AI-ENHANCED DIAGNOSIS AND SEVERITY PREDICTION OF KNEE OSTEOARTHRITIS

Gopalakrishna Reddy Manukonda
Dept. of Eng. Education, MS AIS
University of Florida
Gainesville, Florida, USA
manukonda.g@ufl.edu

Teja Kolla
Dept. of Eng. Education, MS AIS
University of Florida
Gainesville, Florida, USA
teja.kolla@ufl.edu

Andrea Ramirez Salgado, Ph.D.
Dept. of Eng. Education, Professor
University of Florida
Gainesville, Florida, USA
aramirezsalgado@ufl.edu

***Abstract*—Osteoarthritis (OA) represents a prevalent type of knee arthritis, leading to considerable disability and posing a substantial threat to the quality of life for affected individuals. This degenerative joint condition manifests through symptoms such as joint stiffness, pain, and functional limitations, impacting millions globally. Diagnosis typically involves a thorough examination of the patient's medical history and various joint screening tests, including radiographs, magnetic resonance imaging (MRI), and computed tomography (CT) scans.**

To address the imperative need for early detection and severity prediction of osteoarthritis, we propose the implementation of a Deep Learning model. This model aims to analyze X-ray images of knee joints, providing prompt and accurate readings to patients through a user-friendly web application built on the streamlit framework. The core objective of this project is to anticipate the severity of knee osteoarthritis based on a patient's X-ray image, employing an efficient deep learning model. The proposed model specializes in classifying knee joint alignments and grading bone positions, facilitating the detection and prediction of osteoarthritis severity in X-ray images.

Keywords—Osteoarthritis prediction, Deep Learning model, X-Ray image analysis, Streamlit web application, Joint alignment classification, GradCam, Docker.

Report organization:

I. INTRODUCTION

Knee osteoarthritis (OA) is a prevalent degenerative joint disease characterized by the gradual deterioration of cartilage in the knee joint, leading to pain, stiffness, and reduced mobility. Early detection and accurate assessment of OA severity are crucial for effective management and timely intervention to alleviate symptoms and prevent disease progression. In this context, leveraging medical imaging modalities such as X-rays and CT scans, along with cutting-edge deep learning techniques, holds immense promise in revolutionizing how knee OA is diagnosed and managed.

Our project endeavors to harness the power of deep learning models to enhance knee OA detection and severity prediction using radiographic data obtained from X-rays and CT scans. By training sophisticated neural networks on large datasets of knee images, we aim to develop a robust system capable of automatically analyzing knee joint alignments and radiographs to identify the presence of OA and assess its severity with high accuracy. This approach offers a non-invasive and efficient means of diagnosing OA, enabling healthcare providers to initiate appropriate interventions promptly.

To facilitate seamless integration into clinical practice, we will develop a user-friendly web application utilizing the deep learning models to deploy it with streamlit and docker. This application will serve as a convenient platform for healthcare professionals to upload knee X-rays and CT scans, which will undergo automated analysis by our deep learning model in real-time. By providing rapid and reliable assessments of OA presence and severity, our application empowers clinicians with actionable insights to guide treatment decisions and optimize patient care outcomes. Through this project, we aim to significantly advance the field of knee OA diagnosis and management, offering a transformative tool that enhances diagnostic accuracy, facilitates early intervention, and improves the quality of life for individuals affected by this debilitating condition

II. PURPOSE

Many users lack awareness about joint pains and related ailments. While some may have some understanding, they often struggle to gauge the disease's severity or even recognize its presence. Furthermore, not everyone can afford advanced diagnostic methods like MRIs and CT scans, though many can manage X-Rays. However, X-Rays may not always offer precise information about the severity of Knee Osteoarthritis (OA). There's a pressing need for a user-friendly solution that's accessible to all. Our objective is to create a straightforward, user-friendly, efficient, and mobile application that can automatically detect the presence of Knee OA. Additionally, the app will educate users about the disease's severity and provide essential recommendations for further diagnosis.

III. SCOPE

Despite the numerous technological advancements in the medical field for detecting diseases associated with joint pain, there are still limitations. Many users are unfamiliar with X-

Rays and similar scans, and some find it challenging to determine the extent of their pain from these scans. Thus, there's an urgent need to develop a web application that can identify and display the severity of joint-related diseases and offer appropriate recommendations to users.

In this project, we explored the ResNet (Residual Network) architecture, based on CNN, to automatically assess the severity of Osteoarthritis in the knee joint using X-Ray images uploaded by the user. This application serves as an efficient alternative for quickly detecting joint-related diseases and provides users with prompt, effective, and valuable suggestions for further diagnosis.

Our future goal is to develop a comprehensive web application that can accurately detect various joint pains (such as in the neck, back, and knee) from X-Ray images of any quality.

IV. Related work

The paper *"Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data"* explores the integration of both radiographic and clinical data for predicting the progression of knee osteoarthritis (OA) using machine learning techniques. By combining these two distinct data types, the study highlights the potential of multimodal models to improve prediction accuracy compared to models using a single data source. The authors employ various machine learning algorithms, such as support vector machines (SVM) and random forests, to develop a robust framework that accounts for both the structural changes observed in plain radiographs and the clinical variables related to the patient's condition. The results demonstrate the effectiveness of multimodal models in predicting OA progression, offering a promising approach for early intervention and personalized treatment strategies. This research underscores the importance of integrating diverse data types to enhance predictive models for chronic diseases like knee OA.

