text{tổng số mẫu}} \quad (5)$$

Kiểm định DeLong test (DeLong và cộng sự, 1988): so sánh AUC của hai mô hình trên cùng một tập kiểm tra. Do hai AUC được tính trên cùng một nhóm bệnh nhân nên không độc lập với nhau, vì vậy không thể áp dụng t-test thông thường. Phương pháp DeLong dựa trên thống kê U - một lớp ước lượng thống kê để xử lý đúng mối tương quan này: thống kê kiểm định Z tuân theo phân phối chuẩn, từ đó tính được giá trị p hai phía; giả thuyết hai AUC bằng nhau bị bác bỏ khi $p < \alpha = 0,05$:

qua thư viện Optuna (Akiba và cộng sự, 2019) với 100 lượt thử. Không gian tham số: $max_depth \in [3, 9]$, $learning_rate \in [0,01; 0,3]$, $n_estimators \in [100, 800]$, $subsample$ và $colsample_bytree \in [0,5; 1,0]$, $gamma \in [0; 4]$, $min_child_weight \in [1, 8]$, reg_alpha và $reg_lambda \in [0; 4]$. Bộ tham số tốt nhất được lựa chọn dựa trên AUC trên tập val sau đó mô hình được huấn luyện lại trên tập train, val trước khi đánh giá trên tập test.

III. Kết quả và thảo luận

3.1. Ảnh hưởng của các nhóm đặc trưng tới kết quả phân loại

Bảng 4 cho thấy rõ ảnh hưởng của từng nhóm đặc trưng DXA đến kết quả

phân loại. Giá trị baseline 0.608 là kết quả AUC của cấu hình XGBoost chỉ sử dụng vector đặc trưng ảnh từ một lần chạy cụ thể, được dùng làm mốc tham chiếu để đo ΔAUC khi bổ sung từng nhóm đặc trưng. Khi bổ sung BMD hoặc T-score, AUC đạt 1.000 một cách nhất quán, do nhân ba lớp được định nghĩa trực tiếp từ T-score.

Bảng 4. Ảnh hưởng của từng nhóm đặc trưng đến AUC phân loại

Bộ thông tin	Số đặc trưng	AUC baseline	AUC sau khi thêm	ΔAUC	Mức độ ảnh hưởng	Giải thích
Demo+Morph	17	0.608	0,705	+0,097	Không	Đề xuất dùng
Area	6	0.608	0.658	+0.050	Không	Diện tích chiếu (cm^2)
Z-score	6	0.608	0.818	+0.209	Gián tiếp	$Z \approx T$ chuẩn hóa tuổi
BMC	6	0.608	0.866	+0.258	Gián tiếp	Tương quan cao BMD
BMD	7	0.608	1.000	+0.392	Trực tiếp	$BMD \rightarrow T\text{-score}$
T-score	7	0.608	1.000	+0.392	Trực tiếp	Biến định nghĩa nhân
BMD+T-score	14	0.608	1.000	+0.392	Trực tiếp	$Nhân \leftarrow T\text{-score}$
Toàn bộ DXA	32	0.608	1.000	+0.392	Trực tiếp	Có nhóm DXA

3.2. Vai trò của thông tin lâm sàng khi không có ảnh

Để định lượng vai trò riêng của thông tin lâm sàng, chúng tôi đánh giá ba cấu hình XGBoost không sử dụng ảnh: Demo, Morph, và Demo+Morph. Bảng 5 cho thấy 5 đặc trưng Demo ($AUC = 0,704$) cao hơn rõ rệt so với 12 đặc trưng Morph

($AUC \approx 0,59$). Tuổi và BMI là yếu tố ảnh hưởng trực tiếp đến loãng xương, do đó nhóm đặc trưng nhân khẩu học mang giá trị tiên đoán mạnh hơn các phép đo hình học đốt sống đơn lẻ. Cấu hình Demo+Morph đạt $AUC \approx 0,71$, qua đó thiết lập ngưỡng tham chiếu để đánh giá đóng góp thực sự của ảnh X-Ray ở các phần phân tích tiếp theo.

