

Multimodal deep learning enables precise early-versus-late ARCO stratification at the critical threshold for hip preservation in osteonecrosis of the femoral head

Siyuan Li

Guangzhou University of Chinese Medicine

Huimin Huang

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Yetong He

Guangzhou University of Chinese Medicine

Xiaowan Zhang

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Xiaoxia Li

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Lin Wang

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Wei He

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Xiaorui Han

Guangzhou First People's Hospital

Xudong Huang

The Third Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine

Gengxin Chen

389207217@qq.com

Guangdong Provincial Hospital of Traditional Chinese Medicine

Research Article

Keywords: Osteonecrosis of the femoral head, Early-stage and late-stage disease, Multimodal deep learning, Early fusion, Hip-preserving surgery

Posted Date: July 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10138595/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Musculoskeletal Disorders on October 13th, 2026. See the published version at <https://doi.org/10.1186/s12891-026-10544-2>.

Abstract

Objective

Precise early- versus late-stage stratification of osteonecrosis of the femoral head (ONFH) is critical for hip-preserving treatment selection, yet conventional MRI interpretation suffers from interobserver variability and limited objectivity. This study aimed to develop a multimodal deep learning model integrating biparametric MRI (STIR and T1-FSE) with clinical parameters (including age, sex, affected side (left, right, or bilateral), and etiology,) to differentiate early- from late-stage ONFH and provide objective imaging evidence.

Methods

This retrospective study included 530 patients with ONFH, yielding imaging data from 851 hips. The collected data comprised STIR sequences, T1-FSE sequences, and clinical information, including age, sex, and etiology. The dataset was randomly divided at the patient level into a training set (424 patients, 675 hips) and a test set (106 patients, 176 hips) at an 4:1 ratio. Seven models were constructed: a clinical model, a STIR model, a T1-FSE model, a bpMRI model, a STIR-clinical model, a T1-clinical model, and a bpMRI-clinical model. Early- and late-fusion strategies were used for multimodal integration. The diagnostic performance of the seven models for classifying early- versus late-stage ONFH was systematically compared, with particular emphasis on their ability to identify borderline cases at the transition between clinical stages. Model interpretability was assessed using Gradient-weighted Class Activation Mapping (Grad-CAM).

Results

The early-fusion bpMRI-clinical model achieved the highest AUC of 0.944 (95% CI 0.905–0.973), accuracy 90.9%, sensitivity 89.0%, and specificity 92.9%, significantly outperforming single-sequence models (all $P < 0.05$). The confusion matrix showed that the model misclassified 10 late-stage cases as early-stage and 6 early-stage cases as late-stage. Grad-CAM visualizations indicated that the model primarily focused on clinically relevant regions, including the necrotic lesion boundary, areas of bone marrow edema, and subchondral structural abnormalities. These attention patterns were highly consistent with the core imaging features used for early- versus late-stage ONFH stratification.

Conclusion

This study presents an interpretable multimodal deep learning framework that integrates dual-sequence MRI with clinical parameters. The framework accurately differentiates early- from late-stage ONFH

according to the ARCO classification and provides objective evidence for determining the critical threshold for hip-preservation decision-making.

Introduction

Osteonecrosis of the femoral head (ONFH) is a disabling skeletal disorder that mainly affects young and middle-aged adults. Disease progression may lead to femoral head collapse, secondary hip degeneration, persistent pain, and functional impairment. Because collapse marks a critical turning point after which joint damage is often irreversible and total hip arthroplasty (THA) becomes the main treatment option, accurate early identification is essential [1]. The Association Research Circulation Osseous (ARCO) staging system is widely used to guide treatment, with hip-preserving procedures generally recommended for early-stage disease (ARCO stages 1–2) and THA for late-stage disease (ARCO stages 3–4)[2–4]. However, imaging-based staging near the transition from stage II to stage IIIA remains challenging, as it requires precise recognition of subtle femoral-head morphological changes, subchondral structural abnormalities, and early signs of collapse [4–6]. Misclassification at this boundary may lead either to unnecessary THA or to failed hip-preserving surgery, highlighting the need for more objective, accurate, and generalizable auxiliary diagnostic approaches.

In recent years, deep learning techniques have shown considerable potential in orthopedic imaging analysis. Previous studies have applied deep learning to ONFH detection, staging, and collapse-risk prediction using radiography, CT, and MRI [7, 5, 8]. The ability of deep learning to detect subtle musculoskeletal abnormalities has also been validated in applications such as labral tears [9], cartilage defects [10], multiple myeloma [11], and occult fractures [12, 13], providing a technical foundation for its use in refined discrimination of borderline ONFH stages. Recently, Klontzas et al. trained a convolutional neural network (CNN) model using coronal STIR sequences alone; this model effectively differentiated early- from late-stage ONFH (ARCO stages 1–2 vs. 3–4) and achieved diagnostic performance comparable to that of musculoskeletal radiologists [14]. However, existing studies are predominantly based on single MRI sequences, and systematic evaluation of the complementary value of multi-sequence information remains limited [15]. STIR sequences are highly sensitive to bone marrow edema and fluid components [16], whereas T1-FSE sequences are advantageous for delineating the necrotic boundary, reflecting bone marrow fat replacement, and assessing subchondral bone integrity [17]. Together, STIR and T1-weighted sequences capture complementary aspects of ONFH progression, reflecting lesion activity and structural change, respectively. Khurana et al. showed that adding T1-weighted images to STIR increased sensitivity for detecting femoral-head avascular necrosis from 85% to 97%, while STIR alone achieved 100% specificity [18]. However, the value of integrating these sequences with clinical variables for defining borderline clinical stages and guiding individualized treatment decisions remains largely unexplored.

Against this background, this study aimed to develop and validate a multimodal deep learning model that integrates bpMRI (STIR and T1-FSE) with key clinical parameters for precise stratification of early-versus late-stage ONFH. The study also systematically compared the classification performance of

different model architectures, including single-sequence imaging models, a clinical model, and multimodal fusion models, with particular emphasis on the value of early- and late-fusion strategies.

