An Explainable Hybrid Deep Learning Framework for Knee Deterioration Level Prediction and Surgery Recommendation

Manimegalai P[1], Dr Kalpana Nagpal[2]

[1] Postdoctoral Researcher,Lincoln University College, Malaysia; [2] Associate Professor & Center Head, PRA, Amity Institute of Pharmacy, AUUP

pdf.manimegalai@lincoln.edu.my, kalpananagpal@gmail.com

Abstract: Osteoarthritis is a degenerative joint disease which primarily affects the knee and has a major effect on quality of life and mobility. The Kellgren-Lawrence (KL) grading system used in conventional diagnosis depends on the subjective and human error-prone manual interpretation of X-rays. To overcome this limitation, this study offers a multimodal, explainable framework that combines radiographic imaging and clinical information to automatically assess knee osteoarthritis using the Kellgren–Lawrence scale and suggest surgery when necessary. The clinical segment achieves 93.3% accuracy using a stacked ensemble of XGBoost, LightGBM, and CatBoost trained on functional, symptomatic, and demographic data such as age, BMI, KOOS pain score, edema, and crepitus. The imaging segment achieves 80% accuracy by fine-tuning EfficientNetV2L with a Transformer head on 5,000 knee X-rays. Decision-level weighted fusion of both produces a hybrid system with 94% accuracy on a test set of 900 samples. Explainability is provided through SHAP for clinical feature and LIME for image , allowing clinicians to examine both numerical and visual explanations for each prediction. The suggested framework gives a dependable, understandable, and clinically significant approach for OA severity classification and surgical recommendation, for real-time hospital implementation, longitudinal disease monitoring, multi-center validation, and integration of biomechanical parameters.

Keywords: Knee Osteoarthritis, Surgery Prediction, Explainable Artificial Intelligence, Clinical Decision Support System, Machine Learning

Introduction

Osteoarthritis (OA), the most commonly recurrent etiology of chronic disability worldwide, is a degenerative disease affecting articular cartilage, subchondral bone, synovium, ligaments, and periarticular muscles. Globally, the estimated burden of knee OA (KOA) has been on an upward trend since 1990 and was an estimated 374.7 million individuals in 2021 [1]. Old age, obesity, previous joint injury, extreme mechanical load on joints, and metabolic abnormalities contribute to the onset and progression of KOA [2]. Clinically, KOA considerably causes loss of function and independent capacity of people to be mobile and therefore limits their quality of life. KOA commonly presents itself in the knee and in other weight-bearing joints. Classification based on the severity of osteoarthritis on a plain X-ray can be done using the Kellgren-Lawrence (KL) scale from grade 0 (no injury) to 4 (severe injury). But classification is still subject to manually interpreting X-rays, leading to inconsistent results between different observers, high time costs, and low sensitivity to detect early changes. The advancement of artificial intelligence (AI) and particularly deep learning (DL) technology paved the way toward the automation of image analysis and providing a reliable and scalable analysis of joint degradation. In parallel, clinical feature-based machine learning models such as individual characteristics (factors of risk, symptom severity scale, and demographic factor) also demonstrated good promise to predict disease state and progression. Nevertheless, most of these models acting as black boxes lacking interpretability lead to a difficulty in their implementation and generalization to the clinical setting. Therefore, a multimodal model combining clinical and imaging data with explainable AI (XAI) is a possible venue for accurately and understandably predicting individual osteoarthritis of the knee, providing treatment guidelines to help the clinical choice.

The Explainable Deep Learning Framework for Knee Deterioration Level Prediction and Surgery Recommendation in this paper integrates radiographic images and structured clinical data using a hybrid framework architecture, whereby the clinical branch employs an ensembled stacked model consisting of XGBoost, LightGBM, and CatBoost to learn complex clinical patterns, and the imaging branch uses an EfficientNetV2L integrated with a Transformer head in a large labelled X-ray imaging database for KL classification. An automated prediction based on both modalities is generated with a decision-level weighted fusion method producing one unified assessment for deterioration level and surgical prescription. To address the interpretability concerns, the LIME algorithm is used on the imaging data, and the SHAP algorithm is implemented on the clinical data to present explainable insights on the rationale behind each classification. Advances in AI and multimodal approaches, as well as the increasing worldwide burden, necessitate explainable, clinically relevant systems to enable customized treatment decisions and provide precise and consistent estimations.