Bảng 5. Kết quả XGBoost chỉ dùng thông tin lâm sàng

Tổ hợp đặc trưng	Số đặc trưng	AUC	F1 macro	Acc
Demo	5	0,704	0,477	0,484
Morph	12	0,589	0,348	0,379
Demo + Morph	17	0,709	0,501	0,500

3.3. Tổng hợp kết quả

Bảng 6 trình bày kết quả các cấu hình kết hợp vector đặc trưng ảnh và thông tin lâm sàng. Cấu hình tốt nhất là OsteoNet + Demo+Morph đạt $AUC = 0,761$ và $F1 \text{ macro} = 0,501$, cải thiện $\Delta AUC = +0,049$ so với OsteoNet Emb only (0,712) và $+0,052$ so với baseline lâm sàng thuần. Bốn quan sát chính được

rút ra: (i) OsteoNet vượt trội hơn FCoTNet ở mọi tổ hợp ($\Delta AUC \approx +0,05$); (ii) với OsteoNet, việc bổ sung thông tin lâm sàng luôn cải thiện kết quả: Demo (+0,016), Morph (+0,019), Demo+Morph (+0,049); (iii) với FCoTNet, Demo (+0,049) đóng góp mạnh hơn Morph (+0,020); (iv) cấu hình FCoTNet Emb only (0,654) thấp hơn baseline lâm sàng thuần (0,709), cho thấy

embedding ảnh đơn thuần không thể thay thế 17 đặc trưng nhân khẩu học cơ bản.

Phân tích F1 từng lớp cho thấy Giảm mật độ xương (F1-GM = 0,360-0,500) luôn là lớp khó nhất do nằm ở vùng

Bảng 6. Kết quả phân loại của các cấu hình kết hợp emb ảnh và thông tin lâm sàng không chứa DXA

Kiến trúc	Tổ hợp đặc trưng	AUC	F1	F1-BT	F1-GM	F1-LX
OsteoNet	Emb only	0,712	0,487	0,571	0,404	0,484
	Emb + Demo	0,728	0,470	0,552	0,360	0,500
	Emb + Morph	0,731	0,502	0,602	0,421	0,484
	Emb + Demo+Morph	0,761	0,501	0,588	0,422	0,492
FCoTNet	Emb only	0,654	0,423	0,473	0,382	0,414
	Emb + Demo	0,703	0,510	0,568	0,463	0,500
	Emb + Morph	0,674	0,505	0,541	0,500	0,475
	Emb + Demo+Morph	0,714	0,517	0,578	0,473	0,500

Bảng 7 trình bày kết quả DeLong test cho các so sánh AUC quan trọng nhất. Đóng góp của thông tin lâm sàng được kiểm chứng có ý nghĩa thống kê chắc chắn với cả 2 kiến trúc: OsteoNet Emb+Demo+Morph cải thiện $\Delta AUC = +0,049$ so với Emb only với $p = 0,0002$; FCoTNet đạt $\Delta AUC = +0,061$ với $p = 0,009$. Khi xét riêng từng nhóm, Demo đóng góp có ý nghĩa thống kê với cả hai kiến trúc ($p \leq 0,032$), trong khi Morph

Bảng 7. Kiểm định DeLong test cho các so sánh AUC chính ($\alpha = 0,05$)

Kiến trúc	So sánh	ΔAUC	z	p	Kết luận
OsteoNet	Emb + Demo + Morph vs Demo + Morph	+0,056	1,959	0,050	Cận biên
	Emb + Demo + Morph vs Emb	+0,049	3,777	0,0002	Có ý nghĩa
	Emb + Demo vs Emb	+0,016	2,146	0,032	Có ý nghĩa
	Emb + Morph vs Emb	+0,019	1,804	0,071	Không
	Emb vs Demo + Morph	+0,007	0,193	0,847	Không
FCoTNet	Emb + Demo + Morph vs Demo + Morph	+0,003	0,103	0,918	Không
	Emb + Demo + Morph vs Emb	+0,061	2,631	0,009	Có ý nghĩa
	Emb + Demo vs Emb	+0,049	2,404	0,016	Có ý nghĩa
	Emb + Morph vs Emb	+0,020	1,054	0,292	Không
	Emb vs Demo + Morph	-0,058	-1,548	0,122	Không

3.4. Thảo luận

OsteoNet vượt trội hơn FCoTNet nhờ sử dụng hai bộ CBAM riêng biệt cho góc AP và LAT, cho phép mạng học được

trung gian. Với OsteoNet, thêm Morph cải thiện F1-BT mạnh nhất (0,571 $\rightarrow$ 0,602). Với FCoTNet, thêm Demo cải thiện F1-GM mạnh nhất (0,382 $\rightarrow$ 0,463) vì tuổi và BMI giúp phân biệt giai đoạn trung gian.

khi đứng một mình không vượt ngưỡng ý nghĩa ($p \geq 0,071$).

