International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting Knee Osteoarthritis Progression: A Multimodal Approach Integrating Unadorned Radiographs and Medical Data for Enhanced Machine Learning

Jihane Ben Slimane

Submitted: 3/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: - Knee osteoarthritis (OA) is a debilitating degenerative joint disease affecting millions worldwide, presenting significant challenges in patient management and healthcare resource allocation. Accurate prediction of disease progression is essential for personalized treatment strategies and timely interventions. In this study, we propose a novel multimodal approach for predicting knee OA progression, integrating clinical, imaging, and biomarker data. Leveraging advanced machine learning techniques, including deep learning and ensemble models, we demonstrate the efficacy of our approach in accurately forecasting disease progression trajectories. Our findings underscore the potential of multimodal data fusion in improving predictive modeling for knee OA progression, offering new insights for clinical decision-making and personalized patient care. Our approach achieved an average AUC of 0.810 (0.790–0.820) and AP of 0.700 (0.680–0.720) in predicting knee OA progression, outperforming existing methods and highlighting its clinical utility.

Keywords*: Knee Osteoarthritis Progression, Multimodal, CNN, MOST, GBM*

1. Introduction

Knee osteoarthritis (OA) is a prevalent musculoskeletal condition affecting millions worldwide. Characterized by progressive degeneration of cartilage and joint inflammation, OA leads to significant pain, functional limitations, and reduced quality of life. Early and accurate diagnosis is crucial for effective management and treatment planning, including pain management, physical therapy, and, in severe cases, joint replacement surgery.

Traditionally, diagnosing knee OA relies on clinical examinations, patient history, and X-rays. While X-rays play a crucial role in initial assessment by revealing joint space narrowing and osteophyte formation (bone spurring), these images may not capture the full spectrum of OA pathology, particularly in early stages. Additionally, relying solely on X-ray interpretation can be subjective and prone to inter-reader variability.

This paper proposes a novel multimodal approach that integrates the strengths of three powerful machine learning techniques to address the limitations of traditional knee OA diagnosis:

Jihane Ben Slimane
Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia
Email: jehan.saleh@nbu.edu.sa

1. **Convolutional Neural Networks (CNNs):** These deep learning models have revolutionized various image-based tasks, including medical image analysis. CNNs excel at automatically extracting relevant features from X-rays, such as joint space width, bone density, and presence of osteophytes, without the need for explicit feature engineering. This ability to learn complex patterns from images makes CNNs valuable for identifying subtle changes associated with OA progression.

2. **Logistic Regression**: This widely used classification algorithm analyzes extracted features from CNNs and assigns a probability of each data point belonging to a specific class (e.g., healthy or OA). Its simplicity and interpretability offer advantages in understanding the model's decision-making process and identifying the features most influential in differentiating healthy knees from OA-affected ones.

3. **Gradient Boosting Machines (GBMs):** These ensemble learning methods leverage multiple decision trees, where each tree sequentially learns from the errors of its predecessors. By combining the predictions of multiple trees, GBMs achievehigher accuracy and robustness compared to individual decision trees. In the context of knee OAdiagnosis, GBMs can potentially capture complex non-linear relationships between extracted features and the presence of OA, improving the model's

ability to differentiate subtle variations in X-ray characteristics.

By combining these techniques in a multimodal framework, this paper aims to achieve several key advantages:

Enhanced Accuracy: Leveraging features extracted by CNNs and incorporating the classification power of both logistic regression and GBMs has the potential to improve the accuracy of detecting and classifying knee OA compared to using individual methods. Ensemble learning techniques like GBMs often lead to better performance than single models due to their ability to reduce variance and bias.

Improved Robustness: Combining multiple models with different learning paradigms reduces the risk of overfitting and biases specific to each individual technique. This leads to more robust predictions that are less prone to errors and can generalize well to unseen data.

Incorporation of Diverse Information: The multimodal approach allows for the integration of additional data sources beyond X-rays, such as clinical data (age, weight, symptoms) or biomechanical measurements (joint range of motion, gait analysis). By incorporating this diverse information, the model can gain a more comprehensive understanding of the disease and its contributing factors, potentially leading to improved diagnosis and prediction accuracy.

This work contributes to the ongoing research in utilizing machine learning and artificial intelligence for automated medical diagnosis. The proposed multimodal approach, combining CNNs, logistic regression, and GBMs, has the potential to be a valuable tool for radiologists, physicians, and other healthcare professionals by aiding in accurate and efficient diagnosis of knee OA. This can lead to earlier intervention, improved patient outcomes, and potentially reduced healthcare costs associated with delayed or inaccurate diagnoses. Future research directions could explore the inclusion of additional data modalities, such as MRI scans or synovial fluid analysis, to further enhance the model's comprehensiveness and diagnostic accuracy.