Materials and Methods

1. Study Population

This single-center retrospective study consecutively collected clinical and imaging data from patients diagnosed with ONFH at the Hospital between January 2018 and December 2023. The inclusion criteria were as follows: (1) a diagnosis of ONFH confirmed by clinical presentation and imaging findings; (2) availability of complete dual-sequence MRI data, including coronal STIR and T1-FSE sequences; and (3) complete clinical data, including etiology, sex, and age. Etiology was classified as steroid-induced, alcohol-induced, idiopathic, or other. The exclusion criteria were as follows: (1) concomitant hip infection, tumor, traumatic fracture, or other hip disorders that could affect femoral-head morphology and disease assessment; (2) prior hip surgery; (3) poor MRI image quality, including severe motion or metallic artifacts that precluded accurate evaluation of the femoral-head region; and (4) availability of only a single MRI sequence.

All cases were staged independently by two experienced musculoskeletal radiologists according to the ARCO staging system, and discrepancies were resolved by consensus. Based on clinical treatment-stratification needs, cases were categorized into an early-stage group (ARCO stages 1–2, 404 hips) and a late-stage group (ARCO stages 3–4, 447 hips). To prevent data leakage, bilateral hip data from the same patient were assigned to the same data subset and were not distributed across the training and test sets. The training set was used for model development and parameter optimization, whereas the independent test set was used for final evaluation of model performance.

This study was approved by the Ethics Committee. The requirement for informed consent was waived by the committee given the retrospective nature of the study.

2. Clinical Data Collection and MRI Acquisition

Clinical data, including age, sex, affected side (left, right, or bilateral), and etiology, were systematically collected from the hospital electronic medical record system. Etiology was categorized as steroid-induced, alcohol-induced, idiopathic, or other.

All enrolled patients underwent standardized hip MRI examinations for subsequent imaging analysis and model construction. MRI examinations were performed on 1.5-T or 3.0-T scanners. Coronal T1-FSE and STIR sequences were acquired for all patients. The T1-FSE sequence parameters were as follows: repetition time (TR), 300–800 ms; echo time (TE), 10–20 ms. The STIR sequence parameters were as follows: TR, 3000–4000 ms; TE, 80–100 ms. Section thickness was 3–4 mm, with an interslice gap of

0.5 mm for all sequences. These dual-sequence MRI images were used as inputs for subsequent deep learning model development and analysis.

3. ARCO Staging and Definition of the Endpoint Event

The ARCO stage for each case was independently determined by two radiologists with more than 10 years of experience in musculoskeletal imaging, in accordance with the 2019 revised ARCO staging system. When the two readers disagreed, a final consensus was reached through discussion; this consensus served as the reference standard for model training and evaluation. ARCO staging was determined primarily using imaging findings from coronal STIR and T1-FSE sequences. In alignment with the clinical stratification objective of this study, ARCO stages 1–2 were defined as the early-stage group and ARCO stages 3–4 as the late-stage group.

4. Image Preprocessing

To prevent data leakage, data partitioning was performed at the patient level, ensuring that bilateral hip data from the same patient were assigned exclusively to the same data subset. Before model training, all MRI images underwent uniform preprocessing. The preprocessing pipeline consisted of the following steps: first, the original images were resampled to an isotropic voxel resolution of $1 \times 1 \times 1 \text{ mm}^3$ to mitigate the effects of varying scanning parameters on input consistency; second, Z-score normalization was applied to signal intensities to reduce inter-individual and inter-scanner signal variability; and finally, gray-scale discretization was performed to further standardize the input data distribution and enhance training stability. The preprocessed images were then used for training and testing the deep learning models.

5. Model Architecture Design

A total of seven models were constructed: a clinical model, a STIR model, a T1-FSE model, a bpMRI model, a STIR-clinical model, a T1-clinical model, and a bpMRI-clinical model. DenseNet-121 pretrained on ImageNet-1K was used as the backbone network for imaging feature extraction[19–21].

Discriminative performance was assessed using the area under the receiver operating characteristic curve (AUC), and the DeLong test was used to compare AUCs between models. All statistical analyses were conducted in Python 3.7; notably, the DeLong test was implemented in Python following the same logic as its R counterpart. Gradient-weighted class activation mapping (Grad-CAM) was used to visualize the imaging regions driving model decisions and to qualitatively verify whether the model focused on clinically relevant pathological features, such as the necrotic lesion, bone marrow edema zone, and subchondral fracture line.

6. Atistical Analysis

All statistical analyses and graphical visualizations were performed in Python 3.7 (Python Software Foundation). The DeLong test for comparing AUCs was implemented using Python code that mirrors the standard R implementation. Continuous variables following a normal distribution were expressed as mean $\pm$ standard deviation, while skewed continuous variables were presented as median (interquartile range). Categorical variables were summarized as frequency (percentage). Comparisons of baseline clinical characteristics (age, sex, and etiology) between the training and test sets, as well as between the early-stage (ARCO 1–2) and late-stage (ARCO 3–4) groups, were performed using the independent samples t-test, Mann-Whitney U test, and chi-squared test, as appropriate for the variable type.

Results

1. Study Cohort and Data Partitioning

This retrospective single-center study included 530 patients with ONFH, yielding 851 hip datasets, each containing dual-sequence MRI (STIR and T1-FSE) and clinical parameters. All cases were staged independently by two senior musculoskeletal radiologists according to the ARCO staging system, with discrepancies resolved by consensus. Of these, 404 hips were classified as early-stage (ARCO stages I–II) and 447 hips as late-stage (ARCO stages III–IV). The dataset was randomly divided at an 8:2 ratio into a training set (424 patients, 675 hips) and an independent test set (106 patients, 176 hips) for model development and final performance evaluation.

2. Comparison of Diagnostic Performance Across Models

We systematically constructed and compared seven models to evaluate the contribution of different data modalities to ONFH staging. The ROC curves of all models on the independent test set are shown in Fig. 1. Table 1 and Table 2 present the AUC and other metrics for each model under the early-fusion and late-fusion strategies, respectively.

Table 1

Diagnostic performance of early fusion models for early-and late-stage ONFH in the test cohort.

Model	AUC (95%CI)	Accuracy	Sensitivity	Specificity	p-value
Clinical model	0.828(0.765–0.884)	0.790	0.769	0.812	0.00013
STIR model	0.886(0.833–0.930)	0.824	0.791	0.859	0.009
T1-FSE model	0.905(0.856–0.947)	0.852	0.835	0.871	0.030
bpMRI model	0.930(0.886–0.966)	0.858	0.868	0.847	0.472
STIR-clinical model	0.892(0.840–0.935)	0.841	0.813	0.871	0.013
T1-FSE-clinical model	0.909(0.860–0.948)	0.858	0.846	0.871	0.047
bpMRI-clinical model	0.944(0.905–0.973)	0.909	0.890	0.929	-

Table 2

Diagnostic performance of late fusion models for early-and late-stage ONFH in the test cohort.

