Fair and Explainable AI for the Automatic Diagnosis of Knee Osteoarthritis

Presenter: Soheyla Amirian

Vladyslav Protsenko, Danylo Reznyk, Yaroslav Klym, Nickolas Littlefield, Kurt R. Weiss, Johannes F. Plate, Ahmad P. Tafti, Soheyla Amirian

PennAITech Collaboratory forHealthy Aging

Background and Motivation

Prevalence and Impact: Knee osteoarthritis (OA) is a major cause of pain and disability, especially among the aging population.
Limitations of Traditional Methods: Kellgren-Lawrence (KL) grading scale is subjective, leading to inconsistent diagnoses and patient care.
Role of AI and Deep Learning:

- AI and deep learning (DL) methods have shown potential in automating knee OA analysis using X-rays and MRIs.
- Existing models rely mainly on single-modal data, ignoring patient-specific factors like sex and race.

Proposed Framework:

- Integrates MRI scans with patient metadata to improve accuracy and fairness in OA diagnosis.
- Aims to standardize diagnosis and reduce biases.

Explainability:

- Focuses on increasing clinician trust by making the AI's decision-making process more transparent.
- Enhances understanding of model predictions for better patient outcomes.

Data

Data Collection

• Data sourced from the NIH Osteoarthritis Initiative (OAI) dataset.
• Dataset includes imaging, biochemical, genetic, and clinical data from an 11-year longitudinal study.
• Focused on baseline visit with ~7,000 samples, incorporating metadata on sex and race.

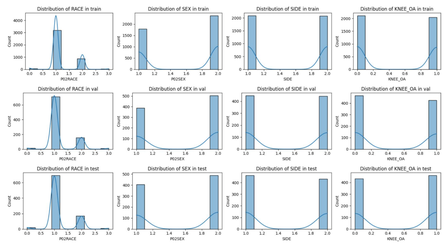
Data Processing

• Clinical data: Extracted KL grades, converted to integers, and removed missing values.
• MRI data:

- Reformatted to right anterior-superior (RAS) orientation.
- Resampled to 1 × 1 × 1 mm resolution and resized to 128 × 128 × 35.
- Normalized intensities and clipped to [0, 2126].

• Training pipeline:

- Pre-scaling and augmentation (Gaussian noise, intensity shifts).
- Data split: 70% training, 15% validation, 15% testing.
- Down-sampled no-OA class by 50% for balance.
- KL scores binarized: ≥2 = 1 (OA), <2 = 0 (no OA).



Data distribution across train, validation, and test splits.

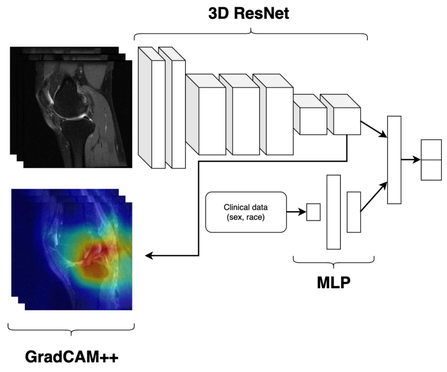
Methods and Results

AI Model(s)

• Used 3D ResNets (ResNet-10, ResNet-18, ResNet-34) for MRI data.
• Multi-modal setup combined ResNet output with MLP for clinical data.
• MLP: Single hidden layer (size 8), output vector size 4.
• Pre-trained weights from MedicalNet (trained on 23 medical datasets) via Monai framework.

AI Explainability and Fairness

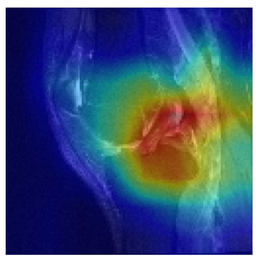
• Used GradCAM++ to visualize influential image regions in model decisions.
• Red regions = high predictive influence; blue = low influence.
• Focused on knee joint space narrowing as a key predictor.
• Aimed to reduce bias for sociodemographic groups (sex, race).