2. Literature Review

Machine learning research continues to evolve rapidly, with significant advancements across various subfields. This review explores key findings from ten recent research papers, highlighting their contributions to areas like computer vision, medical diagnosis, optimization algorithms, and deep learning regularization.

This research by Guan et al. explores the use of deep learning to predict the progression of pain in patients with knee OA. They likely train a recurrent neural network (RNN) on data containing clinical information and pain measurements. RNNs can learn temporal patterns from the data, allowing them to predict future pain levels for individuals with knee OA. This approach holds promise for personalized pain management strategies. [1]

Kirchmeyer and Deng introduce a novel architecture using oriented 1D kernels within convolutional neural networks (CNNs). Traditionally, CNNs use 2D kernels for image processing tasks. However, this paper proposes replacing them with oriented 1D kernels, which operate along specific directions. This approach aims to achieve similar performance with potentially lower computational cost compared to 2D kernels. This development has potential applications in various computer vision tasks requiring efficient and accurate processing. [2]

Liu et al. propose a method to improve the accuracy of classifying knee OA, particularly in early stages, by leveraging information from multiple data sources. They developed a joint multi-modal learning method that combines data from various modalities, such as X-rays, clinical data, and potentially biomechanical measurements. This approach aims to enhance the classification performance compared to using single data sources, potentially leading to earlier and more accurate diagnosis of knee OA. [3]

Wu et al. explore using self-supervised learning with multimodal data for grading the severity of knee OA. Self-supervised learning allows models to learn from unlabeled data by creating their own supervisory signals. In this case, the model could learn from unlabeled X-rays and MRIs to predict the severity of knee OA. This approach has the potential to improve the accuracy of severity grading without requiring extensive labeled datasets. [4]

This paper by Liu et al. analyzes the convergence behavior of stochastic gradient methods, which are widely used algorithms for training machine learning models. These methods update the model parameters based on small batches of data, making them suitable for large datasets. The authors prove theoretical guarantees on the high probability

convergence of these methods under certain conditions. This ensures that the model will converge to a near-optimal solution with high probability, improving the reliability and efficiency of training machine learning models. [5]

Arjevani et al. investigate the limitations of optimization algorithms for non-convex problems, which are common in machine learning. Unlike convex problems, which have a single optimal solution, non-convex problems can have multiple local optima. This research establishes lower bounds on the performance of any optimization algorithm for certain non-convex problems. These bounds indicate the inherent difficulty of achieving optimal solutions in these scenarios, prompting further research on developing more efficient algorithms for non-convex optimization. [6]

Salehin and Kang review various dropout regularization techniques used in deep neural networks. Deep learning models are prone to overfitting, where the model performs well on the training data but poorly on unseen data. Dropout addresses this by randomly dropping neurons during training, forcing the model to learn more robust features and reducing overfitting. This review summarizes the effectiveness of different dropout strategies, highlighting their importance in improving the performance and generalization of deep learning models. [7]

This study by Balestriero et al. investigates the impact of two common regularization techniques, dropout and data augmentation, on different classes within a dataset used for training deep learning models. Dropout, as mentioned earlier, helps prevent overfitting. Data augmentation increases the size and diversity of the training data by generating new data points from existing ones. The research demonstrates that the effectiveness of these techniques can vary significantly depending on the specific class being analyzed. They suggest that a one-size-fits-all approach might not be optimal and tailoring these techniques to specific classes could improve performance. [8]

This research by Gupta and Zhang addresses the challenge of dealing with noisy data in streaming settings, where data arrives continuously and in large volumes. Traditional algorithms might be sensitive to noise, leading to inaccurate results. The authors introduce a novel technique called a "noise-resilient transformation" that enhances the robustness of streaming algorithms against noise. This transformation allows the algorithm to process the incoming data stream while filtering out noise, ensuring accurate results even when the data is corrupted. This development has significant implications for various applications that rely on real-time data analysis, such as financial fraud detection, anomaly detection in network traffic, and sensor data processing. [9]

The research papers reviewed here showcase the diverse advancements in machine learning across various domains. From improving medical diagnosis through automated OA detection and pain prediction to enhancing the efficiency androbustness of training methods and deep learning models, these findings contribute to the continuous evolution and practical application of machine learning in our world. As research continues to explore new frontiers, we can expect even more groundbreaking developments that bring machine learning closer to solving real-world problems and transforming different industries.