Model	AUC (95%CI)	Accuracy	Sensitivity	Specificity	p-value
Clinical model	0.828(0.765–0.884)	0.790	0.769	0.812	0.0005
STIR model	0.886(0.833–0.930)	0.824	0.791	0.859	0.023
T1-FSE model	0.905(0.856–0.947)	0.852	0.835	0.871	0.042
bpMRI model	0.913(0.867–0.953)	0.835	0.824	0.847	0.480
STIR-clinical model	0.896(0.845–0.938)	0.830	0.780	0.882	0.051
T1-FSE-clinical model	0.914(0.868–0.951)	0.818	0.813	0.824	0.453
bpMRI-clinical model	0.927(0.885–0.960)	0.847	0.813	0.882	-

The clinical-only model, based solely on etiology, sex, and age, achieved the lowest performance (AUC, 0.828), indicating that clinical parameters alone had limited capacity to differentiate early- from late-stage ONFH. Among the single-sequence imaging models, the STIR-only and T1-FSE-only models demonstrated favorable diagnostic performance, with AUCs of 0.886 and 0.905, respectively, suggesting that deep learning can extract staging-relevant features from a single MRI sequence. Adding clinical parameters to the single-sequence models improved the AUCs of the STIR-clinical and T1-clinical models to 0.892 and 0.909, respectively, indicating the supplementary value of clinical information for imaging-based ONFH staging. The bpMRI model, which integrated complementary information from STIR and T1-FSE sequences, further improved performance, reaching an AUC of 0.930 and significantly outperforming either single-sequence model. This finding suggests that multi-sequence MRI provides a more comprehensive representation of the complex pathological changes in ONFH. Among all models, the bpMRI-clinical model, which fused dual-sequence MRI with clinical parameters, achieved the best diagnostic performance. With the early-fusion strategy, it attained an AUC of 0.944 (95% CI: 0.905–0.973), an accuracy of 90.9%, a sensitivity of 89.0%, and a specificity of 92.9%, significantly outperforming all single-modality models (DeLong test, all $P < 0.05$). These results demonstrate that integrating multidimensional information further improves diagnostic performance for early- versus late-stage ONFH stratification.

3. Misclassification Analysis and Overfitting Assessment

All models were first evaluated on the internal test set. The bpMRI-clinical model, which integrates dual-sequence MRI and clinical parameters, achieved the best performance, with an AUC of 0.944 (95% CI: 0.905–0.973). In contrast, the single-modality models, particularly the clinical-only model, showed the poorest performance and produced the highest number of misclassifications, including 16 early-stage cases misclassified as late-stage and 21 late-stage cases misclassified as early-stage (Fig. 2A). We

further examined model accuracy on the training set to assess generalizability and monitor the risk of overfitting; detailed results are presented in Tables 1 and 2.

4. Early Fusion Outperformed Late Fusion for This Task

To compare different multimodal integration approaches for stratifying early- versus late-stage ONFH, we further analyzed the performance of early- and late-fusion strategies. Overall, early fusion demonstrated comparable or marginally superior diagnostic performance. For the biparametric MRI models, early fusion showed a consistent advantage regardless of whether clinical parameters were incorporated. Specifically, the bpMRI-clinical model achieved an AUC of 0.944 with early fusion, compared with 0.927 with late fusion. For models combining a single sequence with clinical parameters (i.e., T1-clinical and STIR-clinical), the performance differences between fusion strategies were minimal, with AUC differences within 0.005; late fusion performed comparably or slightly better than early fusion. Taken together, when biparametric MRI was integrated with structured clinical data for early- versus late-stage ONFH stratification, early fusion demonstrated superior overall performance in this study.

5. Detailed Performance and Interpretability Analysis of the Optimal Model

Based on the above comparisons, the early-fusion bpMRI-clinical model was identified as the optimal model. The confusion matrix for this model on the test set (Fig. 3A) showed an overall misclassification rate of 9.09% (16/176 cases). Specifically, 10 late-stage ONFH cases were misclassified as early-stage (false negatives), and 6 early-stage cases were misclassified as late-stage (false positives). These results indicate that the model achieved high overall classification performance, although misclassification still occurred in some late-stage cases with atypical imaging manifestations. To enhance interpretability, Grad-CAM was used to visualize the decision-making regions of the optimal model. The visualizations showed that the model's attention was predominantly focused on key imaging regions, including the necrotic lesion of the femoral head, the surrounding bone marrow edema zone, and the subchondral fracture line (Fig. 4). These findings suggest that the model's predictions were based primarily on imaging features closely associated with ONFH progression. Thus, in addition to high diagnostic performance, the model showed a meaningful degree of clinical interpretability.

Discussion

This study found that multimodal models integrating dual-sequence MRI with clinical parameters outperformed single-modality models for ONFH staging, particularly at the critical boundary between ARCO stage II and stage IIIA. The early-fusion bpMRI-clinical model achieved the highest AUC (0.944), significantly exceeding the single-sequence and clinical-only models. These findings suggest that multimodal integration can improve the accuracy and robustness of risk stratification when treatment decisions between hip preservation and arthroplasty are most clinically consequential.

This finding is supported by the pathological and imaging characteristics of ONFH. Disease progression involves necrotic lesion expansion, bone marrow edema, subchondral fracture, and femoral-head contour remodeling [22–24]. STIR sequences are sensitive to active changes such as edema and inflammation, helping identify early collapse in ARCO stage IIIA [16], whereas T1-FSE sequences better depict the necrotic band and osseous structures [17]. Their complementary emphasis on lesion activity and structural damage provides the rationale for multi-sequence fusion, particularly for distinguishing stage II from stage IIIA.

To date, most deep learning studies on ONFH have focused on lesion detection or collapse-risk prediction. For example, Chai et al. and Shinohara et al. primarily used radiography or related models to detect early-stage ONFH[25, 26], Shen et al. focused on early MRI-based identification[27], and Gao et al. emphasized collapse-risk prediction[1]. Studies specifically targeting accurate stratification at the pivotal decision-making boundary between ARCO stage II and stage IIIA remain relatively limited. Although the single-sequence STIR convolutional neural network developed by Klontzas et al. showed potential for distinguishing ARCO stage II from stage IIIA in internal testing, its performance declined substantially in external validation (AUC decreasing from > 0.99 to 0.855), suggesting insufficient robustness of a single modality when confronted with variations in scanning parameters and population heterogeneity[14].