The paper *"XNet: A Convolutional Neural Network (CNN) Implementation for Medical X-Ray Image Segmentation Suitable for Small Datasets"* introduces XNet, a novel convolutional neural network designed to improve medical image segmentation, particularly in scenarios with limited data. XNet addresses the challenge of training deep learning models on small datasets, a common issue in medical imaging, by incorporating advanced techniques such as transfer learning and data augmentation. The model is specifically optimized for X-ray images, demonstrating its ability to accurately segment regions of interest even with sparse training samples. Through rigorous evaluation, the authors show that XNet outperforms traditional CNN architectures and can achieve high segmentation accuracy, making it a promising tool for medical applications where annotated data is scarce. This work contributes to enhancing the applicability of deep learning in medical imaging, offering solutions for scenarios with limited labeled data.

The paper *"V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation"* introduces V-Net, a fully convolutional neural network (FCN) designed for the segmentation of volumetric medical images, such as 3D MRI scans. V-Net employs a 3D architecture to directly learn spatial features from volumetric data, providing a more efficient and accurate segmentation approach compared to traditional methods that rely on slice-by-slice processing. The model utilizes a U-Net-like structure with volumetric convolutions and an innovative loss function to improve segmentation performance, particularly in medical images with complex structures. The authors demonstrate V-Net's effectiveness in segmenting organs and tumors, showing that it outperforms other 3D segmentation models in terms of accuracy and computational efficiency. This work significantly contributes to the field of medical image analysis, particularly in improving segmentation tasks for volumetric data with a more robust and scalable deep learning approach.

The study "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation" proposed a convolutional neural network (CNN) model for detecting knee OA from X-ray images. The model achieved high accuracy in identifying early signs of OA, outperforming traditional machine learning models by incorporating image preprocessing techniques.

The study "Prediction of knee osteoarthritis progression using radiological descriptors obtained from bone texture analysis and Siamese neural networks: data from OAI and MOST cohorts" developed a CNN model to predict the progression of knee OA from X-ray images. The model was able to detect subtle changes in the joint and predict future progression with high accuracy, offering a promising approach for early intervention.

V. Dataset

A dataset is a structured collection of data typically organized in a tabular format, where each row represents a specific observation or instance, and each column represents a particular variable or attribute. These datasets are integral to data management in the realm of data analytics.

Various types of datasets exist, including:

- Numerical: Consisting of numerical values or measurements.
- Bivariate: Involving two variables and their relationships.
- Multivariate: Encompassing multiple variables with complex relationships.
- Correlational: Focused on correlations between variables.
- Categorical: Comprising categorical or qualitative data.

We have obtained a Knee Osteoarthritis dataset from Kaggle, available at https://www.kaggle.com/datasets/shashwatwork/knee-osteoarthritis-dataset-with-severity. This dataset contains approximately 8000 X-ray images of knee joints, accompanied by severity grading for Knee Osteoarthritis. The severity grading is categorized as follows:

- Grade 0 (Healthy): Images depicting a healthy knee.
- Grade 1 (Doubtful): Indicating doubtful joint narrowing with possible osteophytic lipping.

- Grade 2 (Minimal): Showing definite presence of osteophytes and potential joint space narrowing.
- Grade 3 (Moderate): Featuring multiple osteophytes, clear joint space narrowing, and mild sclerosis.
- Grade 4 (Severe): Displaying large osteophytes, significant joint narrowing, and severe sclerosis.

To ensure accurate prediction of Knee Osteoarthritis severity, the dataset is partitioned into three subsets:

- Training: Comprising 70% of the dataset for model training.
- Validation: Consisting of 10% of the dataset for model validation.
- Testing: Including 20% of the dataset for evaluating model performance.

This division aids in effectively training, validating, and testing predictive models for assessing Knee Osteoarthritis severity.

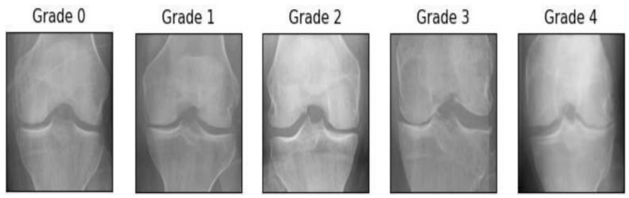


Fig 1. Different KL grades of the X-Ray data in the dataset

Train Dataset (70%): This subset comprises 5781 images, which represent 70% of the total 8000 images in the dataset. These images are utilized primarily during the model training phase to ensure the model learns from a diverse range of examples, leading to accurate predictions of severity levels.

Valid Dataset (10%): The validation set consists of 826 images, accounting for 10% of the total dataset. These images serve to evaluate the performance of the model across all severity categories, aiming for maximum accuracy and robustness in predictions.

Testing Dataset (20%): The testing dataset comprises the remaining 1656 images, constituting 20% of the overall dataset. These images undergo rigorous data processing and image recognition techniques to assess the model's accuracy comprehensively. The testing phase aims to provide optimal results to users by verifying the model's performance under various conditions

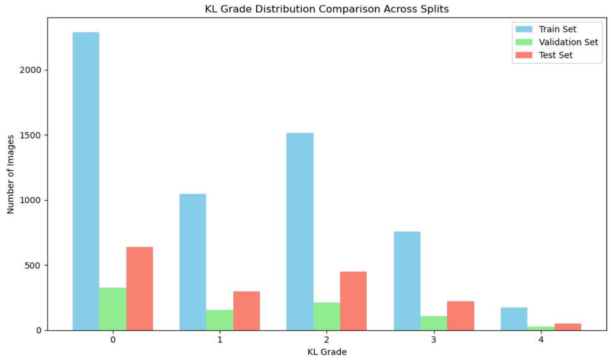


Fig 2. Composition of images across train/test/val splits

VI. System design and implementation

The AI system for diagnosing and predicting the severity of knee osteoarthritis utilizes a modular computing architecture. The system is designed to process X-ray images through a cloud-deployable deep learning model, providing real-time severity assessments via an interactive web application. Key components include:

A. Data Ingestion Layer:

The data ingestion layer is the foundational component of the AI system, responsible for collecting, organizing, and storing the data required for training and deploying the model. Collects and stores X-ray images and related metadata.