Để đánh giá đóng góp riêng của ảnh X-Ray khi đã có thông tin lâm sàng, OsteoNet đạt $\Delta AUC = +0,056$ ($p = 0,050$) sát ngưỡng ý nghĩa thống kê, gợi ý có đóng góp nhưng cỡ mẫu chưa đủ khẳng định chắc chắn. FCoTNet không cho thấy đóng góp ($\Delta AUC = +0,003$, $p = 0,918$). Ảnh X-Ray chỉ có giá trị bổ sung khi kiến trúc khai thác tốt cả hai góc AP và LAT.

các đặc trưng bổ sung từ mỗi góc chụp, tương đồng với các dấu hiệu hình thái mà bác sĩ thường dùng để đánh giá loãng xương trên ảnh X-Ray. Đóng góp đặc thù

của nghiên cứu nằm ở việc phân tích định lượng có hệ thống ba mức độ rõ rĩ nhãn DXA, kết hợp với phép thử baseline không sử dụng ảnh được kiểm định thống kê bằng DeLong test, qua đó cho phép tách bạch rõ ràng vai trò của ảnh X-Ray và của thông tin lâm sàng trong kết quả phân loại.

Nghiên cứu này còn tồn tại một số hạn chế: (i) bộ dữ liệu LUMOS là dữ liệu đơn trung tâm, chỉ thu thập trên người Châu Á, nên khả năng khái quát hóa sang các quần thể khác chưa được kiểm chứng; (ii) quy trình hai bước có thể chưa tối ưu bằng huấn luyện end-to-end; (iii) tập kiểm tra chỉ gồm 120 mẫu khiến đóng góp của ảnh X-Ray chỉ đạt mức ý nghĩa cận biên ($p = 0,050$), cần tập dữ liệu lớn hơn (≥ 300 mẫu) để xác nhận; (iv) nhóm đặc trưng Morph hiện được đo thủ công, do đó cần phát triển một module phân đoạn (segmentation) tự động để có thể triển khai trong thực tế.

So với các nghiên cứu trước, điểm chính của bài báo này không phải là khẳng định đạt AUC cao nhất, vì các công trình khác nhau về bộ dữ liệu, vị trí chụp, tiêu chí gán nhãn và số lớp phân loại. Đóng góp nổi bật của nghiên cứu nằm ở việc chỉ ra và định lượng hiện tượng rõ rĩ nhãn từ các biến DXA, bổ sung phép thử baseline không sử dụng ảnh, đồng thời áp dụng kiểm định DeLong để đánh giá đóng góp thực sự của ảnh X-Ray khi kết hợp với thông tin lâm sàng không chứa DXA.

IV. Kết luận

Nghiên cứu đã đề xuất và đánh giá một quy trình phát hiện loãng xương hai bước, không yêu cầu máy DXA, dựa trên ảnh X-Ray cột sống kết hợp với thông tin lâm sàng. Từ kết quả thực nghiệm, có thể rút ra bốn kết luận chính:

(1) Nhóm đặc trưng DXA làm sai lệch nghiêm trọng kết quả: bất kỳ bộ thông tin nào chứa BMD hoặc T-score đều cho $AUC = 1,000$, không phải do mô hình có hiệu năng vượt trội mà do đang đọc lại chính dữ liệu dùng để định nghĩa nhãn.

(2) Thông tin lâm sàng đóng vai trò quan trọng và được kiểm định có ý nghĩa thống kê: cấu hình Demo+Morph cải thiện rõ rệt so với chỉ embedding ảnh.

(3) Đóng góp của ảnh X-Ray phụ thuộc kiến trúc: OsteoNet đạt $\Delta AUC = +0,056$ ($p = 0,050$) sát ngưỡng ý nghĩa, gợi ý có đóng góp nhưng cần xác nhận với tập kiểm tra lớn hơn. FCoTNet không cho thấy đóng góp ($p = 0,918$) do cơ chế chia sẻ backbone giữa hai góc chụp.

(4) OsteoNet với CBAM riêng cho hai góc chụp là kiến trúc phù hợp nhất, đạt $AUC = 0,761$ với Demo+Morph, không sử dụng thông tin DXA và được kiểm định bằng DeLong test.

Các hướng nghiên cứu bao gồm: (i) huấn luyện end-to-end kết hợp đồng thời ảnh và thông tin lâm sàng; (ii) kiểm nghiệm trên dữ liệu đa trung tâm với tập kiểm tra từ 300 mẫu trở lên để xác nhận đóng góp của ảnh X-Ray; (iii) tích hợp các phương pháp giải thích mô hình tự động như Grad-CAM (Gradient-weighted Class Activation Mapping) và SHAP (SHapley Additive exPlanations); (iv) phát triển module phân đoạn (segmentation) đốt sống tự động nhằm giảm sự phụ thuộc vào việc gán nhãn thủ công cho đặc trưng hình thái đốt sống (Morph).