The present findings indicate that multimodal integration provides more than a simple aggregation of information; rather, it substantially improves discrimination at the critical staging boundary. Although the STIR and T1-FSE models showed favorable performance (AUCs: 0.886 and 0.905), both produced considerable bidirectional misclassifications. In contrast, the early-fusion bpMRI model achieved a higher AUC of 0.930 and reduced misclassification, whereas the clinical-only model performed worst (AUC: 0.828), suggesting that clinical variables alone are insufficient for accurate staging. These results are consistent with previous multimodal studies showing improved diagnostic or prognostic performance in hip disease and ONFH when imaging and clinical information are combined [28][29]. In this study, clinical parameters such as etiology and age were incorporated to complement, rather than replace, imaging assessment by accounting for heterogeneity in imaging phenotypes. Ultimately, the bpMRI-clinical model achieved the best performance (AUC: 0.944), with a misclassification rate of 9.09% (16/176 cases), including 10 false negatives and 6 false positives, underscoring the value of integrating complementary information to reduce critical staging errors.

Fusion strategy also affected model performance. Early fusion consistently outperformed late fusion in both dual-sequence and multimodal settings (bpMRI-clinical AUC: 0.944 vs. 0.927; bpMRI AUC: 0.930 vs. 0.922), whereas differences were minimal when a single sequence was combined with clinical parameters. This suggests that early fusion may better capture joint representations when modalities have strong anatomical correspondence and complementary pathological information. In this study, STIR and T1-FSE sequences were acquired from the same region and reflected different aspects of ONFH progression, while clinical variables such as etiology and age may further relate to imaging phenotypes. Therefore, early fusion was better suited to modeling cross-modal associations. This finding is consistent with previous evidence showing that multimodal fusion can improve diagnostic,

predictive, and segmentation performance across medical imaging tasks. For example, a systematic review by Zubair et al. noted that multimodal medical image fusion can improve diagnostic, detection, and segmentation performance by exploiting complementary information across modalities[30]. In other diseases, fusing multiparametric MRI sequences has improved the accuracy of brain glioma grading [31], integrating CT imaging with clinical risk factors has enhanced malignancy-risk prediction for pulmonary nodules [32], and combining radiography, CT volumes, and clinical text has improved diagnostic performance for hip disorders [28]. Furthermore, Roest et al., Pluciennik et al., and Lu et al. reported that early fusion may improve heterogeneous data integration, feature-selection stability, or diagnostic sensitivity in studies of prostate cancer, thyroid cancer, and breast tumors, respectively[33–35].

From a clinical perspective, the significance of this study lies not only in the improvement of statistical metrics but also in its potential to optimize the decision-making workflow for ONFH staging. In clinical practice, radiologists and orthopedic surgeons must review multiple MRI sequences and integrate information on etiology, age, and disease course to reach a judgment. However, substantial imaging overlap and interpretive variability exist at the boundary between ARCO stage II and stage IIIA [5, 6], making decision-making susceptible to individual experience and contributing to wide variation and a lack of consensus in treatment strategies [36, 37]. The bpMRI-clinical model provides a unified framework for integrating this information and generating a more objective judgment at this critical point. It may support confidence in hip-preserving treatment for patients classified as early-stage ARCO and help avoid futile hip-preserving attempts in those classified as late-stage ARCO. Thus, multimodal fusion not only improves model performance but also reflects the clinical logic of ONFH staging and may serve as an intelligent decision-support tool.

This study has several limitations. First, its retrospective design makes it difficult to eliminate selection bias completely. Second, despite the use of data augmentation and transfer learning, the overall sample size remained relatively limited. Third, the model has not yet undergone adequate external validation in prospective multicenter cohorts involving different MRI scanners and acquisition parameters; therefore, its generalizability requires further confirmation. Future research should include prospective multicenter external validation to systematically evaluate model stability and generalizability; exploration of additional modalities, such as CT, biochemical markers, and genomic information, on top of the existing dual-sequence MRI and clinical parameters; detailed subgroup analyses centered on the critical boundary between early- and late-stage ARCO; and expanded interpretability analyses using Grad-CAM and feature visualization to clarify the model's decision-making rationale, strengthen clinical interpretability, and facilitate acceptance in practice.

In summary, this study developed a multimodal deep learning model integrating dual-sequence MRI and clinical parameters, improving the accuracy and robustness of ONFH staging. By focusing on the clinically critical distinction between early- and late-stage ARCO disease, the model demonstrates the potential of multimodal fusion for musculoskeletal diagnosis and may support more precise hip-preservation decision-making.

Abbreviations

ONFH Osteonecrosis of the femoral head

ARCO Association Research Circulation Osseous

bpMRI biparametric MRI

STIR Short tau inversion recovery

T1-FSE T1-weighted fast spin-echo

Grad-CAM Gradient-weighted Class Activation Mapping

AUC Area under the receiver operating characteristic curve

THA Total hip arthroplasty

CNN Convolutional neural network

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Institutional Review Board (IRB) of the Third Affiliated Hospital of Guangzhou University of Chinese Medicine (Approval No: PJ-KY-20230919-001). The trial was prospectively registered with the Chinese Clinical Trial Registry (Registration No: ChiCTR2300075354, Sep 1, 2023) and documented in the National Medical Research Registration System. All participating centers complied with the ethical standards established by the primary institution. Additionally, each collaborating institution obtained independent ethical review and approval for their respective cases prior to study initiation. The study was approved by the ethics committee of the primary institution and conducted in accordance with the Declaration of Helsinki. A waiver of informed consent was granted due to the retrospective nature of the study. All patient information was anonymized and de-identified.

Consent for publication

Not applicable.

Competing Interests

All authors declare no competing financial interests or personal relationships that could have influenced this work.

Funding Statement

This work was supported by No. 3 [2018] of the General Office, National Administration of Traditional Chinese Medicine (Medical Administration); the Guangzhou Science and Technology Program (No. 2024A04J5056); the Medical Scientific Research Foundation of Guangdong Province (No. B2020030); the Guangdong Medical Science and Technology Research Fund (No. A2025259) and the Characteristic Innovation Project of Guangdong Provincial Universities (No. 2025KTSCX024).

