

Large Language Model-based Textualization of Gradient-weighted Class Activation Mapping Explanations for Osteoporosis Prediction from Sensor-acquired Hip Radiographs

Shang-Wen Feng,¹ Hao-Chun Chiang,² Yu-Sheng Huang,³
Chen-Ni Hu,³ Yi-Hung Chiang,¹ and Szu-Yin Lin^{3*}

¹Department of Orthopedics, National Yang Ming Chiao Tung University Hospital,
No. 169, Xiaoshe Rd., Yilan City, Yilan County 260, Taiwan

²Department of Computer Science and Information Engineering, National Ilan University,
No. 1, Sec. 1, Shennong Rd., Yilan City, Yilan County 260, Taiwan

³Department of Management Science, National Yang Ming Chiao Tung University,
No. 1001, Daxue Rd., East Dist., Hsinchu City 300, Taiwan

(Received June 4, 2026; accepted October 13, 2026)

Keywords: osteoporosis prediction, sensor-acquired hip radiographs, explainableAI, Grad-CAM, large language models, prompt engineering

Osteoporosis screening from hip radiographs requires both predictive performance and interpretable model outputs. In this study, we developed a three-component computational framework that separates binary image classification from *post hoc* visual and textual explanation. Four ImageNet-pretrained models, namely, VGG16, ResNet50, InceptionV3, and ViT-B16, were evaluated using patient-level fivefold cross-validation. VGG16 achieved the highest observed aggregate accuracy of 0.80 and was selected for subsequent explainability analysis. Gradient-weighted class activation mapping (Grad-CAM) was then used to localize image regions associated with the classifier output, and ChatGPT-4o converted the activation patterns and level-specific contextual information into structured textual explanations under five prompt-engineering conditions. The five prompt-engineering conditions progressively incorporated symbolic representation, anatomical context, structured reasoning, and critical clinical review. In the representative case, Grad-CAM activation was mainly observed around the femoral neck, proximal femur, and adjacent cortical regions. The generated outputs illustrate how prompt design can alter the detail, organization, and clinical framing of textualized explainable artificial intelligence (XAI) explanations. These findings demonstrate the technical feasibility of using a large language model (LLM) as a *post hoc* textualization layer for sensor-acquired hip radiographs, although further multicase and clinician-based validation is required.

*Corresponding author: e-mail: stan@nycu.edu.tw
<https://doi.org/10.18494/SAM6463>

1. Introduction

With the growing diversity of diseases and the increasing level of automation in healthcare, AI has been progressively integrated into various medical applications, including targeted therapy, personalized medicine, and clinical diagnosis.⁽¹⁾ Among these applications, medical imaging has experienced particularly rapid advancement because X-ray, ultrasound, magnetic resonance imaging, and computed tomography provide rich visual information for diagnostic assessment. The incorporation of AI and deep learning into imaging workflows has promoted more precise and personalized healthcare, especially when complex visual patterns must be interpreted efficiently.⁽²⁾

As AI systems become more involved in clinical decision support, their usefulness depends not only on predictive performance but also on whether their outputs can be understood and trusted by medical professionals. In healthcare, black-box systems often face barriers to adoption because clinicians must be able to interpret model outputs and judge whether they are appropriate for diagnosis and treatment planning. For this reason, explainable artificial intelligence (XAI) has become an important direction in healthcare AI, as it improves the transparency of model decisions and helps clinicians better understand the basis of AI predictions.⁽³⁾

This requirement is especially important for osteoporosis assessment using hip radiographs. Dual-energy X-ray absorptiometry (DXA) remains the clinical standard for bone mineral density measurement,⁽⁴⁾ while hip radiographs are widely available and have been explored as inputs for deep-learning-based osteoporosis prediction.⁽⁵⁾ X-ray-based assessment can still be affected by incomplete clinical information, reader variability, or diagnostic bias.⁽⁶⁾ Therefore, the intelligent processing of radiographic image data requires not only classification accuracy but also interpretable evidence that can be reviewed by clinicians.

