Osteoarthritis and Gout Detection using Machine Learning

Murlidhar[1], Laxmi[2], Arun Kumar[3], Hemant[4], Dr. Karuna Sharma[5], Dr. VISHAL SHRIVASTAVA[6], Dr. AKHIL PANDEY[7],
[1]B.TECH. Scholar, [2]B.TECH. Scholar, [3]B.TECH. Scholar, [4]B.TECH. Scholar, [5]Assistant Professor,
[6]Assistant Professor, [7]Professor,
Computer Science & Engineering
Arya College of Engineering & I.T. India, Jaipur

1. Abstract:

Joint disorders such as osteoarthritis and gout are afflictions of people that ultimately reduce quality of life. Identifying the disease early is essential to avoid complications during later stages of the disease. This paper proposed an automated system for detecting osteoarthritis and gout utilizing deep learning techniques from X-ray medical imaging. The study constructed a pre-trained MobileNetV2 model that classified osteoarthritis and gout (determining probability for diagnosis using symptom data). Additionally, the automated detection included a web-based user interface for users to upload X-ray images and describe symptoms to determine risk for both maladies. The systems performance will be measured using accuracy, precision, recall, and the F1 score. The study reported and achieved a validation accuracy of 82.32% for MobileNetV2 study, indicating the viability of usefulness for deep learning medical imaging in medical diagnosis. Next steps for enhancement were provided in the study, which may include an increased sophistication of augmentation methods and/or tuning hyperparameters and better explain the varying degrees of augmentation methods. Keywords: Osteoarthritis, Gout, Deep surveillance learning, Standardized medical imaging, MobileNetV2, Disease analysis, Disease classification.
Keywords: Machine Learning, AI, Supervised Learning, Reinforcement Learning

2. Introduction

Osteoarthritis (OA) and gout are chronic diseases affecting millions of people across the globe, and while they are two fundamentally different conditions (OA is a degenerative joint disease due to cartilage degeneration; gout is an inflammatory arthritis due to uric acid deposition), the current approach to diagnosing OA and gout remains reliant on clinical evaluation and radiological investigation, both of which are slow and rely on expertise interpretation. Advances in deep learning have allowed for the diagnostic evaluation processes to be automated utilizing image-processing methods for disease diagnosis. In this study, we develop a convolutional neural network (CNN) model with a MobileNetV2 architecture to detect OA and use a symptom-based probabilistic model for the diagnosis of gout. This paper will present an open web-based diagnostic tool aimed at improving the accessibility and efficiency of medical diagnostics.

3. Literature Review

1. Deep Learning in Medical Imaging

The revolution that has taken place in medical imaging has been largely due to deep learning with the capability to diagnose medical images quickly, reliably, and consistently. Convolutional neural networks have been used mostly in feature extraction, segmentation, and classification scenarios. Research shows transfer learning could increase performance both cross-domain & on medical datasets with less training data. Some researchers have been able to assist in the detection of disease from medical images with backbone models like ResNet, VGG, and MobileNet for predicting and assisting in clinical practice.

2. **Detecting osteoarthritis (OA) via deep learning** is quite prevalent. There are numerous articles confirming the ability of CNN architectures to detect OA severity through X-ray image results. The Kellgren-Lawrence grading system has primarily been used as the gold standard for modeling. The articles have indicated that more lightweight architectures that pay attention to speed (such as MobileNetV2) allow for near real-time classification and effort/resource conservation. Accuracy is cited around 70%-80%, depending on data quality and inputs into the pre-processing and data stages.

3. The Application of Machine Learning to Diagnose Gout:

Diagnosing gout relies solely on symptoms and lab evaluations, and is different than OA. Some recent publications have published research studies using Machine Learning in attempts to classify gout, that incorporated symptoms, uric acid levels, and demographic information to account for affinity of diagnosis of gout. Hybrid models are also emerging that combine symptom-based Machine Learning and imaging-based Machine Learning, which presents a strong opportunity moving forward. These models used feature weighting and probabilistic inference models to improve accuracy in diagnosing gout.

4**. Challenges in Automated Medical Diagnosis**

While AI-based medical diagnostic systems have advanced considerably, many obstacles remain.

- Insufficient data: Not all medical data is publicly available or open-source, which challenges model training.

- Imbalanced data: If the targeted data is not increased, imbalanced data sets can be problematic for the model and different resampling methods could be utilized.

- Bias interpretability: The covert nature of a AI-based diagnostic model requires questions related to bias in clinical care.