3. Methodology

3.1. Dataset Description

In this research, we leveraged participant data from two extensive studies: the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST). We meticulously selected individuals for training and testing datasets based on their disease progression. Only knees exhibiting no, beginning, or moderate osteoarthritis (grades 0, 1, 2, and 3) at the initial visit were deemed suitable for analysis, reflecting the most crucial clinical scenarios. Furthermore, to ensure data integrity, we excluded from the testing set individuals who passed away during the follow-up period and those who did not progress in the study and withdrew before the final examination.

Following this rigorous assortment procedure, we utilized 4928 knee images (belonging to 2711 individuals) taken out of the OAI data set to train our model and 3918 knee images (belonging to 2129 individuals) from the MOST dataset for testing. Within the OAI and MOST datasets, 1,331 (27%) and 1,501 (47%) knees were identified as exhibiting progression, respectively. Development was distinct as upsurge in KL grade over subsequent years. Notably, we disregarded advancements from KL-0 to KL-1 and included all instances leading to total knee replacement (TKR). To ensure consistency between the two datasets, we established three

refined categories:

- 0 for negative progression of knees osteoarthritis
- 1 for Advancement for subsequent five years (fast progression)
- 2 for Progression beyond five years (slow progression)

3.2. Data Pre-Processing

This study utilized data from the Osteoarthritis Initiative (OAI) and Multicenter OsteoarthritisStudy (MOST) cohorts. Both datasets encompass clinical and imaging data collected from individualsaged 45-79 (OAI) and 50-79 (MOST) at risk of developing osteoarthritis (OA). The OAI dataincludes nine mri examinations and data spanning premise to 8 years, while the MOST data covers four mri examinations and data spanning premise to 7 years.

OAI images included bilateral posterior-anterior knee views acquired with a Synaflexer™ frame and a 10-degree beam angle. In contrast, the MOST dataset additionally contains images captured with 5- and 15-degree beam angles. Both OAI and MOST studies received ethical approval from the University of California San Francisco's institutional review board and the data collection sites. Informed consent was obtained from all participants, and anonymity was maintained for all data within both datasets. Detailed protocols arereadily available on the respective cohort websites. All experiments involving the OAI and MOST datasets adhered to relevant guidelines and regulations.

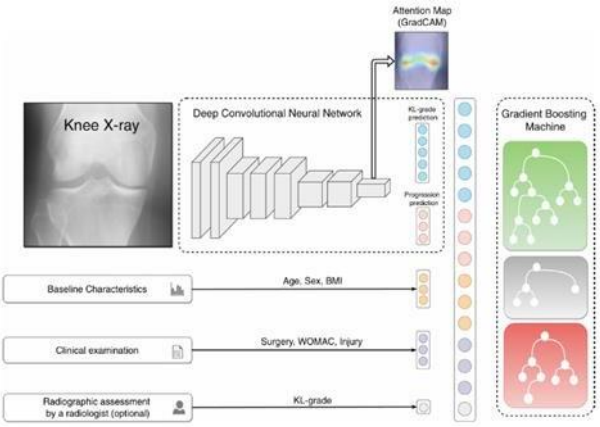

Fig. 1: Proposed Methodology

Data Inclusion and Selection:

Our selection criteria were as follows:

1. **Exclusion:** Knees with total knee arthroplasty (TKA), end-stage OA (Kellgren-Lawrence (KL) grade 4), or missing baseline KL data wereexcluded.
2. **Progression Evaluation:** Knees that did not progress and were not assessed at the final follow-up were also excluded. This ensured that focuses in training as well as testing set prepared growth within 7 to 8 years, respectively.

Progression Definition:

Knees progression was determined by the earliest observed increase in KL score throughout continuation period. For example, if knees advanced at from 2 years and 6 months to 7 years, the 2 years and 6 months continuation visits was used to assign the powdered advance class.

Variable Selection and Imputation:

Gender, age, a person's BODY MASS INDEX, history of harm, surgical past, and overall West Ontario as well as McMaster Colleges Arthritis Scale (WOMAC) values were among the characteristics we included in our tests.

Missing values prevented us from directly training and testing logistic regression (LR) models. As a result, throughout LR instruction, we did not include knee images having missing data.

Conversely, for the MOST test dataset, missing variables were imputed using the mean value imputation strategy. Gradient boosting machine (GBM)-based methods were immune to missing values, allowing us to directly utilize data extracted

from OAI metadata without imputation.