Visual XAI methods such as class activation mapping (CAM) and gradient-weighted class activation mapping (Grad-CAM) can highlight image regions that contribute to model predictions.^(7,8) Nevertheless, heatmaps alone may still be difficult for clinicians to interpret, particularly when activations appear near overlapping anatomy, cortical margins, or background structures. DXA-based *T-scores* indicate an individual's bone mineral density compared with that of a healthy young reference population and remain central to osteoporosis classification.^(9,10) Prior work has shown that deep learning can predict osteoporosis from simple hip radiographs, supporting the feasibility of using X-ray images as accessible screening material.⁽⁵⁾

In this study, the candidate computational intelligence models included VGG, ResNet, Inception, and Vision Transformer architectures, representing both established convolutional designs and transformer-based image classification approaches.^(11–14) In recent studies, language models have begun to be combined with visual explanation methods to produce more accessible explanations for computer vision models.⁽¹⁵⁾ Prompt engineering has also been proposed as a practical way to control the structure and reasoning depth of generated medical explanations.^(16,17) In our previous work, our group developed deep learning models for *T-score*-based osteoporosis classification using whole-hip and anatomical-region hip radiographs.⁽¹⁸⁾ A subsequent sensor-based framework combined Grad-CAM, Shapley additive explanations (SHAP), and local interpretable model-agnostic explanations (LIME) with orthopedic-surgeon evaluation to assess

visual interpretability and clinical trust.⁽¹⁹⁾ However, these studies focused on image classification and visual XAI rather than the conversion of XAI outputs into structured clinical text. Therefore, the value of using large language models (LLMs) to textualize XAI outputs for osteoporosis prediction from hip radiographs remains insufficiently examined.

In this study, we treat hip radiographs as X-ray-derived biomedical images generated using a conventional X-ray imaging system, in which sensor-acquired anatomical signals are transformed into clinically meaningful visual representations. On this perspective, we develop a three-component computational workflow comprising deep-learning-based osteoporosis classification, Grad-CAM-based visual explanation, and LLM-based textual explanation. In this workflow, computational intelligence methods are used to convert sensor-derived visual signals into structured explanatory outputs intended for subsequent clinical review, linking noninvasive biomedical sensing with image-based explanation. The five prompt-engineering conditions were designed to examine how progressively adding symbolic, anatomical, contextual, and reasoning constraints changes the structure and clinical framing of LLM-generated explanations. By connecting sensor-derived image classification, visual XAI, and language-based explanation, we present a proof-of-concept method for making model outputs more accessible for subsequent clinical review.

2. Data, Materials, and Methods

2.1 Research framework

This study was organized into two phases. In Phase I, hip radiographs and DXA-derived *T-scores* were used to train and compare deep learning models for binary osteoporosis classification. ImageNet-pretrained models, including VGG16, ResNet50, InceptionV3, and ViT-B16, were evaluated, and the model with the highest observed aggregate accuracy was selected for subsequent explainability analysis. In Phase II, Grad-CAM was applied to the selected model to identify regions associated with the classifier output. Five prompt-engineering strategies, designated L1–L5, were then used to convert the visual explanation into structured text. These strategies progressively incorporated basic heatmap description, symbolic representation, patient and anatomical context, structured reasoning, and critical clinical review. The generated explanations were examined qualitatively to illustrate how prompt design affected explanatory detail, organization, anatomical grounding, and cautionary content.

2.2 Data and materials

In this retrospective image-based study, we used an expanded institutional cohort of 477 hip radiographs from 467 unique patients collected over a five-year period at a regional hospital in Yilan, Taiwan. The present cohort includes all 359 radiographs from the 359 patients reported in our previous sensor-based XAI study.⁽¹⁹⁾ That cohort, in turn, included the 134 patients analyzed in our earlier hip-radiograph segmentation and osteoporosis-classification study,⁽¹⁸⁾ which contained 139 radiographs from 134 patients. Accordingly, the present dataset represents an

expanded and reanalyzed institutional cohort rather than an independent cohort. Compared with the 359-image cohort, the current cohort was expanded by 118 radiographs and 108 unique patients. We applied a distinct preprocessing and evaluation pipeline and addressed a different research objective, namely, the LLM-based textualization of Grad-CAM explanations, in the present study. Previously published segmentation, consensus-region retraining, and clinician-survey results were not reused in the present analysis. The retrospective use of radiographic and associated clinical data was approved by the Institutional Review Board of National Yang Ming Chiao Tung University Hospital (IRB No. 2022A007). Among the 467 patients, 457 contributed unilateral radiographs (97 males and 360 females) and 10 contributed bilateral radiographs (2 males and 8 females). The mean patient age was 74 years.