Related work

AI- and machine learning-based approaches for automatic KOA diagnosis were analysed by the works. Even though Jose et al. presented how prediction of the disease progression is improved when longitudinal clinical data is used, Choi et al. trained an AI-based system that performs automatic KL grades using X-ray images of the knee [3]. The opening of the AI was made higher using explainable AI by Wang et al. and Ahmed & Imran by applying deep learning with SHAP value [4, 5], whereas Touahema et al., Kinger, and Lee et al. improve the classification by utilizing CNNs and transfer learning through various deep learning frameworks [6, 7]. The performance was further analysed by combining explainable AI with fuzzy feature selection from Kokkotis et al. and ensemble CNNs performance shown by Pi et al. [8].

Key Contribution

The primary contribution of this work is the creation of a multimodal framework that combines clinical features and radiographic images for better evaluation of knee OA. With the interpretable nature, it becomes understandable and verifiable the models' decisions by mapping to meaningful patient factors and structural variations of the knee OA.

Method

The Explainable Deep Learning Framework is a hybrid and multimodal approach, which quantifies the knee osteoarthrosis severity and gives surgery recommendations, in particular by incorporating clinical data with imaging. The clinical model, imaging model, and hybrid model are the main three components of the overall system. Demographic and symptoms of knee health are in the clinical dataset, which includes Side, Age, Height, Weight, Maximum Weight, BMI, Frequent Pain, Surgery, Risk, SxKOA, Swelling, Ability to Bend Fully, Symptomatic State, Crepitus, and KOOS Pain Score. The clinical model finds deep and non-linear patterns in the structured data using stacked models of XGBoost, LightGBM, and CatBoost. Important factors like BMI, Frequent Pain, Crepitus, Swelling, and KOOS Pain Score show high relation with OA severity. The data are normalized and encoded before the training. The calculated OA severity and chances of needing surgery are provided by the model. The image model also collects structural and spatial information from knee X-rays using EfficientNetV2L with a Transformer head. Grayscaling , 224x224 pixel scaling, contrast correction, and random augmentation are among the preprocessing steps done to a dataset with a count 5,000 X-ray images (1,000 photos for each KL grade 0–4). The hybrid model allows both modalities to contribute based on their level of identification by combining the prediction probabilities from the clinical and imaging models using a weighted average at the decision level. Explainable AI (XAI) techniques: LIME identifies important areas in the X-ray that affect the imaging model's assessment, whereas SHAP finds the clinical characteristics that most affect OA severity identification. As a result, the suggested approach gives surgical advice along with a complete and understandable explanation of OA severity.

Results

Performance Comparison of ML Models

Clinical Model

Several machine learning models were trained and compared on the clinical dataset using accuracy, precision, recall, and F1-score as shown in the table.

Table 1. Performance Comparison across multiple ML Models

S.No.	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
1	Random Forest	0.88	0.93	0.54	0.68	0.93
2	XGBoost	0.86	0.83	0.47	0.60	0.90
3	LightGBM	0.81	0.75	0.27	0.40	0.85
4	Gradient Boosting	0.79	0.69	0.14	0.24	0.75
5	Logistic Regression	0.78	0.56	0.14	0.22	0.71
6	SVM	0.77	0.57	0.076	0.13	0.66
7	MLP	0.76	0.45	0.29	0.35	0.70
8	KNN	0.74	0.38	0.20	0.26	0.62

The table shows that Random Forest has the highest accuracy (88%) and F1-score (0.68), whereas LightGBM and XGBoost have low recall. Random Forest shows the best individual performance, but overall, the tree-based ensemble models perform better than the other methods. Combining many boosting models into a stacked ensemble model reached 93.3% accuracy and 0.98 ROC-AUC, as shown in the figure. The KOOS Pain Score, BMI, Age, Swelling, and Crepitus were the most significant features, according to the feature-importance analysis, which is in relation their clinical significance to the progress of OA. The choice of the stacked model for the hybrid framework is supported by these results.

