OA-MultiFuse: A Multimodal Multitask Deep Learning Framework for Knee Osteoarthritis Detection and Severity Assessment

Prof. Ayesha Asif Sayyad[1], Dr. F Rahman[2],
[1]Research Scholar, Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India
[2]Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India

***Abstract*—Knee osteoarthritis is a progressive joint disorder in which cartilage deterioration, osteophyte formation, meniscal change, and joint-space loss develop across multiple anatomical structures. Existing computer-aided approaches often use one imaging modality or optimize classification without explicitly linking the prediction to measurable joint anatomy. This paper proposes OA-MultiFuse, a multimodal and multitask framework that combines a two-dimensional X-ray branch with a three-dimensional MRI branch. The X-ray branch uses a transfer-initialized ConvNeXt encoder and attention pooling for radiographic severity cues. The MRI branch uses a three-dimensional attention U-Net with boundary refinement to segment femur, tibia, cartilage, and joint-space regions. A cross-modal gated fusion module combines the learned representations, while three task heads estimate Kellgren-Lawrence grade, joint-space narrowing, and osteoarthritis probability with uncertainty. The proposed objective joins segmentation, ordinal classification, regression, and calibration losses. This design is motivated by the complementary information provided by X-ray and MRI: radiographs are practical for grading, whereas MRI exposes soft-tissue changes that may precede advanced radiographic damage. The paper defines a reproducible evaluation protocol using patient-level splits, external validation, ablation studies, calibration analysis, and explainability checks. The framework is presented as a proposed methodology; new performance claims require validation on the local and Osteoarthritis Initiative datasets.**

***Index Terms*—Osteoarthritis, knee imaging, MRI segmentation, X-ray grading, joint-space narrowing, multitask learning, multimodal fusion, explainable AI.**

I. INTRODUCTION

Osteoarthritis (OA) is a chronic joint disorder that progressively changes cartilage, subchondral bone, menisci, synovium, and joint geometry. The knee is frequently affected because it carries repeated mechanical loads. Pain, stiffness, reduced mobility, and difficulty with daily activities may appear gradually, while structural damage can progress before a patient seeks specialist care. The clinical problem is therefore not limited to deciding whether OA is present. A useful diagnostic system should also estimate severity, identify the anatomical evidence supporting the prediction, and communicate uncertainty to the clinician.

Radiography remains widely used for routine grading because it is relatively accessible and inexpensive. The Kellgren-Lawrence (KL) scale organizes radiographic findings into grades from 0 to 4, with increasing osteophytes, joint-space narrowing, sclerosis, and deformity. MRI provides a different view. It can reveal cartilage defects, meniscal abnormalities, bone-marrow lesions, and synovial changes that are not fully represented in a projection radiograph. A system that learns from both modalities may therefore combine clinical practicality with soft-tissue sensitivity.

The first uploaded paper describes an intelligent OA diagnostic system centered on MRI processing, Discrete-MultiResUNet segmentation, boundary leakage correction, discrete wavelet transform features, and KL-related analysis [1]. The second uploaded paper reviews X-ray and MRI diagnosis,

transfer learning, 3-D CNNs, DenseNet and ResNet families, PSO-DNN methods, explainability, and clinical workflow [2]. These studies motivate a unified architecture, but they do not provide a single multimodal model that jointly segments anatomy, estimates joint-space narrowing (JSN), predicts ordinal severity, and quantifies uncertainty.

This work proposes OA-MultiFuse, a new multimodal multitask framework. Its purpose is not to claim a measured improvement without a completed experiment. Instead, it specifies a reproducible architecture and evaluation plan that can be implemented on the local dataset and the Osteoarthritis Initiative (OAI) data. The design uses X-ray images for global radiographic evidence and MRI volumes for anatomical localization. The two streams are fused only after modality-specific feature extraction, reducing the risk that one imaging type dominates the representation.

The main contributions are: (1) a two-stream X-ray and MRI architecture; (2) a 3-D attention segmentation module with explicit boundary refinement; (3) a cross-modal gated fusion layer; (4) joint prediction of KL grade, JSN, OA probability, and uncertainty; (5) an ordinal loss that respects the ordered nature of KL categories; and (6) a patient-level validation protocol designed to reduce leakage and assess generalization.