- Generalizability: Variances in medical imaging may compromise the reliability of a trained model when exposed to a different dataset.

AI & Machine Learning in Rheumatic Disease Diagnosis

The advancements in artificial intelligence (AI) and machine learning (ML) have made great progress in gynaecologic disorders as they apply to automated image analysis, enhanced prediction capabilities, and early disease diagnosis. Deep learning models employing convolutional neural networks (CNNs) have shown high accuracy in identifying osteoarthritis (OA) and rheumatoid arthritis (RA) on medical imaging. For example, Ali (2025) created a CNN model that included an edge-detection method with 98.2% accuracy in the diagnosis of knee OA. Kadu & Pawar (2025) created Bilinear CNN (BiCNN) for the grading of severity in knee OA with 94.28% accuracy. Most recently, ML models have been integrated into the analysis of MRI along with other x-ray-based images. Hu et al. (2024) utilized a temporal-regional graph-convolutional network (TRGCN) that predicted the progression of knee OA with area under the curve (AUC) scores greater than 0.84. Chen et al. (2025) demonstrated that deep learning algorithms could be used to identify and classify structures of the joint on MRI images to improve accuracy of procedures and decrease human dependency on interpretation.

The application of Artificial Intelligence (AI) in rheumatology also includes the innovation of Natural Language Processing (NLP). Natural language processing methods can scan for symptoms in electronic health record (EHR) data, and identify trends and patterns in presenting symptoms in patients that have already been diagnosed with rheumatic disease, to help to improve the overall clinical decision making and delivery of care for patients. For example, Osborne et al. (2024) used an NLP based approach on EHRs to detect gout flares, finding an F-score of 87% in regards to its effectiveness within documentation of an emergency department. NLP methods or applications do not replace or remove the clinical reasoning that is part of clinical practice, but they support or supplement clinical decision making by using real time healthcare data, while also helping to lessen the a priori

diagnostic burden placed on providers. Despite even newer innovations in AI diagnostic models, best practice in applied AI involves monitoring data bias, guaranteeing interpretability of the models, and model external validation based on the standard of practice for devising diagnostic models or methods; all of these factors are important to best practice-based implementation of applied AI in clinical settings. A clinical practice should evaluate and address both issues, as studies suggest that having more diverse datasets based on larger cohort sizes, and use of direct clinical validation to address generalizability for clinical application is a significant value for applied AI models.

Multi-Modal Data Integration in Healthcare

One innovative idea is to combine interpretable information from clinical, imaging, biomarkers, and genomic data sets together into a collaborative framework to strengthen evidence for identifying a disease marker for diagnosis and treatment. AI-based model frameworks that integrate multiple data modalities into single models have demonstrated significant benefits with regard to diagnostic performance and subsequent clinical outcomes. Kokkotis et al. (2020) showed that ML models that incorporated clinical, imaging, kinematic and biochemical data yielded strong diagnostic performance, with AUCs of 0.95 (for SLE) and 0.89 (for NPSLE). Multi-panel models were also found to have similar or in some cases superior diagnostic performance to multi-modal models. Even if multi-modality data may improve performance compared to unimodal models, a multitude of methods to ensure multi-modality information will achieve the expected potential level of improvement, can be used.

Genomic and/or biomarker AI approaches are also instrumental in advancing rheumatology research. For example, Chen et al. (2025) have created ML models that used single cell RNA sequencing, and found WDR74 and TNFRSF12A to be early OA biomarkers, with an AUC of >0.9. This study emphasizes the potential of genomic and imaging data to add to the classification of disease and enhance disease classification more broadly. Despite that promise, the integration of multi-modal data is still a challenge. The variability of data, standards, and even complexity of computation all pose problems for market uptake. Therefore, it is vital that we have interoperable health information systems and/or standards.

Proposed Methodology

Dataset Collection and Preprocessing

- The dataset is made up of X-ray images that have been labelled according to their severity level and the processing is accomplished through the following steps:
- Automated extraction. The dataset is compressed and will extract automatically at run time. This way, the loading of images will be easier. It will also eliminate the risk of human error associated with extraction. The automated extractor will also sort the extracted files into necessary folders for viewing and processing.
- Image resizing. Medical images are taken at varying resolutions which cause problems during the model training phase. In an effort to keep input images in a consistent dimension, the images will be resized to have a size of 224 x 224 pixels. This size is compatible with MobileNetV2 since it has fixed-size input dimensions and can extract features from the images without distortion.