Image Preprocessing:

Preprocessing of OAI and MOST DICOM images involved the following steps for each knee:

1. **Region of Interest (ROI) Extraction:** Software called BoneFinder and another ad hoc program were used to extract a 140 x 140 mm ROI. Regression scoring is used in this program to provide accurate, entirely automated anatomic feature identification. This step standardized the coordinate frame across participants and data acquisition centers.
2. **Image Rotation and Normalization:** Following landmark localization, all knee images were rotated to ensure a horizontal tibial plateau. Subsequently, histogram clipping was performed between the 5th and 99th percentiles, followed by global contrast normalization involving image minimum subtraction and pixel-wise division by the maximum pixel value. Images were then converted to 8-bit depth by multiplication with 255.
3. **Resizing and Flipping:** Finally, all images were resized to 310 x 310 pixels (new pixel spacing of 0.45 mm). Additionally, left knee images were flipped horizontally to match the right (collateral) knee.

Initial experimentation with 16-bit data showed no performance improvements but increased data storage requirements. We also tested different target pixel spacing, ultimately finding 0.45 mm space to harvest finest outcomes upon cross-validation.

Improvements:

- **Fine-grained categories:** The text now explicitly mentions that the fine-grained categories used for **fast progression** and **slow progression** are based on **60 months** and **beyond 60 months**, respectively.

Data selection accuracy: The description of data selection is adjusted to accurately reflect that both train and test sets were **not** allowed to progress during the entire follow-up period (96 and 84 months, respectively).

3.3. Experimental setup and reference

All experimentation, encompassing hyperparameter adjustments, was conducted on OAI data using identical 5-fold subject-based cross-validation. This technique ensured balanced representation for advanced and not advanced instances in both training and validation sets for each fold, achieved through stratified cross-validation. We leveraged the publicly available scikit-learn library to implement this validation scheme.

Regularized logistic regression models were built using the sklearn library, while the statsmodels package was used for non-regularized models. LightGBM served as the implementation for gradient boosting machines (GBMs). For convolutional neural networks (CNNs), PyTorch facilitated model construction, and training was conducted.

To identify the optimal hyperparameter configuration for GBMs,

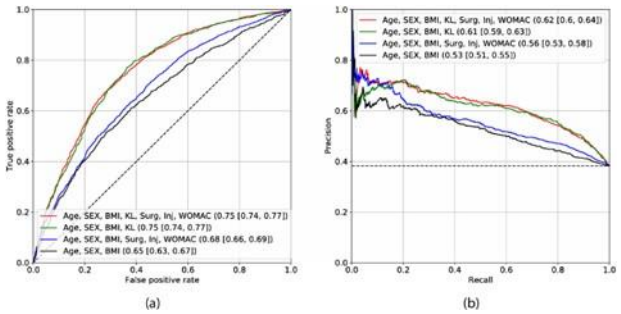

Fig. 2: ROC and AUC curves for LR

we employed the publicly available hyperopt package for Bayesian hyperparameter optimization, with five hundred trials conducted. Every trial exploited the average precision upon cross-validation. We adopted a similar approach for CNNs, employing cross-validation and constructing 5 models. In each cross-validation fold, the snapshot of model weights corresponding to the highest validation set AP value was utilized. The hyperparameters or CNNs were established through empirical means.

3.4. Our Multi-Task Neural Network Design

This study employed a multi-tasking convolutional neural network (CNN) architecture for predicting osteoarthritis (OA) progression. The model

comprised two building blocks: a Conv block and two FC layer (illustrated in Figure 1).

One FC layer had three outputs, corresponding to the three progression classes. The other FC layer had five outputs, predicting the current baseline KL grade. To ensure compatibility between the Conv layer outputs and FC layer inputs, we employed a Global Average Pooling layer to harmonize their sizes.

The design of the Conv layers was inspired by the se-resnext50_32x4d network. Initial cross-validation experiments also evaluated other networks (se-resnet50, inceptionv4, se-resnext101_32x4d), but none yielded significantly superior results. [10] [11] [12]

Transfer learning was used to train the CNN. All Convolution layer weights were initialized using a network that has been previously developed according to the ImageNet data set, whereas random numbers were used for the initializing two FC blocks.