II. RELATED WORK AND MOTIVATION

A. Imaging-Based OA Assessment

X-ray imaging is useful for identifying osteophytes, joint-space narrowing, sclerosis, and gross alignment changes. Its limitations include projection overlap, two-dimensional representation, and reduced visibility of early cartilage damage. MRI provides richer soft-tissue information but is more expensive and less available for routine screening. A multimodal design should therefore treat the modalities as complementary rather than interchangeable.

The second uploaded paper summarizes multiple imaging modalities and notes that X-ray, MRI, CT, ultrasound, nuclear medicine, and optical methods expose different OA features and have different costs and limitations [2]. Its literature survey includes CNN, transfer learning, 3-D DenseNet, ResNet, PSO-DNN, Siamese networks, and multimodal signals. The first paper emphasizes MRI denoising, bone-region initialization, boundary correction, and DWT-based feature extraction [1].

B. Segmentation and JSN Measurement

Segmentation is clinically meaningful because it transforms a black-box image prediction into an anatomical measurement problem. The first paper uses a MultiResUNet-derived strategy to initialize femur and tibia regions, correct boundary leakage across adjacent slices, and construct a mask for joint-space analysis [1]. In the proposed framework, segmentation is retained but redesigned as a 3-D attention U-Net with a boundary-aware refinement head. The target regions are femur, tibia, cartilage, and joint space.

Joint-space narrowing is treated as a continuous target rather than only an implicit feature. After segmentation, the system estimates the minimum or robust percentile distance between opposing cartilage and bone boundaries in the medial and lateral compartments. The measurement is normalized by image spacing and anatomical scale, allowing comparison across scanners and acquisition protocols.

C. Deep Learning and Explainability

The reviewed papers report the use of transfer learning and deep architectures for OA classification and grading, including ResNet, DenseNet, CNN, Siamese networks, and other feature extractors [1], [2]. These approaches can learn useful image features, but a classification score alone does not prove that the model used the correct anatomical evidence. Explainability is therefore included in OA-MultiFuse through X-ray Grad-CAM, MRI segmentation overlays, and a structured reason-for-referral summary.

A central design decision is to make explanation and measurement part of the model output rather than a post-hoc visualization added after training. The MRI branch must identify anatomical structures, and the JSN head must produce a measurable output. The system can then show whether a high predicted KL grade is consistent with narrowing, cartilage loss, or an osteophyte-related attention region.

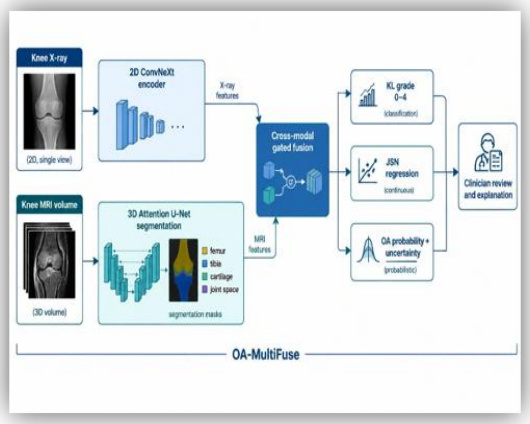

Fig. 1. Proposed OA-MultiFuse architecture combining X-ray grading, MRI segmentation, multimodal fusion, and clinical review.

III. PROPOSED OA-MULTIFUSE ARCHITECTURE

A. Input and Preprocessing

The system accepts a weight-bearing knee radiograph and an MRI volume when both are available. Each case is linked by a patient-level identifier, but the identifier is not passed to the neural network. X-ray preprocessing includes orientation correction, intensity normalization, removal of irrelevant borders, and crop standardization around the tibiofemoral joint. MRI preprocessing includes DICOM-to-volume conversion, voxel-spacing normalization, bias-field correction, intensity standardization, and removal of slices with insufficient anatomical content.

Augmentation is applied separately to the modalities. X-ray augmentation can include small rotations, translations, contrast variation, and controlled horizontal reflection when laterality is explicitly managed. MRI augmentation can include small spatial transformations, intensity perturbations, simulated noise, and slice-consistent deformation. The augmentation pipeline must not create anatomically implausible cartilage or joint-space patterns.