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.95	0.93	1485
1	0.95	0.92	0.93	1484
accuracy			0.93	2969
macro avg	0.93	0.93	0.93	2969
weighted avg	0.93	0.93	0.93	2969

Figure 1. Classification Report

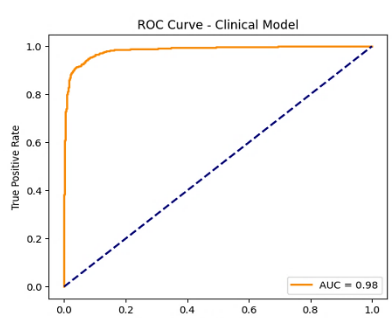


Figure 2. ROC Curve

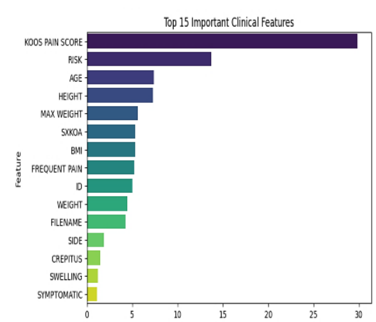


Figure 3. Feature Importance Chart

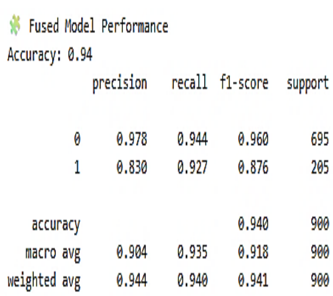
Image Model

As shown in Table 2, accuracy, precision, recall, and F1-score were used to compare the performance of several deep learning models. With an F1-score of 0.764 and an accuracy of 80.7%, EfficientNetV2L with a Transformer Head outperformed the other evaluated models. So, in the suggested explainable hybrid framework, it was chosen as the imaging model for combining with the clinical branch.

Table 2. Performance Comparison across multiple Image Models

S.No.	Model	Accuracy	Precision	Recall	F1-Score
1	Xception	0.540	0.580	0.590	0.550
2	EfficientNetV2B3	0.573	0.561	0.573	0.563
3	EfficientNetV2	0.608	0.596	0.608	0.597
4	EfficientNetB0	0.640	0.630	0.640	0.630
5	Swin Base	0.718	0.722	0.730	0.724
6	Swin Tiny	0.720	0.710	0.720	0.710
7	EfficientNetV2L + Transformer Head	0.807	0.771	0.762	0.764

The fusion model uses weighted decision-level fusion to combine the predictions of the EfficientNetV2L + Transformer Head with the stacked clinical model. The model takes both clinical characteristics and radiographic changes by combining the confidence scores. The hybrid model performed better than the separate models and provided more accurate OA severity prediction, with an AUC of 0.935 and 94% accuracy on a different test set of 900 samples.


Fused Model Performance
Accuracy: 0.94

	precision	recall	f1-score	support
0	0.978	0.944	0.960	695
1	0.830	0.927	0.876	205
accuracy			0.940	900
macro avg	0.904	0.935	0.918	900
weighted avg	0.944	0.940	0.941	900

Figure 4. Fusion Model Performance

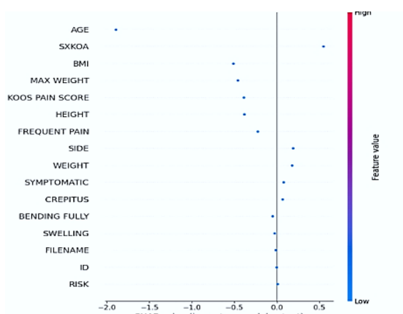


Figure 5. SHAP XAI

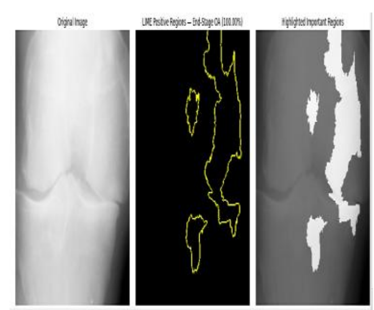


Figure 6.LIME XAI

Explainable AI (XAI) is used by the framework to increase clinician confidence and transparency. BMI, edema, crepitus, and KOOS Pain Score are top factors in SHAP's identification of the clinical characteristics that most affect the prediction. LIME shows the key locations in the X-ray pictures, especially those related to osteophytes, cartilage degeneration, and joint-space constriction. When combined, SHAP and LIME offer both visual and numerical explanations that help clinicians in validating the model's conclusions.