• Data Augmentation: Because medical datasets are often small, we will augment data in an effort to artificially raise the size of the dataset, and help generalization for the model being trained. The data augmentation methods will include:

◦ Random Rotation– Each image will be rotated a small amount, or degrees, to randomize orientation.

◦ Flipping – To account for the inherent variation of X-ray images, images were horizontally flipped.

◦ Brightness – Brightness was adjusted when necessary to improve contrast of some subtler features of the images.

◦ Contrast – Contrast was adjusted if needed to improve visibility of structures of importance. All of these adjustments, help with preventing the model from becoming over reliant on particular structures that may bring

down the overall accuracy and generalization. • Overfitting can be reduced as a function of training a model to begin to learn features that will generalize, not just memorizing instances.
Normalization: Pixel values for raw image data is usually between 0-255. Normalization rescales to a [0,1] range, normalizes, and helps to stabilize training and speed up convergence
training. because there is an instability that can make it slower. Normalizing pixel values provides numerical stability, allows for the most optimization, and will lead to less chance of instability and faster convergence.

Model Architecture

- For the classification of osteoarthritis, the transport will be MobileNetV2 in this case for feature extraction as it is able to decrease the computational costs:

- **Input Layer**: The input layer is X-ray images resized to fit the input size of MobileNetV2 and will be in dimension size of either grey scale or RGB colour of (224, 224, 3) for a feature extraction.

- **Feature Extraction**: MobileNetV2 allows for extractions of features as it is a series of depth wise separable convolutions in a closed computational bill of that of other traditional convolution methods to capture and learn clinical important spatial hierarchy of osteoarthritis clinical modelling context. The essential situation here is MobileNetV2 depth wise convolutional layers are taking the important patterns that the training set showed or even some that were rarely present of osteoarthritis severity level and doing this with lessened computational costs.

- **Global Average Pooling Layer:** It is routine derived from existing literature to see fully connected layers added with no forethought, and the addition of burden too early with parameters which adds back convolved feature maps dimensionality to reduce considerable amounts down to a modest averaged unit. This helps with generalization and reduce the risk of overfitting, after passing the learned feature extraction from convolutional layers.

- **Fully Connected Layers:** We proceed from layer 4 where features were extracted to the fully connected layer with ReLU rectified linear unit, output response. The edge of the ReLU would allow the model nonlinearities that extend beyond a simpler model to learn complex booster valued osteoarthritis severity levels.

- **Dropout Regularization:** Dropout regularization has been implemented in the fully connected layer to reduce the level of overfitting during training. Dropout means randomly choosing some percentage of the neurons to cease from being activated

- **Output Layer:** The output layer of the model utilizes a Softmax activation function to classify the input into a specific osteoarthritis severity level (0-4). The Softmax activation function represented a probability function, which allowed for the more natural decision boundary response to classes of severity levels.

Training Process

The training plan aims to optimize the performance while reducing overfitting:

• **Loss Function:** We will use Categorical Cross entropy for multi-class classification, which allows the model to penalize incorrect guesses properly within classes.

• **Optimizer:** The use of Adam as the optimizer with a learning rate of 0.0001 is intentional in terms of speed vs. accuracy. Adam allows for adaptive learning rates that make the training process relatively stable small sized datasets.

• **Batch Size:** The use of a batch size of 32 images will optimize GPU memory usage on calculations made in each iteration, but also provide some stability to our training process.

• **Epochs, and Early Stopping:** The model will be trained for 50 DoN, but may stop early based on validation performance. If the performance of the model is not improving on the validation dataset we will stop training after a certain predetermined amount DoN. This is to reduce the potential for overfitting the training set.

• **Evaluation Metrics:** We will evaluate the performance by using accuracy, precision, recall and F1-score: these metrics will allow us to have a more comprehensive overview of the reliability of classification for each level for severity.

Gout Detection Mechanism

In contrast to osteoarthritis, where deep learning on image analysis is the basis of the detection process for the disease, the detection of gout must use a symptom-based disease model of probability. The probability of gout is determined by combining clinical knowledge with calculation into the probability derived from the symptoms provided by the patient. The scenarios for the gout detection process will utilize sensitivity weighted symptoms, probabilities, and dynamic visualization, which will be explained further.

1. Weighted Symptom Assessment

• Gout is a complex disorder that has variability in different clinical features, but these clinical features do not have equal diagnostic value. Joint pain is an important clinical feature of gout; however, the intensity of pain and the duration of pain are both important factors that can influence the diagnosis of gout. A weighted clinical feature evaluation accounts for the variability in clinical features.