B. X-Ray Branch

The X-ray branch uses a transfer-initialized ConvNeXt encoder followed by attention pooling. ConvNeXt is selected as a modern convolutional backbone with large receptive fields and hierarchical feature extraction. The final feature map is converted into a compact vector by learned spatial attention rather than simple global average pooling. The attention map is retained for Grad-CAM-style explanation and for checking whether the model focuses on the tibiofemoral compartments.

The X-ray branch produces a feature vector x and auxiliary logits for KL grade. Auxiliary supervision prevents the fusion layer from discarding radiographic severity information. If a case lacks MRI, the X-ray branch can still generate a provisional prediction, but the output should be marked as single-modality and assigned an uncertainty penalty.

C. MRI Branch

The MRI branch is a three-dimensional attention U-Net. Its encoder extracts volumetric context, while its decoder reconstructs voxel-level masks through skip connections. Attention gates suppress irrelevant background and emphasize the femoral condyle, tibial plateau, cartilage layer, and joint-space region. The model predicts a four-class anatomical mask: femur, tibia, cartilage, and joint space.

A boundary refinement module operates on the decoder features and predicted contour. It estimates boundary confidence and penalizes discontinuities between adjacent slices. This redesign addresses the boundary leakage issue discussed in the first paper, where adjacent-slice point distances are used to identify abnormal contour displacement [1]. The proposed model performs the correction through a learned boundary head combined with a deterministic post-processing check.

D. Cross-Modal Gated Fusion

The X-ray and MRI embeddings are projected to the same dimensionality. A gated fusion layer learns how much each modality should contribute for each case:

$$g = sigmoid\,(Wg\,[x \,||\, m] + bg)$$

$$z = g \odot x + (1 - g) \odot m$$

where x is the X-ray embedding, m is the MRI embedding, || denotes concatenation, ⊙ is element-wise multiplication, and z is the fused representation. The gate can give greater weight to MRI when cartilage or meniscal evidence is available, and greater weight to X-ray when the radiograph provides clearer KL-related features. A missing-modality

mask is included so that the model does not interpret absent MRI as a normal MRI finding.

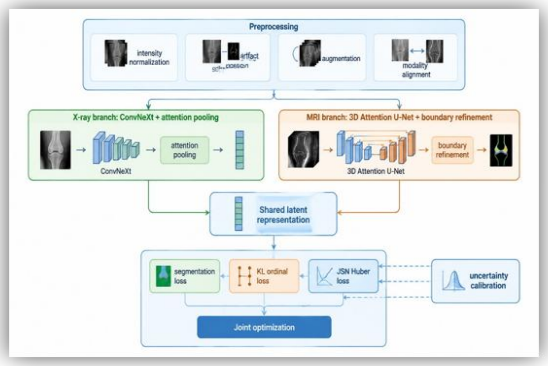

Fig. 2. Multitask training design with modality-specific branches, joint optimization, and uncertainty calibration.

E. Output Heads

The fused representation feeds four outputs. The first is an ordinal KL head. Instead of treating grades 0 through 4 as unrelated classes, the model estimates four cumulative probabilities, such as P(y > 0), P(y > 1), P(y > 2), and P(y > 3). This penalizes a grade-0 to grade-4 error more strongly than a grade-2 to grade-3 error.

The second head predicts JSN using a Huber regression loss. The third estimates OA probability for a binary referral decision. The fourth estimates predictive uncertainty using a calibrated probabilistic head or an ensemble of stochastic forward passes. These outputs are reported together rather than reducing the case to a single class label.

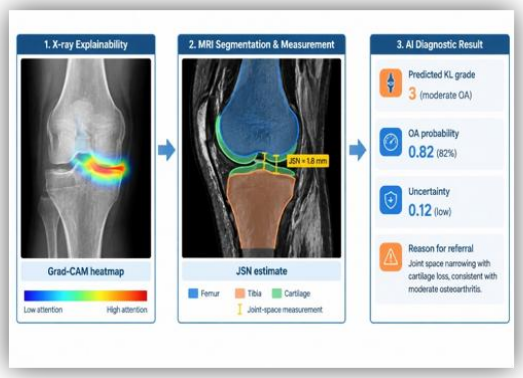

Fig. 3. Explainability and measurement output: radiographic attention, MRI anatomical segmentation, JSN estimate, and calibrated diagnostic summary.