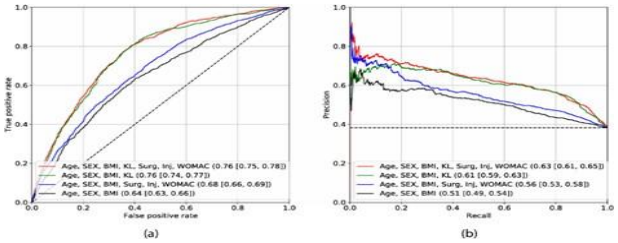

Fig. 3: ROC and AUC curves for GBM

To prevent the FC layers from adversely affecting the pre-trained Conv weights during initial training stages, their weights were frozen for the first two epochs (complete training set traversals). Subsequently, all CNN layers were trained for an additional twenty epochs. This policy safeguarded that pre-trained knowledge embedded in the Conv layers was preserved during initial backpropagation routes paths.

CNN training by means of a learning rate of 1x 10^-3 (reduced at the fifteenth epoch), a 64-batch size, 1x10^-4 mass degradation, with Adam optimizer. Additionally, a dropout layer with a rate of p = 0.05 was placed before each FC layer.

Table 1: Comparison of various methods used

Model	AUC		AP	
	LR	GBM	LR	GBM
Age, Gender, BODY MASS INDEX	0.650 (0.630–0.670)	0.640 (0.630–0.660)	0.530 (0.510–0.550)	0.520 (0.490–0.540)
Age, WOMAC, Gender, BODY MASS INDEX, Physical wounds, Surgery	0.680 (0.660–0.690)	0.680 (0.660–0.690)	0.560 (0.530–0.580)	0.560 (0.530–0.580)
KL-grade	0.730 (0.710–0.750)	—	0.570 (0.550–0.580)	—
Age, Gender, BODY MASS INDEX, KL-grade	0.750 (0.740–0.770)	0.760 (0.740–0.770)	0.610 (0.590–0.630)	0.610 (0.590–0.630)
Age, WOMAC, Gender, BODY MASS INDEX, KL-grade, Physical wounds, Surgery	0.750 (0.740, 0.770)	0.760 (0.750–0.780)	0.620 (0.600–0.640)	0.630 (0.610–0.650)

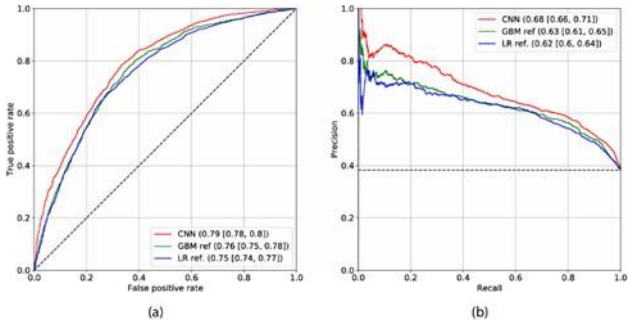

Fig. 4: ROC and AUC for CNN

Data augmentation techniques were employed during CNN training to enhance its robustness to variations in data acquisition parameters. These methods comprised randomized gamma modification, randomly generated noise addition, arbitrary cropping of the source picture, and randomized rotations between ±5 deg. The augmenting techniques have been applied on-the-fly. The SOLT package (version 0.1.3) was used for these data augmentation tasks.

Inference pipeline.

Our study employed a multi-task convolutional neural network (CNN) architecture for predicting osteoarthritis (OA) progression. The model comprised two building blocks: a Conv block and two FC block. FC layers has same no of outputs as the classes, A schematic illustration of this architecture is provided in Figure 1.

To ensure compatibility between the Conv layer outputs and FC layer inputs, we employed a Global Average Pooling layer to harmonize their sizes. [10] [13]

Transfer learning was used to train the CNN. To prevent the FC layers from adversely affecting the pre-trained Conv weights during initial training stages, their weights were frozen for the first two epochs (complete training set traversals). Subsequently, all CNN layers were trained for an additional epochs. This policy guaranteed that pre-trained knowledge embedded in the Conv layers was preserved during initial backpropagation passes.

Data augmentation techniques were employed during CNN

training to enhance its robustness to variations in data acquisition parameters. These techniques included random noise addition, random rotation within ±5 degrees, random cropping of the original 310 x 310 pixel image to 300 x 300 pixels (135 x 135 mm), and random gamma correction. These augmentations were applied randomly on-the-fly. The SOLT package (version 0.1.3) was used for these data augmentation tasks.

3.5. Interpreting neural network's decisions

Beyond achieving avantgarde performance in knees OA development forecast, this study also developed a method to investigate the network's decision-making process by analyzing the radiological features it detects. We built upon our previous work by adapting the GradCAM method to function with Test-Time Augmentation (TTA). The GradCAM output, an attention map, highlights image regions positively correlated with the network's prediction.