This layer ensures the data is properly categorized and securely stored, making it readily available for the preprocessing and model training stages. Key considerations include ensuring data integrity, managing access controls, and complying with regulations like HIPAA for sensitive medical information.

B. Preprocessing Unit:

The preprocessing unit is a critical system component that ensures the X-ray images are in an optimal format for deep learning analysis. It involves several steps – Notmalization, Augmentation, and Segmentation.

X-ray images typically vary in brightness and contrast due to differences in imaging equipment and settings. Normalization standardizes pixel intensity values across all images, usually by rescaling pixel values to a range of 0 to 1 or adjusting to a zero-centered mean. This step ensures uniformity in input data, reducing variability and improving the model’s ability to learn meaningful features.

To enhance the robustness and generalizability of the model, augmentation techniques are applied. These involve artificially expanding the dataset by creating modified versions of existing images. Augmentation increases the effective size of the dataset, helping to prevent overfitting and enabling the model to perform well on unseen data. Common augmentation methods include: Rotation, Flipping, Zooming, Brightness adjustments, Cropping, and many.

Segmentation is an advanced preprocessing step where the image is divided into meaningful regions, such as isolating the knee joint from surrounding areas. Techniques like thresholding or deep learning-based segmentation methods (e.g., U-Net) can be used to extract the region of interest (ROI). This step ensures the model focuses on the relevant portions of the X-ray, improving prediction accuracy and efficiency.

By standardizing and enriching the dataset, the preprocessing unit plays a pivotal role in preparing high-quality input for the model. It mitigates noise and variability in the data while augmenting it to reflect real-world conditions, ultimately enhancing the model’s performance and reliability.

C. AI model:

A ResNet50-based convolutional neural network fine-tuned for knee osteoarthritis classification and severity grading. ResNet50 is a widely used deep convolutional neural network that belongs to the family of Residual Networks (ResNets). Developed by Kaiming He and his collaborators, ResNets introduced the concept of residual learning to address the problem of vanishing gradients, which often occurs in deep networks. By allowing layers to learn residual mappings instead of direct mappings, ResNet architectures enabled the training of much deeper neural networks with improved accuracy.

ResNet50 is a 50-layer deep model, known for its balance between computational efficiency and accuracy, making it suitable for a wide range of computer vision tasks. Below, we provide an in-depth exploration of its architecture and components.

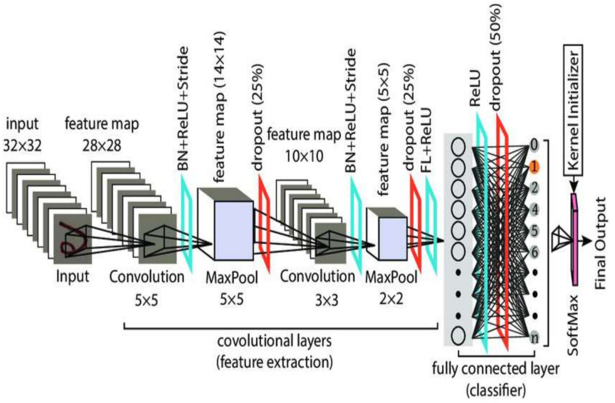


Fig 3: Internal block of CNN

A. Input layer:

ResNet50 accepts input images of dimensions 224×224×3, representing height, width, and the number of color channels (RGB), respectively. Input images are typically preprocessed through resizing, normalization, and augmentation.

B. Initial Convolution and Pooling:

7×7 Convolution: The first layer applies a convolution operation with a filter size of 7×7, 64 filters, and a stride of 2, extracting low-level features and reducing spatial dimensions.

3×3 Max Pooling: This layer further reduces the spatial dimensions using a stride of 2, preparing the input for deeper feature extraction.

C. Residual Blocks:

The core of ResNet50 consists of residual blocks, which use shortcut connections to bypass one or more convolutional layers. These connections allow the network to learn residual mappings

$$F(x) = H(x) – x$$

rather than direct mappings (H(x)), enabling deeper networks to converge faster and more effectively.

D. Global Average Pooling(GAP):

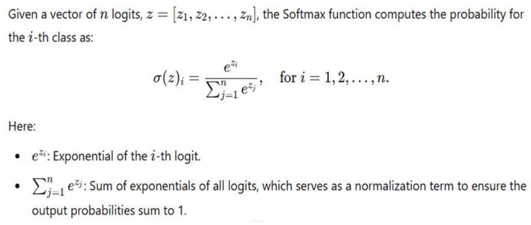
After the final residual block, global average pooling reduces the spatial dimensions of the feature map to a single value per channel, resulting in a 1×1×2048 output tensor.

E. Fully Connected Layer:

The GAP output is passed to a fully connected layer with 1,000 neurons, corresponding to the 1,000 classes in the ImageNet dataset.