DXA-derived *T-scores* served as the reference standard. The image-classification task was binary: non-osteoporotic ($T\text{-score} > -2.5$) versus osteoporotic ($T\text{-score} \leq -2.5$). At the image level, 219 radiographs were assigned to the non-osteoporotic group (213 patients), whereas 258 radiographs were assigned to the osteoporotic group (255 patients). One bilaterally imaged patient contributed one hip to each diagnostic group; therefore, the group-level patient counts overlap by one, while the total number of unique patients remains 467. Exact patient and image counts within individual validation folds could not be reliably reconstructed retrospectively, and only aggregate validation performance is reported. The normal, osteopenia/low-bone-mass, and osteoporosis categories included in L3–L5 were used solely as a clinical background for textual explanation. They were not separate prediction targets and did not alter the classifier output. The clinical classification criteria are summarized in Table 1. The cohort and image-level allocation are summarized in Fig. 1.

2.3 Phase I: model development and selection

Hip radiographs were employed as the primary dataset to train and evaluate multiple deep learning models for osteoporosis classification. Model training and analysis were implemented in Python using TensorFlow and Keras on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU. Four ImageNet-initialized pretrained models were adopted to leverage transfer learning: VGG16, ResNet50, InceptionV3, and ViT-B16.

Whole-hip radiographs were resized directly to 600×600 pixels for the convolutional neural network (CNN)-based models and 592×592 pixels for ViT-B16. Pixel values were rescaled by a factor of $1/255$. The originally grayscale radiographs were loaded in the RGB mode and represented as three-channel inputs. No additional cropping, left–right orientation harmonization, geometric or intensity-based data augmentation, class weighting, oversampling,

Table 1
T-score classification criteria.

Level of osteoporosis	Classification criterion
Normal	$T\text{-score} \geq -1.0$
Osteopenia (low bone mass)	$-2.5 < T\text{-score} < -1.0$
Osteoporosis	$T\text{-score} \leq -2.5$
Severe osteoporosis	$T\text{-score} \leq -2.5$ with one or more fragility fractures

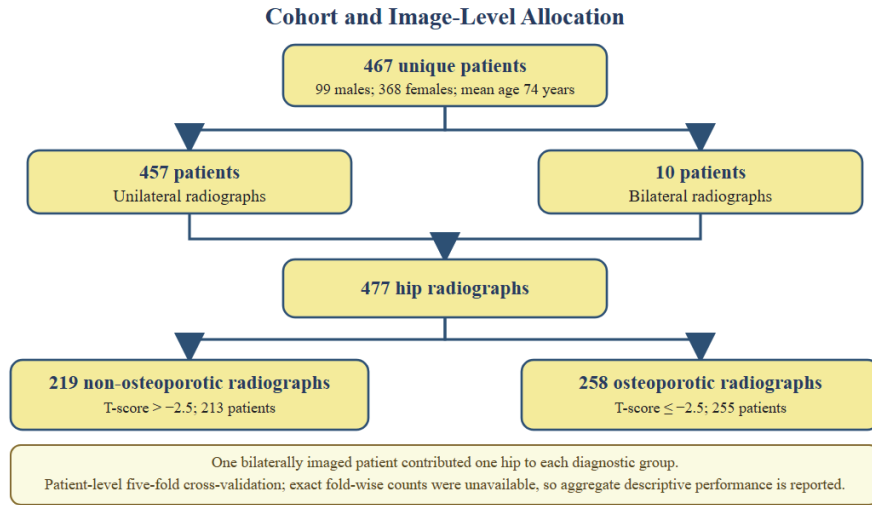


Fig. 1. (Color online) Data flow of the retrospective cohort. The dataset included 467 unique patients and 477 hip radiographs: 457 patients contributed unilateral radiographs and 10 contributed bilateral radiographs. At the image level, 219 radiographs were classified as non-osteoporotic ($T\text{-score} > -2.5$; 213 patients) and 258 as osteoporotic ($T\text{-score} \leq -2.5$; 255 patients). One bilaterally imaged patient contributed one hip to each diagnostic group. Patient-level fivefold cross-validation assigned all radiographs from the same patient to the same fold.