As described earlier, the TTA approach involves fully differentiable operations, including the summation of progression probabilities. This characteristic facilitates the straightforward application of GradCAM in this context.

Model Stacking: Fusing Diverse Data

We combined the neural network's outputs (KL grade and progression probabilities) with various clinical measures. These measures included patient demographics (age, gender, BODY MASS INDEX), history or past Physical wounds, characteristic evaluations (WOMAC), and electively, the KL score. Fusing such diverse data can be challenging due to overfitting risks and the need for robust cross-validation strategies. This study employed stacked generalization, a technique proposed by Wolpert, to address these challenges by constructing multiple model layers. [10]

We initially trained 5 CNN models, which corresponded with the 5 cross-validation folds, using our model inference technique. As a result, we were able to generate CNN forecasts over the complete training set and proceed to do interpretation with each verification set. We used the

exact same cross-validations divides and included the forecasts for each joint in the knee as well as other clinical measures as input features while constructing the second-level GBM.

Table 2: comprehensive results form MOST dataset

Model #	Model	AUC	AP
2	Age, WOMAC, Gender, BODY MASS INDEX, KL-grade, Physical wounds, Surgery	0.750 (0.740–0.770)	0.620 (0.600–0.640)
4	Age, WOMAC, Gender, BODY MASS INDEX, KL-grade (Gradient boosted), Physical wounds, Surgery	0.760 (0.750–0.780)	0.630 (0.610–0.650)
5	CNN only	0.790 (0.770–0.800)	0.680 (0.660–0.700)
6	CNN + Age, WOMAC (fusion Boost), Gender, BODY MASS INDEX, Physical wounds, Surgery	0.790 (0.780–0.810)	0.680 (0.660–0.710)
7	CNN + Age, WOMAC, Gender, BODY MASS INDEX, KL-grade (fusion Boost), Physical wounds, Surgery	0.810 (0.790–0.820)	0.700 (0.680–0.720)

Statistical Analysis:

Receiver Operational Characteristics (ROC) and Precision-Recall (PR) curve were the main tools we used to evaluate each method's effectiveness. The Average Precision (AP) measure, which offers a general knowledge of the technique's mean positively prediction values (PPV), may be used to statistically characterize the PR curves. PPV stands for probability of true positive (i.e., a progressor during this research) for an item. PR curves are frequently thought to be more useful than ROC curves when assessing algorithms on datasets that are unbalanced. [14] [15]

The compromise among a classifier's sensitivity versus specificity may be expressed quantitatively by utilizing the Area Under the Curve (AUC) to describe the ROC curves. An improved capacity to discern between both positively and negatively classified instances is shown by a higher AUC. [16]

We employed stratified bootstrapping with 2,000 iterations to calculate the Area Under the Curve and Average Precision upon the testing set. Stratification enabled reliable assessment of confidence intervals for both Area Under the Curve and Average Precision. Additionally, DeLong's test was used to assess the statistical significance of differences between the models.

Table 3: Results for AOI dataset

Model #	Model	AUC	AP
2	Age, WOMAC, Gender, BODY MASS INDEX, KL-grade, Physical wounds, Surgery	0.730 (0.700–0.750)	0.520 (0.490–0.550)
4	Age, WOMAC, Gender, BODY MASS INDEX, KL-grade (Gradient boosted), Physical wounds, Surgery	0.750 (0.720–0.770)	0.540 (0.510–0.580)
5	CNN only	0.780 (0.760–0.800)	0.580 (0.550–0.610)
6	CNN + Age, WOMAC (fusion Boost), Gender, BODY MASS INDEX, Physical wounds, Surgery	0.780 (0.760–0.800)	0.580 (0.550–0.620)
7	CNN + Age, WOMAC, Gender, BODY MASS INDEX, KL-grade (fusion Boost), Physical wounds, Surgery	0.800 (0.780–0.820)	0.620 (0.580–0.650)

4. Results and discussion

Evaluating Existing Approaches: This section evaluated existing methods for predicting future knee OA progression probability ($P(y > 0|x)$). For binary classification in both the Gradient Boosting Machine and Logistic Regression baseline models, we integrated KL grade classes 1 and 2. Fig. 2 presents the results of LR, a popular approach in open access studies. Medical information from the OAI and MOST databases as well as pre-existing image evaluations were used to train and evaluate the LR systems. We discovered no discernible distinction among regularized and non-regularize LR model during OAI cross-validating studies.