Author Contribution

X. Han, X. Huang, and G.C. conceived and designed the study. S.L., H.Y., and X.Z. acquired the data. X. Han, H.H., G.C., X. Huang, and X.L. developed the model and performed the statistical analysis. X. Han, S.L., W.H., and L.W. interpreted the results. S.L., H.H., H.Y., X. Han, and G.C. drafted the manuscript, and X. Han, X. Huang, and G.C. critically revised it. All authors approved the final version of the manuscript and are accountable for all aspects of the work.

Data Availability Statement

Due to the privacy of patients, the data related to patients cannot be available for public access but it can be available from the corresponding author upon request.

References

1. Gao S, Zhu H, Wen M, He W, Wu Y, Li Z, et al. Prediction of femoral head collapse in osteonecrosis using deep learning segmentation and radiomics texture analysis of MRI. *BMC Med Inf Decis Mak*. 2024;24:320. <https://doi.org/10.1186/s12911-024-02722-w>.
2. Petek D, Hannouche D, Suva D. Osteonecrosis of the femoral head: pathophysiology and current concepts of treatment. *EFORT OPEN Rev*. 2019;4:85–97. <https://doi.org/10.1302/2058-5241.4.180036>.
3. Wu C-T, Yen S-H, Lin P-C, Wang J-W. Long-term outcomes of phemister bone grafting for patients with non-traumatic osteonecrosis of the femoral head. *Int Orthop*. 2019;43:579–87. <https://doi.org/10.1007/s00264-018-4013-9>.
4. Zhao D, Zhang F, Wang B, Liu B, Li L, Kim S-Y, et al. Guidelines for clinical diagnosis and treatment of osteonecrosis of the femoral head in adults (2019 version). *J Orthop Transl*. 2020;21:100–10. <https://doi.org/10.1016/j.jot.2019.12.004>.
5. Kim J, Lee SK, Kim J-Y, Kim J-H. CT and MRI findings beyond the subchondral bone in osteonecrosis of the femoral head to distinguish between ARCO stages 2 and 3A. *Eur Radio*. 2023;33:4789–800. <https://doi.org/10.1007/s00330-023-09403-8>.

6. Shi S, Luo P, Sun L, Zhao Y, Yang X, Xie L, et al. Analysis of MR signs to distinguish between ARCO stages 2 and 3A in osteonecrosis of the femoral head. *J Magn Reson Imaging*. 2022;55:610–7. <https://doi.org/10.1002/jmri.27860>.
7. Rakhshankhah N, Abbaszadeh M, Kazemi A, Rezaei SS, Roozpeykar S, Arabfard M. Deep learning approach to femoral AVN detection in digital radiography: differentiating patients and pre-collapse stages. *BMC Musculoskelet Disord*. 2024;25:547–58. <https://doi.org/10.1186/s12891-024-07669-7>.
8. Shen X, Luo J, Tang X, Chen B, Qin Y, Zhou Y, et al. Deep learning approach for diagnosing early osteonecrosis of the femoral head based on magnetic resonance imaging. *J Arthroplasty*. 2023;38:2044–50. <https://doi.org/10.1016/j.arth.2022.10.003>.
9. Marka AW, Meurer F, Twardy V, Graf M, Weiss K, Makowski MR, et al. Deep learning and conventional hip MRI for the detection of labral and cartilage abnormalities using arthroscopy as standard of reference. *Eur Radio*. 2025;35:6065–78. <https://doi.org/10.1007/s00330-025-11546-9>.
10. Zhuang Z, Si L, Wang S, Xuan K, Ouyang X, Zhan Y, et al. Knee cartilage defect assessment by graph representation and surface convolution. *IEEE Trans Med Imag*. 2023;42:368–79. <https://doi.org/10.1109/TMI.2022.3206042>.
11. Faghani S, Baffour FI, Ringler MD, Hamilton-Cave M, Rouzrokh P, Moassefi M, et al. A deep learning algorithm for detecting lytic bone lesions of multiple myeloma on CT. *Skeletal Radiol*. 2023;52:91–8. <https://doi.org/10.1007/s00256-022-04160-z>.
12. Bützow A, Anttila TT, Haapamäki V, Ryhänen J. A novel segmentation-based deep learning model for enhanced scaphoid fracture detection. *Eur J Radiol*. 2025;191:112309. <https://doi.org/10.1016/j.ejrad.2025.112309>.
13. Yoon AP, Chung WT, Wang C-W, Kuo C-F, Lin C, Chung KC. Can a Deep Learning Algorithm Improve Detection of Occult Scaphoid Fractures in Plain Radiographs? A Clinical Validation Study. *Clin Orthop*. 2023;481:1828–35. <https://doi.org/10.1097/CORR.0000000000002612>.
14. Klontzas ME, Vassalou EE, Spanakis K, Meurer F, Woertler K, Zibis A, et al. Deep learning enables the differentiation between early and late stages of hip avascular necrosis. *Eur Radio*. 2023;34:1179–86. <https://doi.org/10.1007/s00330-023-10104-5>.
15. Shah AJ, Bt P. Neoplastic and non-neoplastic vertebral marrow pathologies: can the conventional and advanced MRI sequences provide a definitive answer? *Indian J Radiol Imaging*. 2023;33:438–9. <https://doi.org/10.1055/s-0043-1775571>.
16. Mayerhoefer ME, Breitsenseher MJ, Kramer J, Aigner N, Norden C, Hofmann S. STIR vs. T1-weighted fat-suppressed gadolinium-enhanced MRI of bone marrow edema of the knee: computer-assisted quantitative comparison and influence of injected contrast media volume and acquisition parameters. *J Magn Reson Imaging*. 2005;22:788–93. <https://doi.org/10.1002/jmri.20439>.
17. Shah LM, Hanrahan CJ. MRI of spinal bone marrow: part 1, techniques and normal age-related appearances. *Am J Roentgenol*. 2011;197:1298–308. <https://doi.org/10.2214/AJR.11.7005>.
18. Khurana B, Okanobo H, Ossiani M, Ledbetter S, Dulaimy KA, Sodickson A. Abbreviated MRI for patients presenting to the emergency department with hip pain. *Am J Roentgenol*. 2012;198:W581–