IV. MULTITASK OBJECTIVE

The total training objective combines segmentation, ordinal grading, JSN regression, binary classification, and uncertainty calibration:

$$Ltotal = \lambda seg\ Lseg + \lambda ord\ Lord + \lambda jsn\ Ljsn + \lambda oa\ Loa + \lambda cal\ Lcal + \lambda reg\ ||\theta||2$$

The segmentation loss combines Dice loss and weighted cross-entropy:

$$Lseg = 1 - Dice + \lambda b\ Lboundary$$

The boundary term compares predicted and reference contours. A distance-transform formulation can penalize errors near the boundary more strongly than errors in uniform interior regions. This is important because a small contour displacement can produce a large JSN measurement error even when the overall Dice score remains high.

For KL grading, the ordinal loss is:

$$Lord = -\Sigma k\ [tk\ log(pk) + (1 - tk)\ log(1 - pk)]$$

where tk is the target for cumulative threshold k and pk is the predicted probability that the grade exceeds threshold k. The JSN head uses Huber loss to reduce sensitivity to occasional annotation errors. The calibration loss compares predicted confidence with observed correctness, and the final uncertainty score is used to determine whether a case should be referred for expert review.

V. TRAINING AND INFERENCE PROCEDURE

A. Training Algorithm

Algorithm 1 summarizes the proposed training procedure.

Input: paired or partially paired X-ray and MRI cases with KL, JSN, and anatomical labels.

1. Normalize each modality and standardize orientation and spatial scale.
2. Generate modality-specific augmentations while preserving anatomy.
3. Encode the X-ray with ConvNeXt and the MRI with 3-D attention U-Net.
4. Predict anatomical masks and refine MRI boundaries.
5. Create X-ray and MRI embeddings and apply the missing-modality mask.
6. Compute the gated multimodal representation.
7. Predict KL grade, JSN, OA probability, and uncertainty.
8. Optimize the weighted multitask objective.

9. Calibrate probabilities on the validation set.
10. Save the model only if patient-level validation improves without a fairness regression.

B. Inference and Referral Logic

At inference, the system first checks image quality and modality availability. Low-quality images are not silently classified; they are routed to a quality-control state. When both modalities are available, the fused output is used. When only one modality is available, the corresponding single-stream output is reported with an explicit uncertainty increase.

A referral rule is defined as a safety layer rather than a treatment recommendation. A case is referred when the uncertainty exceeds a threshold, when segmentation confidence is low, when the predicted grade conflicts with the JSN estimate, or when the model detects a distribution shift. The clinician receives the image, heatmap, segmentation, predicted values, confidence, and reason for referral.

VI. DATASETS AND EVALUATION DESIGN

The two source papers use local and OAI-related datasets and describe training, validation, and testing partitions [1], [2]. The proposed evaluation should preserve patient-level separation. Slices from the same patient must never appear in both training and test sets. If multiple images are available for one knee, all images from that knee should remain within the same partition.

TABLE I. Proposed evaluation protocol

Component	Recommended design	Purpose
Data split	Patient-level train/validation/test split; external test site when available	Reduces identity leakage and tests generalization
Primary tasks	KL grade, OA probability, JSN regression, four-class anatomy segmentation	Evaluates clinical label and anatomical evidence
Baselines	CNN, ResNet, DenseNet, 3-D CNN, single-modality models	Measures the value of the proposed fusion
Ablations	No MRI, no X-ray, no boundary head, no ordinal loss, no calibration	Identifies which components contribute
Explanations	Grad-CAM localization, segmentation overlap, clinician review	Tests whether evidence is anatomically plausible
Robustness	Noise, contrast shift, missing modality, scanner/site shift	Assesses deployment reliability

A. Classification Metrics

KL grading should be evaluated with macro-F1, balanced accuracy, quadratic weighted kappa, sensitivity, specificity, and confusion matrices. The quadratic weighted kappa is useful because KL categories are ordered. Binary OA referral should report AUROC, area under the precision-recall curve, sensitivity, specificity, negative predictive value, and calibration.