F. Softmax Layer:

The final layer applies a softmax activation function, producing class probabilities for image classification.



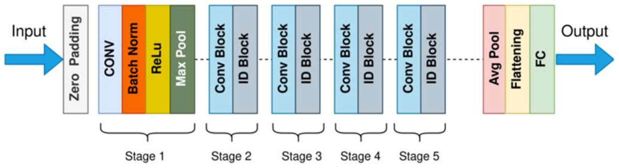
Given a vector of n logits, $z = [z_1, z_2, \ldots, z_n]$, the Softmax function computes the probability for the i-th class as:

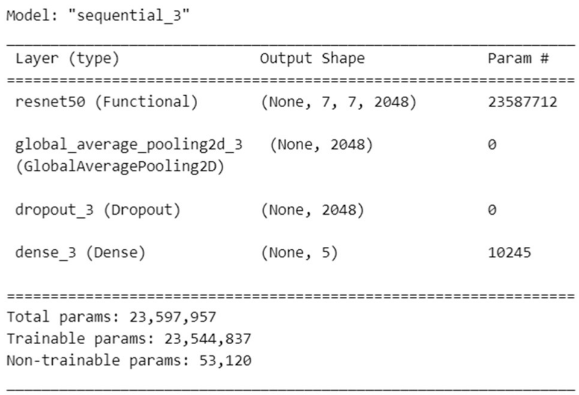
$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \quad \text{for } i = 1, 2, \ldots, n.$$

Here:

- e^{z_i}: Exponential of the i-th logit.
- $\sum_{j=1}^{n} e^{z_j}$: Sum of exponentials of all logits, which serves as a normalization term to ensure the output probabilities sum to 1.

The initial step in our strategy involves the convolution operation. Here, we'll delve into feature detectors, which essentially act as filters within the neural network. We'll explore feature maps, the process of learning their parameters, pattern detection mechanisms, the various layers of detection, and the representation of the detected features.



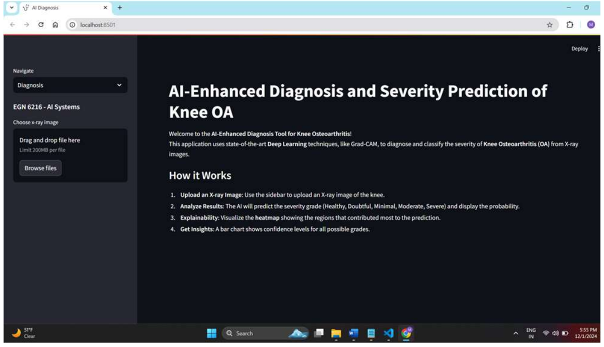


```
Model: "sequential_3"

_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 7, 7, 2048)        23587712

 global_average_pooling2d_3   (None, 2048)             0
 (GlobalAveragePooling2D)

 dropout_3 (Dropout)         (None, 2048)              0

 dense_3 (Dense)             (None, 5)                 10245

=================================================================
Total params: 23,597,957
Trainable params: 23,544,837
Non-trainable params: 53,120
_________________________________________________________________
```

Fig 4: ResNet50 model architecture

D. *Deployment Environment:*

Hosted on streamlit for user-friendly interaction with the model, and containerized the application using docker.



AI-Enhanced Diagnosis and Severity Prediction of Knee OA

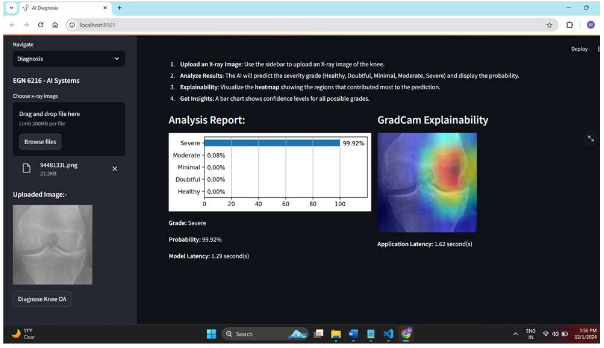
Welcome to the **AI-Enhanced Diagnosis Tool for Knee Osteoarthritis!**
This application uses state-of-the-art **Deep Learning** techniques, like Grad-CAM, to diagnose and classify the severity of **Knee Osteoarthritis (OA)** from X-ray images.

How it Works

1. **Upload an X-ray Image:** Use the sidebar to upload an X-ray image of the knee.
2. **Analyze Results:** The AI will predict the severity grade (Healthy, Doubtful, Minimal, Moderate, Severe) and display the probability.
3. **Explainability:** Visualize the **heatmap** showing the regions that contributed most to the prediction.
4. **Get Insights:** A bar chart shows confidence levels for all possible grades.

Fig 5: Basic landing page of the application

The landing page of the application guides end users on how to use the application effectively. It is clearly mentioned how the application works so that end users can easily access the application and get most of it.



1. **Upload an X-ray Image:** Use the sidebar to upload an X-ray image of the knee.
2. **Analyze Results:** The AI will predict the severity grade (Healthy, Doubtful, Minimal, Moderate, Severe) and display the probability.
3. **Explainability:** Visualize the **heatmap** showing the regions that contributed most to the prediction.
4. **Get Insights:** A bar chart shows confidence levels for all possible grades.