8. <https://doi.org/10.2214/AJR.11.7258>.
19. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 2261–9. <https://doi.org/10.1109/CVPR.2017.243>
20. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput Biol Med*. 2021;128:104115. <https://doi.org/10.1016/j.compbiomed.2020.104115>.
21. Chan H-P, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. In: Lee G, Fujita H, editors. *Deep learning in medical image analysis*. Cham: Springer International Publishing; 2020. pp. 3–21. https://doi.org/10.1007/978-3-030-33128-3_1.
22. Hines JT, Jo W-L, Cui Q, Mont MA, Koo K-H, Cheng EY, et al. Osteonecrosis of the Femoral Head: an Updated Review of ARCO on Pathogenesis, Staging and Treatment. *J Korean Med Sci*. 2021;36:e177. <https://doi.org/10.3346/jkms.2021.36.e177>.
23. Chen Y, Miao Y, Liu K, Xue F, Zhu B, Zhang C, et al. Evolutionary course of the femoral head osteonecrosis: histopathological - radiologic characteristics and clinical staging systems. *J Orthop Transl*. 2022;32:28–40. <https://doi.org/10.1016/j.jot.2021.07.004>.
24. Ayabe Y, Motomura G, Yamaguchi R, Utsunomiya T, Sakamoto K, Nakashima Y. Histopathologic confirmation of subchondral fracture in symptomatic pre-collapse osteonecrosis of the femoral head with bone marrow edema on magnetic resonance imaging. *Skeletal Radiol*. 2025;54:1275–81. <https://doi.org/10.1007/s00256-024-04846-6>.
25. Chai R, Tian N, Wan G, Liu S, Zhan J, Li X, et al. Automated detection of early-stage osteonecrosis of the femoral head in adult using YOLOv10: multi-institutional validation. *Eur J Radiol*. 2025;184:111983. <https://doi.org/10.1016/j.ejrad.2025.111983>.
26. Shinohara I, Inui A, Hwang K, Murayama M, Susuki Y, Uno T, et al. Leveraging AI models for lesion detection in osteonecrosis of the femoral head and T1-weighted MRI generation from radiographs. *J Orthop Res*. 2025;43:650–9. <https://doi.org/10.1002/jor.26026>.
27. Shen X, He Z, Shi Y, Yang Y, Luo J, Tang X, et al. Automatic detection of early osteonecrosis of the femoral head from various hip pathologies using deep convolutional neural network: a multi-centre study. *Int Orthop*. 2023;47:2235–44. <https://doi.org/10.1007/s00264-023-05813-x>.
28. Zhang J, Gong H, Ren P, Liu S, Jia Z, Wang L, et al. Intelligent attention-driven deep learning for hip disease diagnosis: fusing multimodal imaging and clinical text for enhanced precision and early detection. *Med (Mex)*. 2026;62:251–72. <https://doi.org/10.3390/medicina62020250>.
29. Yang T, Sun S, Zhang L, Zhang X, He H. A delphi-based model for prognosis of femoral head collapse in osteonecrosis: a multi-factorial approach. *J Orthop Surg Res*. 2024;19:762–76. <https://doi.org/10.1186/s13018-024-05247-0>.
30. Zubair M, Hussain M, Albashrawi MA, Bendeche M, Owais M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical

- image fusion for improved diagnosis. *Comput Methods Programs Biomed.* 2025;272:109014. <https://doi.org/10.1016/j.cmpb.2025.109014>.
31. Naser MA, Deen MJ. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. *Comput Biol Med.* 2020;121:103758. <https://doi.org/10.1016/j.compbiomed.2020.103758>.
 32. Wang C, Shao J, He Y, Wu J, Liu X, Yang L, et al. Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. *Nat Med.* 2024;30:3184–95. <https://doi.org/10.1038/s41591-024-03211-3>.
 33. Roest C, Yakar D, Renner Sitar DI, Bosma JS, Rouw DB, Fransen SJ, et al. Multimodal AI combining clinical and imaging inputs improves prostate cancer detection. *Invest Radiol.* 2024;59:854–60. <https://doi.org/10.1097/RLI.0000000000001102>.
 34. Płuciennik A, Płaczek A, Wilk A, Student S, Oczko-Wojciechowska M, Fajarewicz K. Data integration—possibilities of molecular and clinical data fusion on the example of thyroid cancer diagnostics. *Int J Mol Sci.* 2022;23:11880. <https://doi.org/10.3390/ijms231911880>.
 35. Lu G, Tian R, Yang W, Liu R, Liu D, Xiang Z, et al. Deep learning radiomics based on multimodal imaging for distinguishing benign and malignant breast tumours. *Front Med.* 2024;11:1402967. <https://doi.org/10.3389/fmed.2024.1402967>.
 36. Schmitt-Sody M, Kirchhoff C, Mayer W, Goebel M, Jansson V. Avascular necrosis of the femoral head: inter- and intraobserver variations of Ficat and ARCO classifications. *Int Orthop (SICO).* 2008;32:283–7. <https://doi.org/10.1007/s00264-007-0320-2>.
 37. Cheng E, Mont MA, Goodman SB, Sierra RJ, Cui Q, Salem HS et al. Diagnosis and treatment of femoral head osteonecrosis: a protocol for development of evidence based clinical practice guidelines. 2020. <https://doi.org/10.21203/rs.3.rs-77450/v1>