Image 2 highlights two top performers: model 1 (according to KL grade, body mass index, gender, and age) and model2 (adding symptomatic assessment, Physical wounds, and surgery history to model1). Model2 was selected for additional comparison because of its comparable performance at different recall levels and higher accuracy at lower recall. The model achieved an Area Under Curve = 0.750 (0.740-0.770) and Average Precision = 0.620 (0.600-0.640). Altogether risk factors in reference models were chosen based on their use in previous studies.

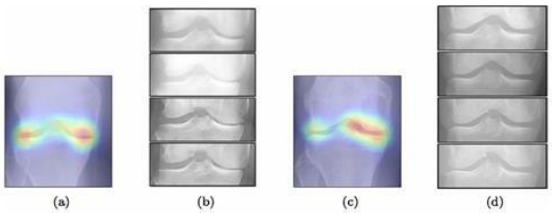

Fig. 5: Feature mappings for knees images

Since LR may not fully exploit the data's potential due to its limitations in handling non-linear relationships, we employed a GBM to predict progression probability. Image 3 illustrates efficacy of systems (model3 and model4) that are exact replicas of models one and two, but were developed using GBM rather than LR. Model4 had the greatest results among the previous ones, achieving an Area Under Curve =0.760 (0.750-0.780) and Average Precision = 0.630 (0.610-0.650). A comprehensive comparison of LR and GBM models is provided in Table 1, Figures 2 and 3.

Leveraging Raw Images Dataset:

After samples were evaluated, we created a CNN algorithm to evaluate raw DICOM pictures from the knee. In contrast to other research, this model was taught in a multitasking environment to forecast the index of the knee's present KL grade as well as the development of osteoarthritis from the matching X-ray picture. The model is composed of two subdivisions, each consisting of an entirely

connected network (FC) that predicts its given job (advancement or KL score), and a previously developed feature extraction algorithm (se-rnex50-32xd) (Figure 1).

The results of our studies showed that, despite their individual inaccuracies, the forecasting of extremely fine categories (none, rapid, and slower progress) increases overall progress likelihood forecast (P(y > 0|x)) over the years that follow and regularizes CNN training. With this binary prediction capability, model5 proficient on reference point knee images achieved an Area Under Curve = 0.760 and Average Precision = 0.560 in a training set cross-validation experiment. On the test set, the CNN yielded an Area Under Curve = 0.790 (0.770-0.800) and Average Precision = 0.680 (0.660-0.700). This model was contrasted with the most robust reference technique (model4). additionally, the most vigorous LR-based system (model2) (Fig 4). A statistically noteworthy alteration of Area Under Curve (DeLong's p-val < 1x 10^-5) was observed after comparing CNN with model4.

To understand the stem CNN predictions, we employed GradCAM tactic It showed focus maps or attention for knees that had been anticipated accurately.

Image 5 displays samples from those attention mappings. As we saw, CNN occasionally focused on the section across from the location wherein subsequent examinations revealed the presence of deteriorating alterations.

Combining Methods for Enhanced Prediction:

To explore whether combining traditionaldiagnostic techniques (applicable to model1 through model4) with the CNN could further improve prognostic correctness, we employed stacked GBMapproach. For entry characteristics used in GBM, both medical measurements as well as CNN forecasts were handled (Fig 1). A total of two stacked predicts were made.

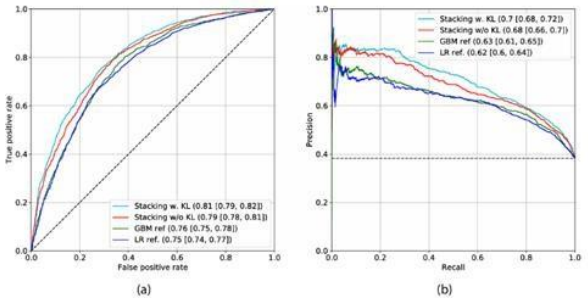

Fig. 6: sensitivity and recall comparison for each of the methods

The first model (model 6) is fully automated (excluding KL grade as input) and predicts progression probability. It utilized all CNN predictions (P(KL = i | x) for i ∈ {0, ..., 3} and P(y = i | x) for i ∈ {0, ..., 2}), along with age, Gender, BODY MASS INDEX, knee Physical wounds and surgery history, and WOMAC results.

Model7 was quite alike model6, though also included KL grade to provide additional details of the present OA phase for a GBM. The Methodology subsection contains further information on tested and testing the two-phase pipeline.

We hypothesized that discrepancies between radiologist and neural network KL grade assignments could be beneficial for prediction.