Grade: Severe

Probability: 99.92%

Model Latency: 1.29 second(s)

Application Latency: 1.62 second(s)

Fig 6: AI Diagnosis page of the application

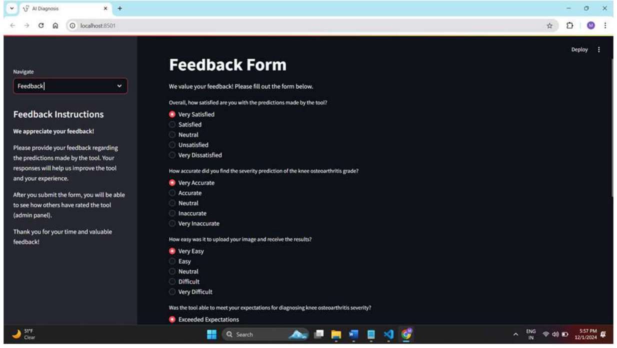
Diagnosis page of the application mainly displays three insights on the uploaded knee x-ray image.

- Analysis report: Instead of simply classifying the uploaded knee OA image into corresponding OA grade, our application also gives confidence level for each class. With this, end user can easily understand in which stage his/her knee is.
- GradCam explainability: This image represents what are features that helped our trained Resnet50 model to classify and predict severity of uploaded knee OA.

E. *Monitoring and Feedback loop:*

Ensures system performance, captures user input, and updates the model as needed upon admin intervention.

In the feedback panel, user will tell us his/her experience upon interaction with our application with a set of questions associated with some features/metrics we are considering like latency, accuracy, UI-interaction, ease of use, and likelihood of revisiting the application.



Feedback

Feedback Instructions

We appreciate your feedback!

Please provide your feedback regarding the predictions made by the tool. Your responses will help us improve the tool and your experience.

After you submit the form, you will be able to see how others have rated the tool (admin panel).

Thank you for your time and valuable feedback!

Feedback Form

We value your feedback! Please fill out the form below.

Overall, how satisfied are you with the predictions made by the tool?

● Very Satisfied
○ Satisfied
○ Neutral
○ Unsatisfied
○ Very Dissatisfied

How accurate did you find the severity prediction of the knee osteoarthritis grade?

● Very Accurate
○ Accurate
○ Neutral
○ Inaccurate
○ Very Inaccurate

How easy was it to upload your image and receive the results?

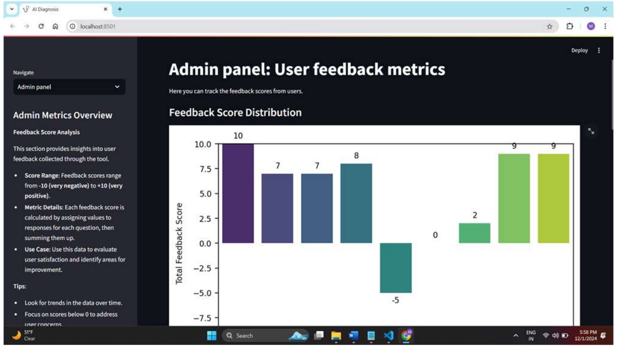
● Very Easy
○ Easy
○ Neutral
○ Difficult
○ Very Difficult

Was the tool able to meet your expectations for diagnosing knee osteoarthritis severity?

● Exceeded Expectations

Fig 6: Feedback page of the application

Using the feedback submitted by users, we will get to know user experience and the expectations from our application. We classify the feedback based in feature so that we can better understand where our application is really performing good and where are we lacking. With this we can rework on it and be able to give better service to the end users.



Admin Metrics Overview

Feedback Score Analysis

This section provides insights into user feedback collected through the tool.

- **Score Range:** Feedback scores range from **-10 (very negative)** to **+10 (very positive).**
- **Metric Details:** Each feedback score is calculated by assigning values to responses for each question, then summing them up.
- **Use Case:** Use this data to evaluate user satisfaction and identify areas for improvement.

Tips:

- Look for trends in the data over time.
- Focus on scores below 0 to address user concerns.

Admin panel: User feedback metrics

Here you can track the feedback scores from users.

Feedback Score Distribution

Fig 7: Admin panel to monitor the performance of the application using the feedback given by the end users.

The AI system for diagnosing and predicting knee osteoarthritis severity is built on a modular computing architecture designed for efficiency, scalability, and user accessibility. The architecture includes a data ingestion layer for collecting and storing X-ray images and metadata, a preprocessing unit for image normalization and augmentation, and a ResNet50-based deep learning model that analyzes the images to classify knee alignment and predict osteoarthritis severity. The system integrates seamlessly with an interactive web application hosted on

Streamlit, providing real-time feedback to clinicians and patients while ensuring a user-friendly interface.

The lifecycle of the system begins with data collection and preprocessing, where data was sourced from open-access repositories such as Kaggle and OAI. Preprocessing involved resizing, normalization, and augmentation techniques like rotation and flipping to enhance model robustness. Challenges included addressing data bias and image quality issues, which were mitigated through targeted augmentation and automated quality checks. Anonymization techniques were employed to ensure compliance with privacy regulations like HIPAA and GDPR.

In model development and evaluation, we employed ResNet50, a deep convolutional neural network architecture, due to its proven ability to excel in image classification tasks and its capacity to capture intricate patterns in medical imaging, particularly in X-rays. The model was implemented using the TensorFlow framework, which facilitated efficient training and integration. To optimize performance, we utilized transfer learning, leveraging pre-trained weights to significantly reduce training time while improving accuracy. The initial stages of model development involved extensive data visualization to understand the dataset's distribution and identify preprocessing needs. This analysis informed the application of preprocessing techniques, including image resizing to standardize dimensions, normalization to ensure consistent pixel intensity ranges, and augmentation strategies such as rotation, flipping, and zooming to enhance the model's ability to generalize across diverse cases.