Discussions

The findings show that when clinical and radiographic data are combined, knee OA can be more accurately assessed than using single modality. EfficientNetV2L with a Transformer Head found significant structural information from X-rays, whereas the stacked XGBoost–LightGBM–CatBoost model found clinical and symptom-related patterns. Their weighted decision-level fusion produced an AUC of 0.935 and an accuracy of 94%. By identifying important clinical characteristics and X-ray locations, the application of

SHAP and LIME additionally gave useful explanations and increased the transparency of the model's performance.

Conclusions

For OA severity evaluation and surgical suggestion, the suggested explainable hybrid model effectively combines multimodal clinical and imaging data. A reliable and therapeutically applicable solution is produced by combining ensemble learning, deep learning, decision-level fusion, and XAI. The framework may help with early detection of knee degradation and treatment planning. Future research may concentrate on multi-center and longitudinal validation along with biomechanical, lifestyle, MRI, and rehabilitation data.

References

1. H. C. Chen J, Chen X, Wang T, Li M, Dai H, Shang S, Cheng L, Tang Z, Liu S, "Global burden of knee osteoarthritis from 1990 to 2021: Trends, inequalities, and projections to 2035," *PLoS One*, vol. 20, no. 6, 2025, doi: PLoS One.
2. Y. Dong, Y. Yan, J. Zhou, Q. Zhou, and H. Wei, "Evidence on risk factors for knee osteoarthritis in middle - older aged : a systematic review and meta analysis," *J. Orthop. Surg. Res.*, vol. 6, pp. 1–16, 2023, doi: 10.1186/s13018-023-04089-6.
3. J. C. Choi, M. Y. Jeong, and Y. J. Kim, "Artificial Intelligence Model Assists Knee Osteoarthritis Diagnosis via Determination of K-L Grade," *Diagnostics*, pp. 0–11, 2025, [Online]. Available: https://doi.org/10.3390/diagnostics15101220
4. Z. Wang, Y. Zhou, X. Zeng, Y. Zhou, T. Yang, and K. Hu, "An explainable machine learning-based prediction model for sarcopenia in elderly Chinese people with knee osteoarthritis," *Aging Clin. Exp. Res.*, vol. 37, 2025, [Online]. Available: https://doi.org/10.1007/s40520-025-02931-x
5. R. Ahmed, A. L. I. Shariq, and I. Member, "Knee Osteoarthritis Analysis Using Deep Learning and XAI on X-rays," *IEEE Access*, vol. PP, p. 1, 2024, doi: 10.1109/ACCESS.2024.3400987.
6. M. Touahema, S.; Zaimi, I.; Zrira, N.; Ngote, M.N.; Doulhousne, H.; Aouial, "MedKnee: A New Deep Learning-Based Software for Automated Prediction of Radiographic Knee Osteoarthritis," *Diagnostics*, vol. 14, no. 993, 2024, [Online]. Available: https://doi.org/10.3390/diagnostics14100993
7. Kinger.S, "Deep Learning for Automatic Knee Osteoarthritis Severity Grading and Classification," *Indian J Orthop*, vol. 58, no. 10, pp. 1458–1473, 2024, doi: 10.1007/s43465-024-01259-4.
8. C. Kokkotis, C. Ntakolia, S. Moustakidis, G. Giakas, and D. Tsaopoulos, "Explainable machine learning for knee osteoarthritis diagnosis based on a novel fuzzy feature selection methodology," *Phys. Eng. Sci. Med.*, vol. 45, no. 1, pp. 219–229, 2022, doi: 10.1007/s13246-022-01106-6.
9. D. W. Lee, D. S. Song, H. S. Han, and D. H. Ro, "Accurate , automated classification of radiographic knee osteoarthritis severity using a novel method of deep learning : Plug - in modules," Knee Surg. Relat. Res., pp. 1–11, 2024, doi: 10.1186/s43019-024-00228-3.
10. S. W. Pi, B. D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep - learning networks for automated osteoarthritis grading in knee X - ray images," Sci. Rep., pp. 1–17, 2023, doi: 10.1038/s41598-023-50210-4.