or undersampling was applied. All models were trained using the Adam optimizer, a learning rate of 1×10^{-5} , a batch size of 16, and 150 epochs. In VGG16, all backbone layers except the final two were frozen. In ResNet50 and InceptionV3, the final 88 and 155 backbone layers were trainable, respectively. In ViT-B16, the first 12 layers were frozen and the remaining layers were trainable. The CNN classification head consisted of batch normalization, global average pooling, a 256-unit dense layer with rectified linear unit (ReLU) activation, and a two-unit softmax output. The ViT-B16 classification head consisted of batch normalization and a two-unit softmax output. Exact trainable-parameter counts and architecture-specific boundary-layer names could not be reliably reconstructed because the serialized model summaries were unavailable. The preprocessing and model-configuration settings are summarized in Table 2.

Model performance was assessed using confusion-matrix-based measures, including accuracy, sensitivity, specificity, and $F1\text{-score}$. True positive (TP) denoted correctly identified osteoporotic cases, false positive (FP) denoted non-osteoporotic cases incorrectly classified as osteoporotic, false negative (FN) denoted osteoporotic cases incorrectly classified as normal, and true negative (TN) denoted correctly identified non-osteoporotic cases. The evaluation metrics were calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

Table 2
Preprocessing and model-configuration settings.

Parameter	Reported setting
Input size, CNN-based models	600 × 600 pixels
Input size, ViT-B16	592 × 592 pixels
Pixel scaling	1/255
Input representation	Three-channel RGB
Additional cropping	None
Left–right orientation harmonization	None
Geometric or intensity-based augmentation	None
Class-imbalance handling	No class weighting, oversampling, or undersampling
Batch size	16
No. of training epochs	150
Learning rate	1×10^{-5}
Optimizer	Adam
Loss function	Binary cross-entropy
Trainable backbone portion, VGG16	Final two backbone layers trainable; all earlier backbone layers frozen
Trainable backbone portion, ResNet50	Final 88 backbone layers trainable
Trainable backbone portion, InceptionV3	Final 155 backbone layers trainable
Trainable backbone portion, ViT-B16	First 12 layers frozen; remaining layers trainable
CNN classification head	Batch normalization → Global average pooling → Dense (256, ReLU) → Dense (2, Softmax)
ViT-B16 classification head	Batch normalization → Dense (2, Softmax)
Exact trainable-parameter counts	Unavailable; not used as a basis for comparison

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \quad (4)$$

2.4 Phase II: Visual and textual explainability

In Phase II, an explainability-driven analytical framework was developed to investigate both the visual and textual interpretability of the osteoporosis classification model. This phase consisted of two sequential components: visual explainability analysis using Grad-CAM and text-based explainability analysis facilitated by LLM-based prompt-engineering strategies. The objective was to examine how visual cues extracted from deep learning models can be translated into structured textual explanations containing varying levels of anatomical and clinical context and reasoning constraints.

Grad-CAM was applied to the selected VGG16-based model. The target feature tensor was retrieved programmatically from the fifth layer from the end of the loaded model, corresponding to the convolutional-backbone output in the implemented architecture. The activation map was

passed through ReLU, normalized by its maximum value, resized to the original image dimensions using OpenCV bilinear interpolation, converted using the JET colormap, and superimposed on the radiograph with a heatmap weight of 0.4. A grayscale threshold of 85 was applied only to the derived binary mask used for activation-area analysis; it did not suppress low activations in the displayed continuous heatmap. Grad-CAM was generated for the reference class associated with each evaluated image.

A representative case was purposively selected from the Grad-CAM candidate outputs for comparison across the five prompt conditions. The archived analysis pipeline calculated the ratio of thresholded activation area to background area as a supplementary ranking measure. Because a fully prespecified final case-selection protocol could not be retrospectively verified, potential case-selection bias is acknowledged as a limitation.