Figures

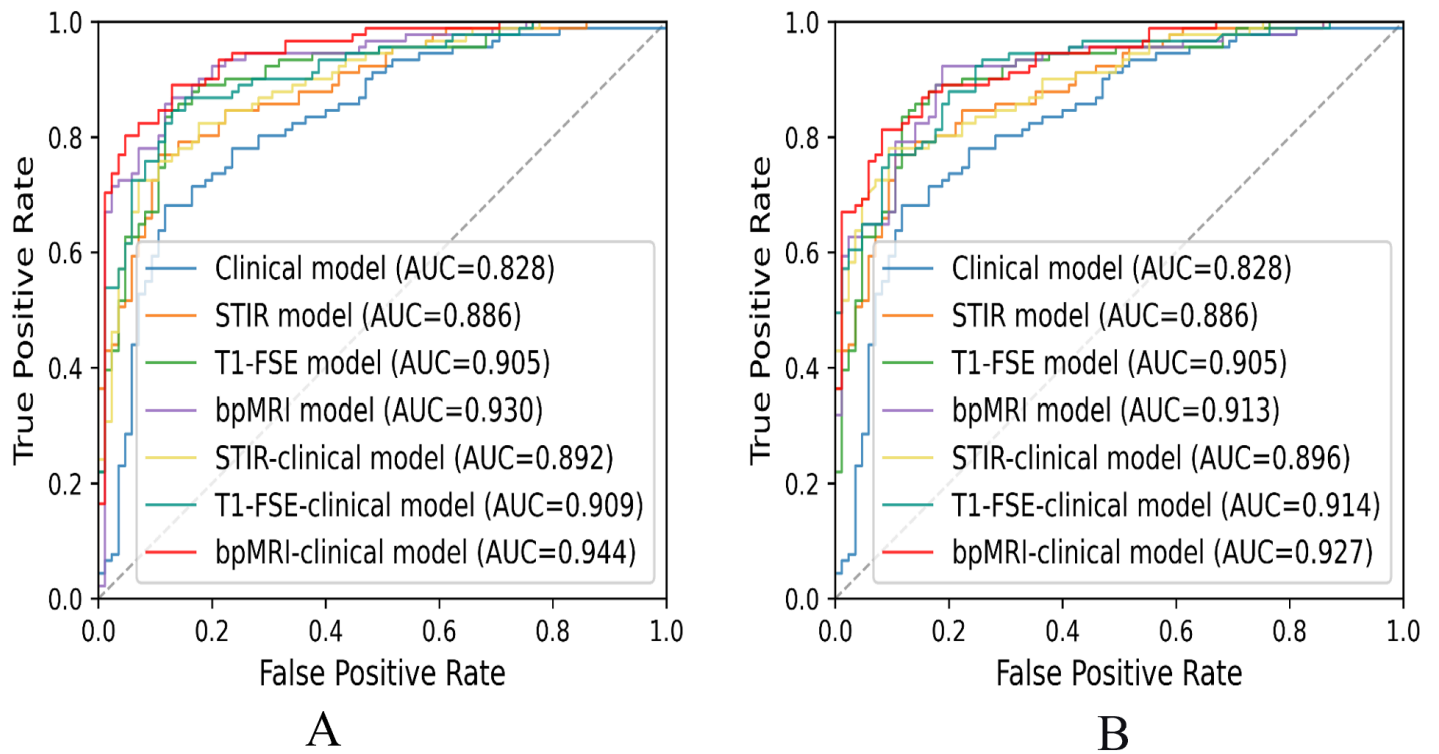


Figure 1

Figure 1

Receiver operating characteristic (ROC) curves.(A) ROC curve under the early fusion strategy. (B) ROC curve under the late fusion strategy.

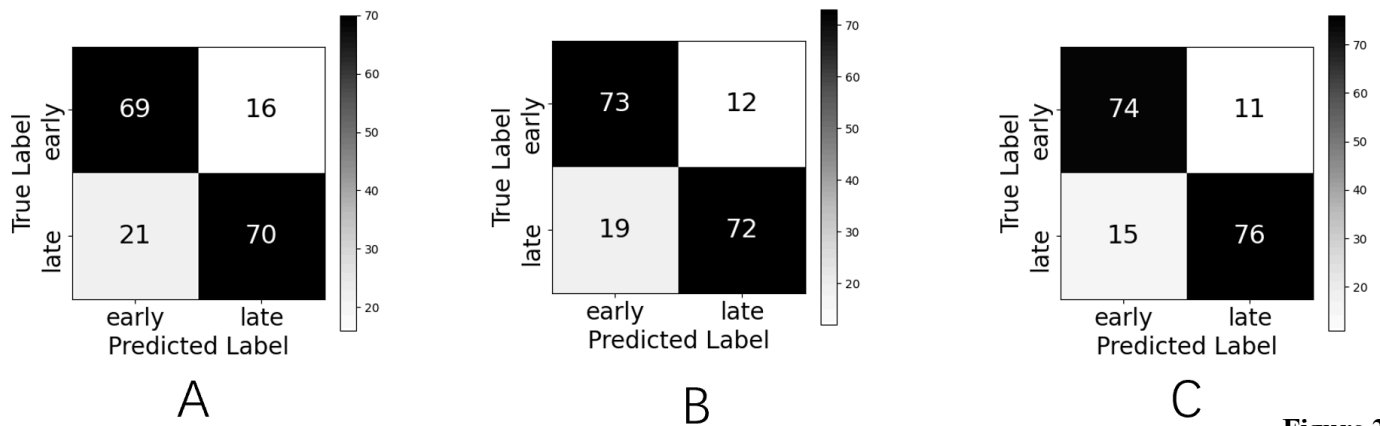


Figure 2

Figure 2

Confusion matrices. (A) Confusion matrix of the Clinical model. (B) Confusion matrix of the STIR (short-TI inversion recovery) model. (C) Confusion matrix of the T1-weighted (T1) model.