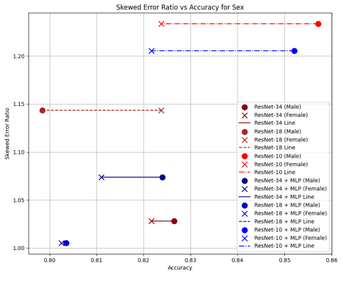
Architecture of the model.

Performance metrics for different mono-modal and multi-modal models. The best performing model is highlighted in bold.

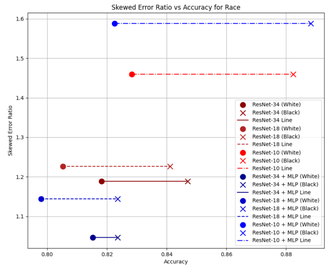
Model	Accuracy	F1-Score	Precision	Recall
ResNet-10	0.8355	0.8353	0.8391	0.8367
ResNet-10 + MLP	**0.8389**	**0.8389**	**0.8392**	**0.8393**
ResNet-18	0.8123	0.8123	0.8124	0.8126
ResNet-18 + MLP	0.8030	0.8022	0.8120	0.8050
ResNet-34	0.8239	0.8239	0.8242	0.8243
ResNet-34 + MLP	0.8169	0.8168	0.8168	0.8168



Example of GradCAM++ output. The red color signifies the regions with the highest predictive influence in the MRI image.



Accuracy and SER for knee OA classification in ResNet-based models with and without an MLP module for sex groups.



Accuracy and SER for knee OA classification in ResNet-based models with and without an MLP module for race groups.

Conclusions

This research developed and validated a multi-modal AI-powered classifier for knee OA classification using ResNet combined with MLP for clinical data integration. Results showed that multi-modality does not always improve classification accuracy but can enhance fairness. The best model, using ResNet-18 and MLP, achieved fairness scores of 1.1448 for race and 1.0052 for sex, indicating reduced bias in sex classification but room for improvement in racial fairness due to dataset imbalance. Despite achieving the lowest accuracy of 0.8030, the model demonstrated that fairness improvements are possible with minimal performance loss. Future improvements may require more advanced deep learning architectures and additional data modalities, such as X-rays or metadata like age, to enhance both accuracy and explainability.

References

[1] URL: https://nda.nih.gov/oai. July 2023
[2] El Jurdi, R., Petitjean, C., Honeine, P. and Abdallah, F., 2020. Bb-unet: U-net with bounding box prior. IEEE Journal of Selected Topics in Signal Processing, 14(6), pp.1189-1198.
[3] Littlefield, N., Moradi H., Amirian, S., Maradit Kremers, H., Plate, JF., Tafti, AP., Enforcing Explainable Deep Few-Shot Learning to Analyze Plain Knee Radiographs: Data from the Osteoarthritis Initiative. IEEE ICHI 2023
[4] Littlefield, N., Plate, JF., Weiss, KR., Lohse, I., Chhabra, A., Siddiqui, I., Menezes, Z., Matorakos, G., Thhakar, S., Abedian, M., Gong, M., Carlson, L., Moradi, H., Amirian, S., Tafti AP. Learning Unbiased Image Segmentation: A Case Study with Plain Knee Radiographs. IEEE BHI 2023
[5] Gao, F., Littlefield, N., Myers, N., Yates Jr., A. J., Weiss, K., Plate, JF., Tafti, AP., Amirian, S. Explainable Contrastive Learning for KL Grading Classification in Knee Osteoarthritis. Submitted at IEEE EMBC 2025.

Acknowledgment

The authors declare that they have no competing interests. This research was supported in part by a grant from the Simons Foundation (SFARI award #1280457, JS) and by the NYU Center for Responsible AI through the RAI for Ukraine program. Also, thank you to the Helene and Grant Wilson Center for Social Entrepreneurship at Pace University for supporting the research that led to this paper and for providing a supportive space for Wilson Center Faculty Fellows to learn from, teach with, and be inspired by innovative research in social entrepreneurship. In addition, **the research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG073105. The content is solely the responsibilityof the authors and does not necessarily represent the official views of the National Institutes of Health.**