For every prompt condition, the LLM received the classifier output, Grad-CAM visualization, and the level-specific textual context summarized in Table 3. The LLM received the Grad-CAM overlay, which retained the radiographic background, but did not receive a separate unoverlaid radiograph. The original radiograph is presented together with the Grad-CAM visualization in Fig. 2 to provide anatomical reference for readers.

The five conditions progressively added contextual and reasoning constraints. L1: Zero-shot requested a basic heatmap description. L2: Chain-of-symbol added prompt-defined color symbols. These symbolic risk labels described Grad-CAM color intensity only and were not calibrated clinical-risk estimates. L3: Context-enriched added patient, anatomical, and *T-score* context. L4: Structured reasoning prompt required a sequential connection among activation, anatomical location, and diagnostic implication. L5: Critical clinical review prompt added cautions, possible confounders, and clinical recommendations. Table 3 shows the inputs and outputs, and Supplementary Table S1 provides the complete prompt templates.

All five prompt strategies were applied to the same representative Grad-CAM case. L3–L5 used the same demographic information: an 80-year-old female patient with a body mass index

Table 3
Inputs and expected outputs of the five prompt-engineering conditions.

Level	Grad-CAM	Class output	Patient data	Anatomy/ <i>T-score</i>	Additional instruction	Expected output
L1	Yes	Yes	No	No	None	Basic heatmap description
L2	Yes	Yes	No	No	Prompt-defined color symbols	Symbolic structured description
L3	Yes	Yes	Yes	Yes	Integrated context	Anatomically grounded description
L4	Yes	Yes	Yes	Yes	Structured sequence	Activation-anatomy-diagnosis reasoning
L5	Yes	Yes	Yes	Yes	Critical review	Reasoning, cautions, and recommendations

Note: The unoverlaid radiograph was not treated as an independent LLM input. It is presented together with the Grad-CAM visualization in Fig. 2 to provide readers with anatomical context. The normal, osteopenia/low-bone-mass, and osteoporosis categories in L3–L5 served only as clinical contextual information and were not classifier outputs.

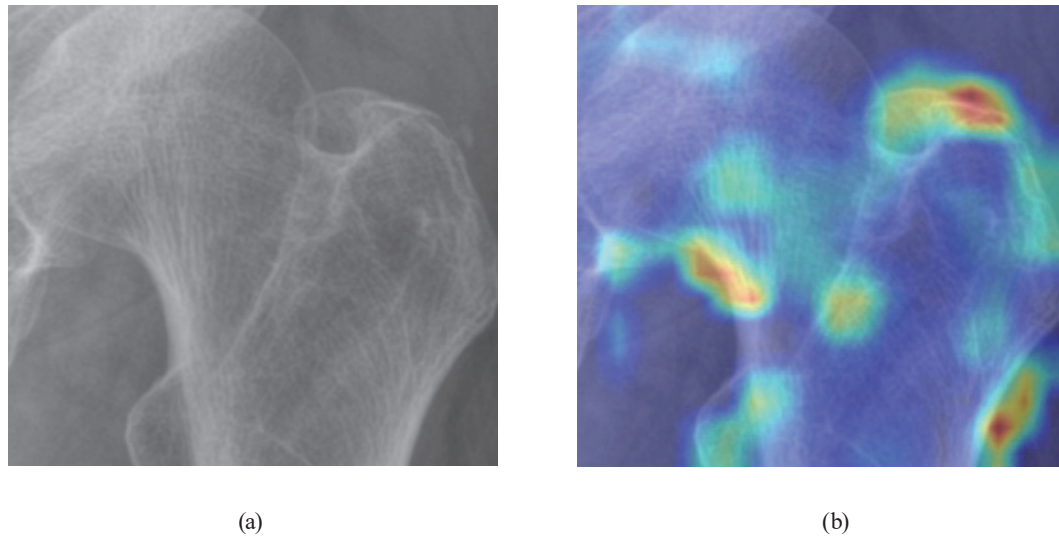


Fig. 2. (Color online) Paired visualization of the representative case used to generate the textual explanation. VGG16 classified the bone in the radiograph as osteoporotic, and the corresponding Grad-CAM heatmap was generated with the osteoporotic class as the target. The original radiograph provides anatomical context, whereas the heatmap highlights image regions relevant to the model's classification: (a) original hip radiograph and (b) Grad-CAM heatmap.