Tài liệu tham khảo

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework.

- Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA. <https://doi.org/10.1145/3292500.3330701>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA. <https://doi.org/10.1145/2939672.2939785>
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44 3, 837-845.
- Genant, H. K., Wu, C. Y., van Kuijk, C., & Nevitt, M. C. (1993). Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research*, 8(9), 1137-1148. <https://doi.org/10.1002/jbmr.5650080915>
- Hsieh, C.-I., Zheng, K., Lin, C., Mei, L., Lu, L., Li, W., Chen, F.-P., Wang, Y., Zhou, X., Wang, F., Xie, G., Xiao, J., Miao, S., & Kuo, C.-F. (2021). Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-021-25779-x>
- Johnell, O., & Kanis, J. A. (2006). An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporosis International*, 17(12), 1726-1733. <https://doi.org/10.1007/s00198-006-0172-4>
- Kanis, J. A., Cooper, C., Rizzoli, R., Reginster, J. Y., & Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO), & Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). (2019). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporosis International*, 30(1), 3-44. <https://doi.org/10.1007/s00198-018-4704-5>
- Muzaffar, A., Riaz, F., & Tahir, M. (2025). OsteoNet-A Framework for Identifying Osteoporosis in Bone Radiograph Images Using Attention-Based VGG Network. *IEEE Access*, PP, 1-1. <https://doi.org/10.1109/ACCESS.2025.3538828>
- Shi, K., Shen, Q., Ye, Z., Jiang, L., Bu, J., & Wang, H. (2025). LUMOS: A Lumbar Multimodal Osteoporosis Screening Dataset with X-ray and CT images. *Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland. <https://doi.org/10.1145/3746027.3758282>
- World Health Organization. (2003). *Prevention and management of osteoporosis*. (WHO Technical Report Series No. 921).
- Xue, L., Qin, G., Chang, S., Luo, C., Hou, Y., Xia, Z., Yuan, J., Wang, Y., Liu, S., Liu, K., Li, X., Wu, S., Zhao, Q., Gao, W., & Yang, K. (2023). Osteoporosis prediction in lumbar spine X-ray images using the multi-scale weighted fusion contextual transformer network. *Artificial Intelligence in Medicine*, 143, 102639. <https://doi.org/10.1016/j.artmed.2023.102639>
- Yamamoto, N., Sukegawa, S., Kitamura, A., Goto, R., Noda, T., Nakano, K., Takabatake, K., Kawai, H., Nagatsuka, H., Kawasaki, K., Furuki, Y., & Ozaki, T. (2020). Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates. *Biomolecules*, 10(11), 1534.

OSTEOPOROSIS DETECTION BASED ON LUMBAR SPINE X-RAY USING DEEP LEARNING WITH CLINICAL FEATURES

Vu Xuan Hanh¹, Vu Thien Vinh², Do Duy Trinh¹, Ngo Van Son¹

Abstract: Osteoporosis increases fracture risk and is commonly diagnosed using DXA, which is costly and not widely available in many lower-level healthcare facilities. This study proposes a two-stage osteoporosis screening pipeline that does not use DXA-derived variables as input, combining lumbar spine X-ray images with clinical features. The pipeline was evaluated on the X-ray subset of the LUMOS dataset, including 803 patients and 1,620 images. In the first stage, OsteoNet and FCoTNet were used to extract 128-dimensional embedding vectors from X-ray images. In the second stage, these embeddings were combined with clinical feature groups and fed into an XGBoost classifier for three-class classification: Normal, Osteopenia, and Osteoporosis. The results show that configurations containing DXA-derived variables, such as BMD or T-score, achieved an AUC of 1.000 because the class labels were directly defined by T-score, indicating label leakage and rendering such results clinically meaningless. After excluding DXA-derived variables, the clinical-only baseline achieved an AUC of 0.709, while the best configuration, OsteoNet + Demo+Morph, reached an AUC of 0.761. DeLong testing showed a statistically significant improvement over the embedding-only configuration ($p = 0.0002$) and a borderline improvement over the clinical-only baseline ($p = 0.050$). These findings suggest that lumbar spine X-ray images can provide complementary value for osteoporosis screening when combined with appropriate non-DXA clinical information.

Keywords: Osteoporosis, deep learning, lumbar spine X-ray, embedding vector, XGBoost, LUMOS dataset, DeLong test

¹ Hanoi Open University, Hanoi, Vietnam

² Hanoi Hospital of Traditional Medical, Hanoi, Vietnam