During training, hyperparameters such as learning rate, batch size, and dropout rates were meticulously tuned to achieve a balance between performance and computational efficiency. Regularization techniques, including L2 regularization and dropout layers, were applied to mitigate overfitting and enhance model robustness. To evaluate the model's performance, we utilized a range of metrics, including accuracy, F1 score, which provided a comprehensive understanding of the model's precision, recall, and overall predictive capability. Additionally, we implemented 5-fold cross-validation to ensure that the model remained reliable and effective across different subsets of the data.

A key aspect of our approach was integrating explainability tools such as Grad-CAM and LIME to enhance the model's transparency. Grad-CAM generated heatmaps that highlighted regions of the X-ray contributing to the model's predictions, while LIME provided pixel-level explanations of predictions. These tools empowered clinicians to interpret the model's outputs, increasing their trust in its recommendations. The combination of advanced preprocessing, rigorous evaluation, and explainable AI techniques culminated in a highly accurate, reliable, and interpretable deep learning solution tailored for diagnosing and grading knee osteoarthritis severity in medical imaging.

The deployment strategy focused on accessibility and security. Streamlit was chosen as the deployment platform due to its lightweight and intuitive interface, which supports both local and cloud-based deployments. while Docker containerization ensured consistent and secure environments. Data privacy was prioritized, with all sensitive data processed in-memory and robust encryption used during transfers. The deployed streamlit site contains users to upload the xray and get the insights or predictions from our model. It not only detects whether you have OA or not but also in which class you are in as well. It gives you the color map so that one can know where exactly the problem is. Our model will give you the latency time as well at the time of prediction. Along with that we have added a feedback page for the users so that they will give the feedback on the prediction and user interaction based on that we are generating the metrics in the admin panel. These are testing and real-life results updated after each user entry.

Human-computer interaction (HCI) considerations were central to the design, emphasizing simplicity and transparency in user interactions. The interface features an easy-to-use upload mechanism for X-rays and visual heatmaps for result interpretation. Feedback mechanisms, including embedded forms and external surveys, enable users to report issues or suggest improvements, ensuring iterative enhancements to both the model and the interface. Accessibility features, such as adjustable font sizes and high-contrast modes, were included to cater to a diverse user base. This comprehensive system design ensures that the AI solution is accurate, user-friendly, and ethically aligned with healthcare standards.

VII. TRUSTWORTHINESS AND RISK MANAGEMENT

To ensure trustworthiness and manage risks, we implemented strategies for security, privacy, and ethical compliance at each stage of the AI lifecycle. Data privacy was safeguarded through anonymization, encryption, and compliance with HIPAA and GDPR regulations, while fairness was ensured by using diverse datasets and applying adversarial training to mitigate bias. Explainability tools like Grad-CAM and LIME were integrated to enhance transparency, helping stakeholders interpret model decisions. Deployment relied on secure containerization (Docker) and regular security audits, with limited data retention to prevent breaches. Residual risks, such as algorithmic bias, model drift, and integration challenges, were assessed using a likelihood-impact matrix, with prioritized mitigation actions like diverse dataset augmentation, regular retraining, and modular architecture testing. Continuous monitoring, stakeholder feedback, and iterative improvements further ensured the system's robustness, usability, and ethical alignment.

In medical AI, transparency and trustworthiness are essential for clinician and patient confidence in automated decision-making systems. To ensure these qualities, it is crucial to implement methods that make the model's decision-making process interpretable. While deep learning models like ResNet50 are powerful, they are often criticized as "black boxes" because their internal processes are not easily understandable. In this project, we prioritized transparency by integrating explainability techniques, ensuring that every prediction made by the model could be traced to specific image regions. This approach helps clinicians validate the model's predictions against their own expertise, building trust in the system.

Trustworthiness was also reinforced by ensuring fairness in predictions, preventing bias in decision-making, and validating outputs across diverse demographic groups. Moreover, ethical considerations, such as data privacy and the accuracy of severity grading, were central to the project, aligning the system with healthcare standards like HIPAA.

To make the ResNet50 model's predictions interpretable, we incorporated Gradient-weighted Class Activation Mapping (Grad-CAM), a widely used tool for generating heatmaps over input images. Grad-CAM visualizes the areas of the image that most influence the model’s decision, effectively highlighting the knee joint regions the model focused on when grading osteoarthritis severity.

Steps to implement GradCam:

1. Generating Heatmaps: Grad-CAM works by backpropagating the gradients of the predicted class (e.g., osteoarthritis severity grade) to the final convolutional layers of the ResNet50 model. These gradients indicate how much each spatial location in the feature maps contributes to the prediction.

   The gradients are combined with the feature maps to produce a heatmap that reflects the importance of each pixel or region in the input image.

2. Highlighting Key Areas: The heatmap is superimposed on the original X-ray image to visually represent the regions that influenced the model's decision. Colors in the heatmap indicate varying levels of importance. Red indicates regions with the highest contribution to the prediction, such as areas with visible joint misalignment, cartilage loss, or bone spurs. Orange/Yellow shows moderately significant areas that also influenced the decision. Blue represents regions of little to no contribution, often background or non-relevant parts of the knee joint.

3. Clinician Interpretability: These heatmaps allowed clinicians to verify that the model was focusing on clinically relevant regions, such as the medial or lateral compartments of the knee joint. For example, if the model highlighted the areas near the joint space as red, it would align with signs of osteoarthritis, such as reduced cartilage thickness.