(*BMI*) of 21.1. The classifier output was osteoporotic under every condition. Textual explanations were generated using ChatGPT-4o through the web-based interface during the original study period, and one archived explanation from each condition was available for illustrative qualitative comparison. Because an API was not used, temperature, top-p, random seed, and the exact backend model snapshot were not exposed for direct user specification. The exact access date, session configuration, complete generation history, and generation-level selection or editing metadata were unavailable for retrospective verification and were not inferred.

3. Results and Discussion

3.1 Phase I: Classification model evaluation

Four models, namely, VGG16, ResNet50, InceptionV3, and ViT-B16, were evaluated through patient-level fivefold cross-validation. Table 4 shows the available aggregate descriptive performance values from patient-level fivefold cross-validation. Because a complete and internally consistent set of fold-specific outputs was unavailable, standard deviations and confidence intervals could not be calculated retrospectively. The model with the highest observed aggregate accuracy was selected for subsequent Grad-CAM analysis, with *F1-score*, sensitivity, and specificity considered as secondary descriptive measures. VGG16 had an

Table 4

Aggregate descriptive performance of the classification models.

Model	Accuracy	Loss	<i>F1-score</i>	Sensitivity	Specificity
VGG16	0.80	0.470197	0.81	0.81	0.80
ResNet50	0.78	0.500996	0.80	0.79	0.78
InceptionV3	0.75	0.564584	0.77	0.78	0.72
ViT-B16	0.79	0.546791	0.80	0.79	0.79

aggregate accuracy of 0.80, a loss of 0.470197, an *F1-score* of 0.81, a sensitivity of 0.81, and a specificity of 0.80. Under this operational selection rule, VGG16 was selected for the subsequent explainability analysis. This rule describes the procedure used in the analysis and was not a prespecified statistical selection criterion.

3.2 Phase II: Visual explainability and prompt-based textualization

To facilitate the presentation of Grad-CAM findings, the original hip radiograph and its Grad-CAM visualization are shown together in Fig. 2. The original radiograph provides the anatomical reference, whereas the heatmap shows the model's class-discriminative attention. This paired presentation allows readers to inspect whether high-activation regions overlap with clinically meaningful structures rather than background, image borders, or unrelated soft tissue.

In the purposively selected case, model attention was concentrated around the femoral neck, proximal femur, and adjacent cortical regions. These structures are clinically relevant to osteoporosis assessment, and cortical thinning or reduced radiographic density may provide indirect cues related to bone mineral status. Limited activation was observed in non-bony background areas in this case. This single visualization does not establish a general pattern of model reliance. Grad-CAM should be interpreted as a localization aid rather than a direct diagnostic measurement and does not replace DXA-derived *T-score* assessment.

The generated textual outputs were examined qualitatively to identify differences in descriptive depth, organizational structure, anatomical grounding, and cautionary content. L1: Zero-shot produced a basic description of color intensity and approximate activation location. L2: Chain-of-symbol introduced prompt-defined symbolic labels for different color regions, resulting in a more standardized output format. L3: Context-enriched incorporated demographic, anatomical, and *T-score* information, and connected the activation pattern with named hip regions. L4: Structured reasoning prompt produced an explicitly ordered explanation linking image background, activation color, anatomical location, possible reasons for activation, and diagnostic implication. L5: Critical clinical review prompt generated the broadest output by additionally discussing possible confounders, image artifacts, degenerative changes, model misattribution, and the need for DXA confirmation. Compared with L4, the L5 output contained more cautionary and recommendation-oriented content, and devoted a larger proportion of the explanation to potential confounders and clinical recommendations rather than to the localized

activation pattern. These observations are descriptive and based on a single representative case. They illustrate differences in output structure across prompt conditions but do not establish the clinical superiority of any prompt strategy.