B. Segmentation and Measurement Metrics

Segmentation should report Dice score, intersection over union, Hausdorff distance, average surface distance, and boundary F1. JSN should report mean absolute error, root mean squared error, correlation, and agreement with expert measurements. A high segmentation Dice score should not be interpreted as sufficient if JSN error remains clinically unacceptable.

C. Calibration and External Validation

The model should be evaluated on data from a different acquisition setting or institution. Calibration curves, expected calibration error, Brier score, and uncertainty-stratified accuracy should be reported. The source paper acknowledges problems related to image quality, label discrepancies, controlled acquisition, and validation bias [1]. The second paper similarly emphasizes the need for robust clinical validation and interpretability [2]. These limitations make external testing essential.

VII. COMPARISON WITH THE SOURCE METHODS

TABLE II. Methodological comparison

Aspect	Source approaches	OA-MultiFuse proposal
Input	Primarily MRI or X-ray-oriented studies	Joint X-ray and MRI when available
Segmentation	MultiResUNet, seed initialization, DWT and boundary correction	3-D attention U-Net, learned boundary head, deterministic contour check
Classification	CNN, transfer learning, ResNet, DenseNet and related models	ConvNeXt branch with ordinal KL head
Fusion	Mostly discussed as separate methods or literature themes	Cross-modal gated feature fusion
Outputs	OA classification, KL grading or JSN-related values	KL grade, OA probability, JSN, anatomy masks and uncertainty
Explainability	Grad-CAM and XAI discussed in related work	Heatmaps and anatomical evidence included in the output design
Validation	Reported values from different studies and datasets	Patient-level, external, ablation and calibration protocol

The comparison is methodological, not a claim that OA-MultiFuse has already exceeded the reported source results. The first paper reports strong values for its proposed architecture, including an overall accuracy of 96.85% in its comparative analysis and sensitivity, specificity, precision, F1, and correlation metrics for the proposed model [1]. The second paper summarizes a range of reported CNN, transfer-learning, 3-D, PSO-DNN, and Siamese approaches [2]. These values cannot be compared directly because the datasets, labels, preprocessing, evaluation splits, and tasks differ. A fair comparison requires re-running the baselines on the same patient-level test set.

VIII. DISCUSSION

The main expected advantage of OA-MultiFuse is not simply a deeper network. It is the alignment of prediction with anatomy and clinical uncertainty. The X-ray branch provides a practical radiographic view, the MRI branch localizes tissue-level changes, and the JSN head creates a measurable bridge between segmentation and severity. This design may be useful when the KL grade is ambiguous or when early structural changes are not obvious in a radiograph.

The multimodal design also creates practical limitations. Many clinical cases will have only one modality. MRI and X-ray may be acquired at different visits, with different knee positions or disease stages. Pairing them may therefore introduce temporal mismatch. The missing-modality mask and single-stream fallback reduce this problem, but they do not replace a carefully designed paired dataset.

Class imbalance is another concern. Severe cases may be less common than mild or doubtful cases, while local datasets may overrepresent selected referrals. The training protocol should use class-balanced sampling, ordinal loss, and external calibration rather than relying only on overall accuracy. The model should also be tested separately by sex, age group, body mass index category, imaging device, and institution when those metadata are available.

Explainability must be evaluated rather than assumed. A heatmap that looks visually plausible may still be unrelated to the decision. Therefore, Grad-CAM should be compared with annotated regions, segmentation masks, perturbation tests, and clinician ratings. Similarly, a segmentation mask should not be treated as clinically meaningful unless its boundaries produce stable JSN estimates.

IX. LIMITATIONS AND ETHICAL CONSIDERATIONS

This paper presents a proposed research framework and does not report new experimental results. The numerical values in the source papers are not transferred as results for the proposed model. The architecture requires paired or partially paired multimodal data, reliable labels, consistent imaging protocols, and clinical review.

The framework is intended for decision support, not autonomous diagnosis or treatment. Patient privacy, de-identification, secure storage, access control, and data-use consent are required. Models should be versioned and monitored after deployment because scanner changes, patient populations, and clinical practices can cause distribution shift. A human expert should review uncertain, low-quality, contradictory, or high-impact cases.