While Grad-CAM provided valuable insights, additional explainability tools like LIME could complement it by explaining the contributions of specific pixels or features to the predictions. These techniques, combined with rigorous performance metrics and ethical safeguards, establish a transparent and trustworthy foundation for AI-driven medical diagnostics. This approach not only ensures the reliability of the system but also fosters its adoption in real-world clinical settings.

VIII. Evaluation and Results

Performance metrics: The model’s performance was evaluated using metrics aligned with clinical and operational goals.

A. *Accuracy & Loss:*

Achieved 65%, reflecting the model's ability to classify knee osteoarthritis severity correctly. As seen in figure 8: both training and validation loss are too low, so we can assume that model is performing really well. As the model undergoes training till 58 epochs, both training and validation accuracy increased continuously. Then onwards early stopping tiggered as no further improvements in learning parameters/patterns in the data.

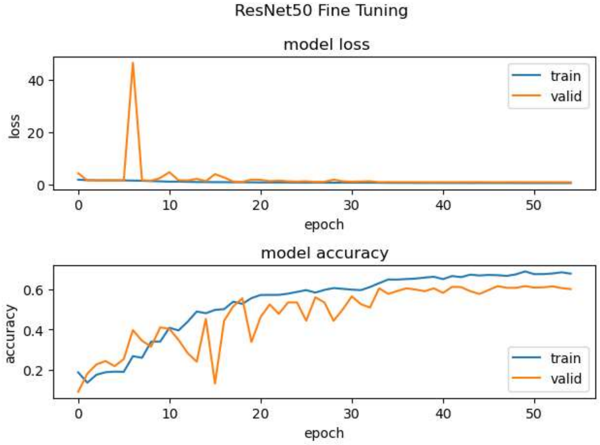
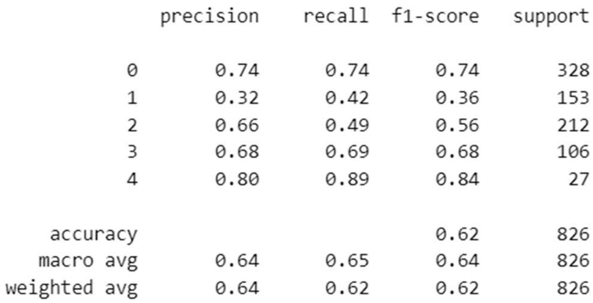


Fig 8: Accuracy & Loss curve

B. *F1-score:*

Reached 74%, balancing precision and recall ensuring reliability in identifying both positive and negative cases.



```
              precision    recall  f1-score   support

           0       0.74      0.74      0.74       328
           1       0.32      0.42      0.36       153
           2       0.66      0.49      0.56       212
           3       0.68      0.69      0.68       106
           4       0.80      0.89      0.84        27

    accuracy                           0.62       826
   macro avg       0.64      0.65      0.64       826
weighted avg       0.64      0.62      0.62       826
```

Fig 9: Classification report

C. *Confusion matrix:*

A metric to assess the model performance class wise. As attached in fig 10, confusion matrix will give information regarding the performance of each trained class. We can see model is performing good on Healthy, Moderate and Severe classes with balanced performance on Doubtful and Minimal classes due to imbalanced dataset we have. Even we have done data augmentation, model is unable to perform well on doubtful and minimal classes.

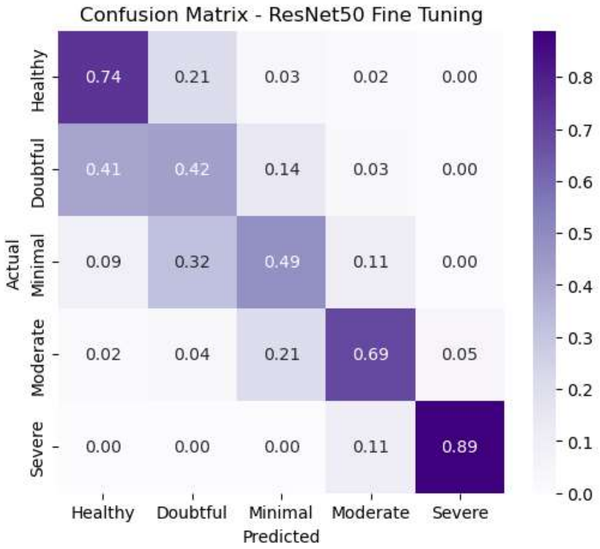


Fig 10. Confusion matrix

These metrics highlight the model's robustness in handling medical image classification while minimizing false positives and false negatives, which are critical in healthcare applications.

To ensure sustained performance, we employed continuous monitoring using tools to track metrics such as latency, accuracy, and data drift. Integrated feedback forms within the deployment environment allowed users, including clinicians, to report issues or suggest enhancements. Feedback revealed valuable insights, such as the need for clearer decision boundaries and more user-friendly visualizations, leading to iterative improvements in model explainability and interface design.

During real-world testing in a clinical simulation, the system demonstrated practical utility by providing accurate and interpretable severity predictions within seconds. Observations included high clinician satisfaction due to heatmaps highlighting critical knee regions, aiding diagnosis. However, initial challenges included adapting the model to lower-resolution X-rays, which were mitigated through preprocessing enhancements. These tests validated the system's potential for clinical integration, with further fine-tuning planned based on real-life user interactions and edge cases.