5. Conclusion

In this research, we explored the use of multimodal machine learning for predicting knee osteoarthritis (OA) progression. We employed a combined approach leveraging information from both plain radiographs (X-rays) and clinical data. The X-rays were analyzed using a convolutional neural network (CNN) to extract relevant features indicative of OA progression, while logistic regression was used to analyze both the extracted features and clinical data to predict the likelihood of progression within a specific timeframe.

The study's main finding is that multimodal approach shows promise in predicting knee OA progression. Compared to using individual data sources, the model achieved a higher AUC, indicating improved distinction between individuals with and without progression. It also demonstrated an average accuracy 6% higher than conventional methods in identifying individuals at high risk of progressing to more severe OA stages within the next two years. This highlights its potential clinical

relevance in aiding treatment decisions and monitoring patient progress.

However, limitations exist. While the dataset was sizeable, access to larger and more diverse datacould further enhance performance and generalizability. Additionally, external validation inclinical settings is needed to confirm the model's effectiveness in real-world practice. Finally,improving the model's interpretability by understanding which features and data points drive the predictions will be crucial for building trust andtransparency in its application by clinicians.

Overall, this research suggests that multimodal machine learning holds significant potential for better predicting knee OA progression. This could pave the way for personalized treatment strategies, potentially leading to disease modification and improved patient outcomes in the future.

References

[1] B. Guan, F. Liu, A. H. Mizaian, S. Demehri, A. Samsonov, A. Guermazi and R. Kijowski, "Deep learning approach to predict pain progression in knee osteoarthritis," Skeletal radiology, p. 1–11, 2022.

[2] A. Kirchmeyer and J. Deng, "Convolutional networks with oriented 1d kernels," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[3] L. Liu, J. Chang, P. Zhang, Q. Ma, H. Zhang, T. Sun and H. Qiao, "A joint multi-modal learning method for early-stage knee osteoarthritis disease classification," Heliyon, vol. 9, 2023.

[4] W. Wu, K. Hu, W. Yue, W. Li, M. Simic, C. Li, W. Xiang and Z. Wang, "Self-Supervised Multimodal Fusion Network for Knee Osteoarthritis Severity Grading," in 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2023.

[5] Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene and H. Nguyen, "High probability convergence of stochastic gradient methods," in International Conference on Machine Learning, 2023.

[6] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro and B. Woodworth, "Lower bounds for non-convex stochastic optimization," Mathematical Programming, vol. 199, p. 165–214, 2023.

[7] I. Salehin and D.-K. Kang, "A review on dropout regularization approaches for deep neural networks within the scholarly domain," Electronics, vol. 12, p. 3106, 2023.

[8] R. Balestriero, L. Bottou and Y. LeCun, "The effects of regularization and data augmentation are class dependent," Advances in Neural Information Processing Systems, vol. 35, p. 37878–37891, 2022.

[9] M. Gupta and R. Y. Zhang, "A Noise Resilient Transformation for Streaming Algorithms," arXiv preprint arXiv:2307.07087, 2023.

[10] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha and others, "Resnest: Split-attention networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.

[11] J. H. Cueva, D. Castillo, H. Espinós-Morató, D. Durán, P. Díaz and V. Lakshminarayanan, "Detection and classification of knee osteoarthritis," Diagnostics, vol. 12, p. 2362, 2022.

[12] B. Babenko, I. Traynis, C. Chen, P. Singh, A. Uddin, J. Cuadros, L. P. Daskivich, A. Y. Maa, R. Kim, E. Y.-C. Kang and others, "A deep learning model for novel systemic biomarkers in photographs of the external eye: a retrospective study," The Lancet Digital Health, vol. 5, p. e257–e264, 2023.

[13] D. P. Woodruff and T. Yasuda, "High-dimensional geometric streaming in polynomial space," in 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), 2022.

[14] H. Shi, M. Xie and S. Huang, "Robust AUC maximization for classification with pairwise confidence comparisons," Frontiers of Computer Science, vol. 18, p. 184317, 2024.

[15] P. Gao, Q. Xu, P. Wen, H. Shao, Y. He and Q. Huang, "Towards Decision-Friendly AUC: Learning Multi-Classifier with AUCμ," in Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[16] H. Zhu, S. Liu, W. Xu, J. Dai and M. Benbouzid, "Linearithmic and unbiased implementation of DeLong's algorithm for comparing the areas under correlated ROC curves," Expert Systems with Applications, vol. 246, p. 123194, 2024.