X. CONCLUSION

This paper proposed OA-MultiFuse, a new multimodal multitask architecture for knee osteoarthritis assessment. The framework combines X-ray radiographic grading with MRI-based anatomical segmentation, gated feature fusion, ordinal KL prediction, JSN regression, OA probability, uncertainty calibration, and explainability. It brings together the clinically practical grading emphasis of radiographic deep-learning studies with the anatomical and boundary-aware MRI processing described in the source work [1], [2].

The proposed contribution is a reproducible methodology rather than an unsupported performance claim. Future work should implement the system on patient-level local and OAI splits, compare it with single-modality CNN, ResNet, DenseNet, 3-D CNN, and MultiResUNet baselines, and report external validation, calibration, subgroup fairness, and clinician-centered explainability. If the framework improves both predictive performance and anatomical reliability under these tests, it may provide a stronger foundation for practical computer-aided knee OA assessment.

REFERENCES

[1] A. A. Sayyad and R. K. Deshmukh, “Intelligent medical diagnostic system for osteoarthritis: A deep learning approach and comparison,” IJIRT, vol. 11, no. 11, 2025.

[2] A. A. Sayyad and R. K. Deshmukh, “Intelligent medical diagnostic system for osteoarthritis using deep learning,” IJSRST, vol. 12, no. 2, pp. 53–64, 2025, doi: 10.32628/IJSRST25121214.

[3] P. S. Q. Yeoh et al., “Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the Osteoarthritis Initiative,” Front. Bioeng. Biotechnol., 2023, doi: 10.3389/fbioe.2023.1164655.

[4] J. Martel-Pelletier, P. Paiement, and J.-P. Pelletier, “Magnetic resonance imaging assessments for knee segmentation and their use in combination with machine/deep learning as predictors of early osteoarthritis diagnosis and prognosis,” Ther. Adv. Musculoskelet. Dis., vol. 15, 2023, doi: 10.1177/1759720X231165560.

[5] Y. X. Teoh et al., “Discovering knee osteoarthritis imaging features for diagnosis and prognosis: Review of manual imaging grading and machine learning approaches,” J. Healthcare Eng., 2022, doi: 10.1155/2022/4138666.

[6] F. Calivà et al., “Studying osteoarthritis with artificial intelligence applied to magnetic resonance imaging,” Nat. Rev. Rheumatol., vol. 18, no. 2, pp. 112–121, 2022, doi: 10.1038/s41584-021-00719-7.

[7] Y. G. Satorras, E. Hoogeboom, and M. Welling, “E(n) equivariant graph neural networks,” in Proc. ICML, 2021, pp. 9323–9332.

[8] Y. Wang et al., “Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing,” Nat. Commun., vol. 14, 2023, doi: 10.1038/s41467-023-43720-2.

[9] A. Tiulpin et al., “Automatic knee osteoarthritis diagnosis from plain radiographs,” in Proc. IEEE ICMLA, 2018, pp. 1–7.

[10] K. D. Allen, L. M. Thoma, and Y. M. Golightly, “Epidemiology of osteoarthritis,” Osteoarthritis Cartilage, vol. 30, no. 2, pp. 184–195, 2022, doi: 10.1016/j.joca.2021.04.020.

[11] Y. X. Teoh, A. Othmani, S. L. Goh, J. Usman, and K. W. Lai, “Deciphering knee osteoarthritis

diagnostic features with explainable artificial intelligence: A systematic review," IEEE Access, vol. 12, pp. 109080–109108, 2024, doi: 10.1109/ACCESS.2024.3439096.

[12] S. M. Ahmed and R. J. Mstafa, "Identifying severity grading of knee osteoarthritis from X-ray images using an efficient mixture of deep learning and machine learning models," Diagnostics, vol. 12, no. 12, p. 2939, 2022, doi: 10.3390/diagnostics12122939.

[13] S. M. Ahmed and R. J. Mstafa, "A comprehensive survey on bone segmentation techniques in knee osteoarthritis research," Diagnostics, vol. 12, no. 3, p. 611, 2022, doi: 10.3390/diagnostics12030611.

[14] K. Leung et al., "Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs," Radiology, vol. 296, no. 3, pp. 584–593, 2020, doi: 10.1148/radiol.2020192091.

[15] Z. Zhao et al., "Identifying significant structural factors associated with knee pain severity in patients with osteoarthritis using machine learning," Sci. Rep., vol. 14, p. 14705, 2024, doi: 10.1038/s41598-024-65613-0.