3.3 Limitations

This study has several limitations. First, exact patient and image counts for individual validation folds and a complete, internally consistent set of fold-level outputs could not be retrospectively verified. Performance was therefore reported as aggregate descriptive values without standard deviations or confidence intervals. Exact trainable-parameter counts were also unavailable. Second, the textualization analysis was based on one purposively selected representative case. Although an activation-to-background-area ratio was available as a supplementary ranking measure, a fully prespecified final case-selection protocol could not be retrospectively verified, and case-selection bias cannot be excluded. The observed activation pattern and prompt-related differences therefore cannot be generalized to the full dataset. Third, explanations were generated through the ChatGPT-4o web interface. API-level parameters, including temperature, top-p, random seed, and the exact backend model snapshot, were not available for direct specification. The exact access date, complete generation history, and generation-level selection or editing metadata were also unavailable for retrospective verification. Fourth, the generated explanations were not evaluated by clinicians. Consequently, their clinical accuracy, usability, interpretability, or trustworthiness could not be determined in the present study. The comparison across prompt conditions should therefore be interpreted as an illustrative qualitative analysis rather than clinical validation. Finally, the image cohort was obtained from a single regional hospital, and no external validation dataset was used. Future studies should include multicenter data, multiple prospectively selected cases, reproducible LLM generation settings, and clinician evaluations.

4. Conclusions

We presented a proof-of-concept framework that integrates binary osteoporosis classification from hip radiographs, Grad-CAM visualization, and prompt-engineered LLM textualization. Among the four evaluated models, VGG16 achieved the highest observed aggregate accuracy and was selected under the stated operational rule for subsequent explainability analysis. Grad-CAM and ChatGPT-4o were applied only after classification and did not modify the classifier output. In the representative case, Grad-CAM activation was mainly observed around the femoral neck, proximal femur, and adjacent cortical regions. The five prompt-engineering conditions generated explanations with different levels of descriptive detail, anatomical context, reasoning structure, and cautionary content. Structured prompts produced more explicitly organized explanations, whereas the critical review prompt incorporated a broader range of potential confounders and clinical cautions. These findings demonstrate the technical feasibility of converting visual XAI outputs into structured textual explanations. However, because the analysis was based on a single representative case and did not include clinician evaluations, no

conclusions can be drawn regarding clinical accuracy, usability, interpretability, trustworthiness, or superiority among prompt conditions. Future studies should be aimed at evaluating of multiple prospectively selected cases and conducting clinician-based validation.

Acknowledgments

This research was supported by the National Science and Technology Council, Taiwan, under Grant Nos. 114-2410-H-A49-066-MY3 and 112-2410-H-A49-087-MY2.