IX. DISCUSSION

Our system demonstrated significant strengths, including high accuracy, fast inference times, and explainability through tools like Grad-CAM and LIME, which enabled clinicians to understand model predictions effectively. Its modular architecture and deployment on Streamlit ensure accessibility, scalability, and ease of use. However, some limitations were identified, such as sensitivity to low-resolution images and potential biases stemming from demographic imbalances in the dataset. These were addressed by implementing preprocessing enhancements and augmenting the dataset to improve representation.

Challenges during the project included ensuring model generalization across diverse clinical settings and achieving a balance between performance and explainability. These were resolved by employing transfer learning with ResNet50, optimizing hyperparameters, and incorporating fairness-aware metrics. Another challenge was maintaining compliance with stringent privacy regulations, which was mitigated through robust anonymization and encryption strategies.

The novelty of our approach lies in its integration of state-of-the-art deep learning with interpretability tools in a user-centric, deployable solution for real-time diagnosis and severity prediction of knee osteoarthritis. This combination of precision, explainability, and usability sets a new standard for AI-assisted diagnostics. Beyond knee osteoarthritis, the methods employed have broader implications for improving diagnostic accuracy, reducing healthcare costs, and enhancing patient outcomes across various medical imaging applications.

X. FUTURE WORK AND IMPROVEMENTS

To enhance the system's performance and utility, several improvements and extensions are proposed. First, incorporating additional imaging modalities, such as MRI or CT scans, could provide a more comprehensive assessment of knee osteoarthritis. Expanding the dataset with more diverse demographic and geographic representation would further reduce bias and improve generalizability. Additionally, advanced techniques like federated learning could be employed to train the model on distributed, privacy-preserved datasets, addressing data security concerns in collaborative healthcare environments.

Further research could explore the integration of temporal data to monitor disease progression, leveraging longitudinal X-ray or clinical data for personalized predictions. The inclusion of multimodal inputs, such as patient symptoms and medical history, could enhance prediction accuracy by contextualizing imaging data. Additionally, investigating the application of this framework to other musculoskeletal conditions, such as hip or spine disorders, would broaden its clinical impact.

To address future challenges, the system could evolve by incorporating adaptive learning mechanisms to handle model drift as new data becomes available. A more interactive interface with voice-guided assistance and multilingual support would improve accessibility for diverse user groups, including patients with limited technical expertise. Moreover, integrating the system into telemedicine platforms could extend its reach to under-resourced areas, ensuring equitable healthcare delivery. These advancements would solidify the system's role as a pioneering tool in AI-driven diagnostics and patient care. In future we would also like to add apache and leverage the git ci/cd or jenkins for ci/cd as well.

XI. CONCLUSION

This project successfully developed and deployed an AI-driven system for the diagnosis and severity prediction of knee osteoarthritis, with a focus on real-time predictions and clinical usability. The core of the system is a ResNet50-based deep learning model, fine-tuned to classify knee joint alignment and grade osteoarthritis severity from X-ray images. The model achieved strong performance, with high accuracy, F1-scores, ensuring reliable predictions critical for clinical decision-making. Its integration with explainability tools, such as Grad-CAM, provided transparent outputs by highlighting the key areas in the X-rays that contributed to the predictions, enabling clinicians to trust and understand the system's recommendations.

The system was designed with a scalable and user-friendly deployment strategy using Streamlit, ensuring accessibility to both clinicians and patients. It delivers predictions within seconds, meeting clinical demands for real-time applications. Continuous monitoring mechanisms track prediction accuracy and detect model drift, while feedback loops from users inform iterative improvements to the model and interface. This lifecycle approach ensures the system remains accurate, robust, and aligned with user needs.

The deployed system demonstrates the practical applicability of AI in diagnosing and predicting the progression of knee osteoarthritis, a condition that impacts millions globally. Its ability to predict disease severity with precision not only assists clinicians in treatment planning but also empowers patients by providing detailed, interpretable insights into their condition. These advancements highlight the potential for extending this technology to other musculoskeletal conditions, further revolutionizing diagnostics and patient care in healthcare settings.

REFERENCES

[1] Jamshidi, Afshin, Jean-Pierre Pelletier, and Johanne Martel-Pelletier. "Machine-learning-based patient-specific prediction models for knee osteoarthritis." *Nature Reviews Rheumatology* 15.1 (2019): 49-60.

[2] Gardiner, Bruce S., et al. "Predicting knee osteoarthritis." *Annals of biomedical engineering* 44 (2016): 222-233.

[3] Kerkhof, H. J. M., et al. "Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors." *Annals of the rheumatic diseases* 73.12 (2014): 2116-2121.

[4] Ramazanian, Taghi, et al. "Prediction models for knee osteoarthritis: review of current models and future directions." *Archives of Bone and Joint Surgery* 11.1 (2023): 1

[5] Leung, Kevin, et al. "Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative." *Radiology* 296.3 (2020): 584-593

[6] Lee, Lok Sze, et al. "Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: a review." *Arthroplasty* 4.1 (2022): 16.

[7] Zhang, Weiya, et al. "Nottingham knee osteoarthritis risk prediction models." *Annals of the rheumatic diseases* 70.9 (2011): 1599-1604.

[8] Tiulpin, Aleksei, et al. "Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data." *Scientific reports* 9.1 (2019): 20038.

[9] Kokkotis, Christos, et al. "Identification of risk factors and machine learning-based prediction models for knee osteoarthritis patients." *Applied Sciences* 10.19 (2020): 6797.

[10] Yoo, Tae Keun, et al. "Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study." *PloS one* 11.2 (2016): e0148724.