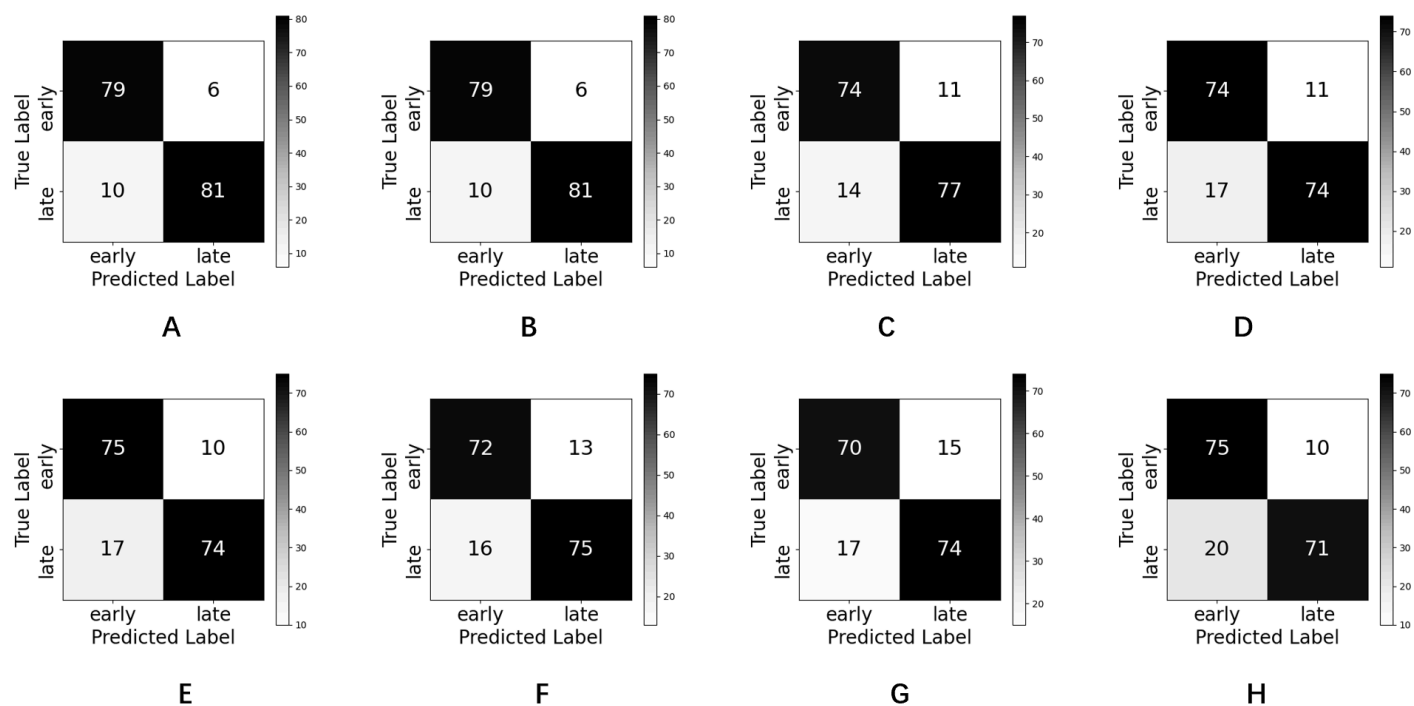


Figure 3

Figure 3

Confusion matrices of models under early and late fusion strategies.(A–D) Early fusion: (A) bpMRI-clinical, (B) bpMRI, (C) T1-clinical, (D) STIR-clinical. (E–H) Late fusion: (E) bpMRI-clinical, (F) bpMRI, (G) T1-clinical, (H) STIR-clinical

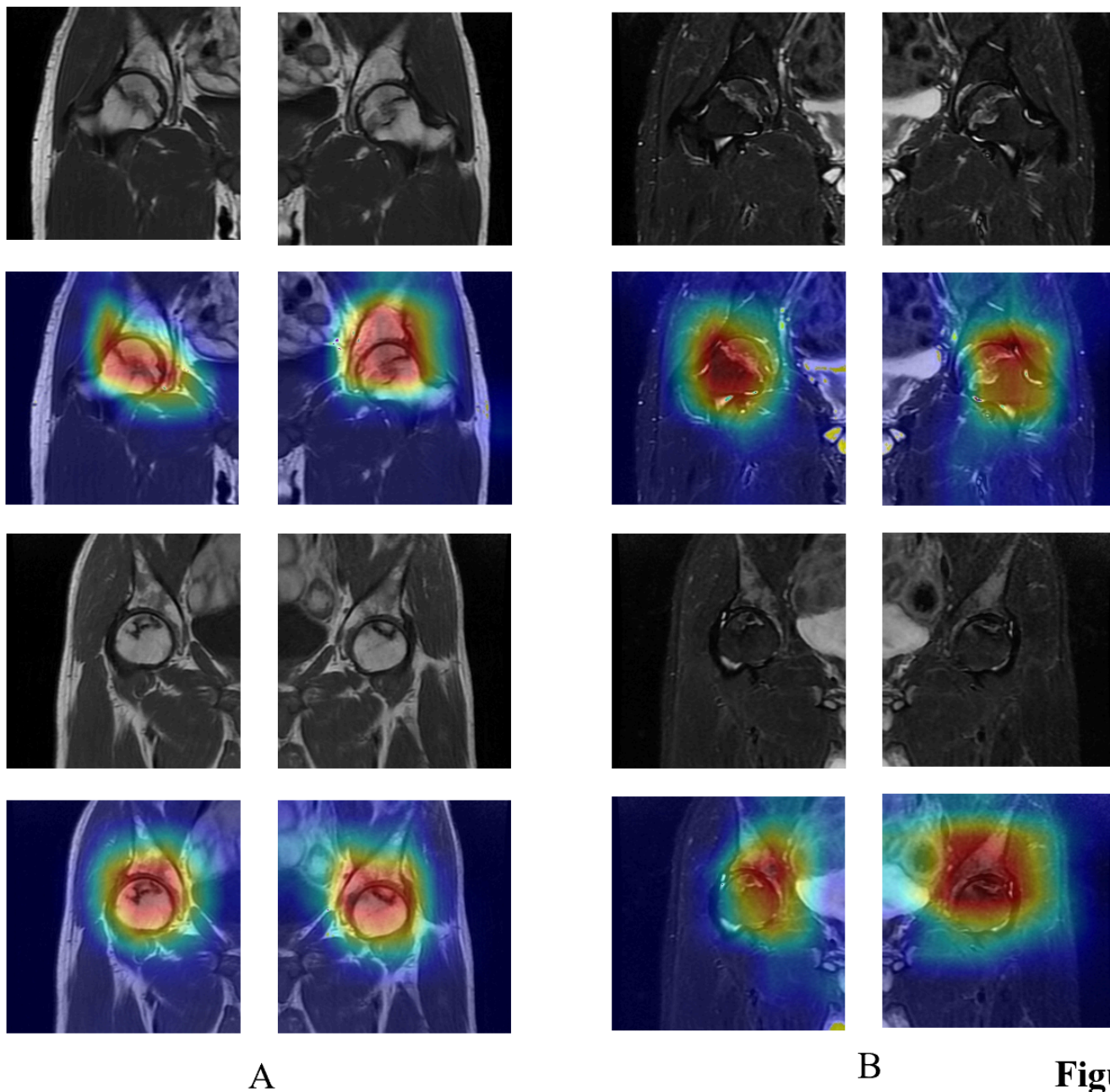


Figure 4

Figure 4

Visualization of feature extraction. (A) T1-FSE MR image and corresponding feature extraction visualization. (B) STIR MR image and corresponding feature extraction visualization.