References

- 1 A. Haleem, M. Javaid, and I. H. Khan: *Curr. Med. Res. Pract.* **9** (2019) 231. <https://doi.org/10.1016/j.cmrp.2019.11.005>
- 2 A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsafaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K. S. Nikita, B. P. Veasey, M. Zervakis, J. H. Saltz, and C. S. Pattichis: *IEEE J. Biomed. Health Inform.* **24** (2020) 1837. <https://doi.org/10.1109/JBHI.2020.2991043>
- 3 S. Bharati, M. R. H. Mondal, and P. Podder: *IEEE Trans. Artif. Intell.* **5** (2024) 1429. <https://doi.org/10.1109/TAI.2023.3266418>
- 4 G. M. Blake and I. Fogelman: *Postgrad. Med. J.* **83** (2007) 509. <https://doi.org/10.1136/pgmj.2007.057505>
- 5 R. Jang, J. H. Choi, N. Kim, J. S. Chang, P. W. Yoon, and C.-H. Kim: *Sci. Rep.* **11** (2021) 19997. <https://doi.org/10.1038/s41598-021-99549-6>
- 6 L. Zhang, X. Wen, J.-W. Li, X. Jiang, X.-F. Yang, and M. Li: *Insights Imaging* **14** (2023) 163. <https://doi.org/10.1186/s13244-023-01521-7>
- 7 B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba: *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016)* 2921–2929. <https://doi.org/10.1109/CVPR.2016.319>
- 8 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra: *Proc. 2017 IEEE Int. Conf. Computer Vision (IEEE, 2017)* 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- 9 H. P. Dimai: *Bone* **104** (2017) 39. <https://doi.org/10.1016/j.bone.2016.12.016>
- 10 S. H. Guelndner, T. N. Grabo, E. D. Newman, and D. R. Cooper, Eds.: *Osteoporosis: Clinical Guidelines for Prevention, Diagnosis, and Management* (Springer Publishing Company, New York, 2007) Chap.7.
- 11 K. Simonyan and A. Zisserman: *arXiv:1409.1556* (2014). <https://arxiv.org/abs/1409.1556>
- 12 K. He, X. Zhang, S. Ren, and J. Sun: *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016)* 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- 13 C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich: *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2015)* 1–9. <https://doi.org/10.1109/CVPR.2015.7298594>
- 14 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby: *Proc. 9th Int. Conf. Learning Representations (ICLR, 2021)*. <https://openreview.net/forum?id=YicbFdNTTy>
- 15 H. Nguyen, T. Clement, L. Nguyen, N. Kemmerzell, B. Truong, K. Nguyen, M. Abdelaal, and H. Cao: *Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI Organization, 2024)* 8754–8758. <https://doi.org/10.24963/ijcai.2024/1025>
- 16 P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha: *arXiv:2402.07927* (2024). <https://arxiv.org/abs/2402.07927>
- 17 T. F. Heston and C. Khun: *Int. Med. Educ.* **2** (2023) 198. <https://doi.org/10.3390/ime2030019>
- 18 S.-W. Feng, S.-Y. Lin, Y.-H. Chiang, M.-H. Lu, and Y.-H. Chao: *Appl. Sci.* **14** (2024) 133. <https://doi.org/10.3390/appl4010133>
- 19 S.-W. Feng, S.-Y. Lin, Y.-T. Shen, B.-Y. Chen, Y.-C. Chen, and Y.-H. Chiang: *Sens. Mater.* **37** (2025) 3803. <https://doi.org/10.18494/SAM5723>

Supplementary Material

Table S1
Complete prompt templates used for the five prompt-engineering conditions.

Prompt condition	Complete prompt template
L1: Zero-shot	This is an X-ray image classified as [Class Result] by [Model]. The image has been visualized by the Grad-CAM method. Please describe the Grad-CAM result of this hip X-ray image.
L2: Chain-of-symbol	This is an X-ray image classified as [Class Result] by [Model]. The image has been visualized by the Grad-CAM method. Please describe the Grad-CAM result of this hip X-ray image using concise symbolic notation: red region: high risk (R); yellow region: medium risk (Y); blue/green region: low risk (B/G).
L3: Context-enriched	This X-ray image of [a male or female patient aged x years with a BMI of x] was classified as [Class Result] by [Model] and visualized by the Grad-CAM method. The hip regions include the femoral neck, total hip, Ward's triangle, and greater trochanter. The T -score classification criteria are $T\text{-score} \geq -1.0$: normal; $-2.5 < T\text{-score} < -1.0$: osteopenia/low bone mass; $T\text{-score} \leq -2.5$: osteoporosis. Please first describe the locations of the hip regions in the image and then provide an integrated description of the Grad-CAM result.
L4: Structured reasoning prompt	This X-ray image of [a male or female patient aged x years with a BMI of x] was classified as [Class Result] by [Model] and visualized by the Grad-CAM method. Please analyze the Grad-CAM result in a structured sequence: (1) describe the background; (2) identify the color regions and corresponding hip locations; (3) explain why these regions may be highlighted; and (4) describe the diagnostic implications in relation to the T -score classification criteria.
L5: Critical clinical review prompt	This X-ray image with [clinical information] was classified as [Class Result] by [Model] and visualized by the Grad-CAM method. Please conduct a critical clinical review covering region description, reasons for highlighting, diagnostic implications, clinical recommendations, and possible abnormal situations such as artifacts, degenerative lesions, or model misattribution.

Note: Square-bracketed expressions indicate case-specific placeholders. The templates are reproduced verbatim from the source study materials, including the phrase “hip X-ray image” used in the original prompts. The risk labels used in L2 were prompt-defined symbolic descriptors of Grad-CAM color intensity and should not be interpreted as calibrated clinical-risk estimates.