• **Concept:** Each clinical feature the user reports will include joint pain, swelling, or redness, and will receive a weight indicating the clinical importance of this feature in diagnosing gout. Clinical feature weights will be based on evidence and expert clinical opinion that describe the strength of association with a specific clinical feature to gout.

• **Why it matters:** For instance, clinical symptoms of gout such as acute pain, redness, and swelling in the joint are much more likely to communicate a diagnosis of gout than mild discomfort. In this way, weighing more clinically significant features with higher weights enables the system to yield better prognostications. For instance, in the case of severe pain in the joint, the distinction in clinical weight for gout diagnosis between severe pain compared to mild discomfort is more important for the purposes of distinguishing diagnoses for the purpose of weight as opposed to the differential between mild discomfort; therefore, that is why the severe pain has a higher weight in the weight system.

• The system would take this input and operationalize it to assign a numeric weight to the individual symptoms based on their clinical weight. For example: The numeric weight for redness of a joint would be .5 because this symptom is highly correlated with gout flare. The numeric weight for severe pain would be .8 because severe pain is one of the most prominent features of acute gout. The numeric weight for swelling of a joint would be .6 because swelling is commonly found in the clinical context of gout flare. The aim of this step is to ensure the system is aligned with clinical reasoning in that it weights symptoms that have the highest probability of synonymous with a gout.

2. Probability Calculation

- After capturing the user's symptoms, we calculate the likelihood of gout based on weighted symptoms. The calculation combines the symptoms reported by users, along with how each one is weighted; to produce a quantitative distribution for the likelihood the user has gout.

- **Concept:** We create a measure of probability based on the weight of symptoms reported by the user. The weighted symptoms will then be statistically combined, whether that be at the statistically-likely average or logistic regression approach, to produce a final statistical probability score. The probability score represents a chosen set of symptoms which the user reported to evidence a potential likelihood of gout.

• **Importance:** Calculating a probability score provides a clinically-relevant degree of uncertainty, instead of just a yes or no answer. For example, gout can present very differently, and the system is able to more accurately represent this clinical nuance by providing a probability to the user for gout, instead of simply saying yes or no.

A user with mild pain and swelling may have a low probability for gout (i.e., 35%), whereas a user with more significant and severe pain swelling and redness, may have a higher probability for gout (i.e., 85%).

- **Implementation:** A simple mathematical function can be used to calculate the probability score of gout. For example, if a user reports three out of five possible symptoms, and where symptoms are weighted, we could calculate a weighted average as the score Gout of disease:

Pgout=∑(weight of symptom×presence of symptom)∑weights of all symptomsP_{\text{gout}} = \frac{\sum (\text{weight of symptom} \times \text{presence of symptom})}{\sum \text{weights of all symptoms}}Pgout =∑weights of all symptoms∑(weight of symptom×presence of symptom)

This allows the system to calculate a value between 0 and 1, where higher values indicate a stronger likelihood of gout.

3. Dynamic Visualization

- The final aspect of the chronic gout detection system is the ability to dynamically visualize the gout probability score. Instead of just indicating a numeric score, you will now have a visual depiction of the probability of gout, which will help you to more easily interpret the result.

- The unique aspect of live visualization is that the detection system can output (for example via a graph, or a chart or even a progress bar) either the gout probability score. This is a more intuitive output to the user, especially for the user who does not have a medical background. This output provides a simple visual representation of the intensity of a condition, as well as the severity of their symptoms.

- The utility of live visualization is in facilitating interpretation, as dispersion provides an easy to see understanding of the effect of clinical features on probability of gout. For example, if you have an 80% probability of gout, you might see a bar chart, with an almost filled bar, representing high probability, with segments with different colours representing the clinical features (e.g. red for pain, blue for swelling).

• **Implementation:** Typical visual displays include Pie charts that offer a percentage representing each symptom's contribution to the probability score. o Bar graphs comparing the calculated probability of gout with the other possible diagnoses Progress bars that fill as the calculated probability of gout increases and give users quick, visually intuitive sense of how likely they are to have gout

5. Results & Discussion

Model Performance

- The assessment of the model pertaining to osteoarthritis detection will utilize multiple standard metrics which are typically used when evaluating standard metrics related to classification settings: accuracy, precision, recall, and F1-score will provide a more comprehensive view of whether the model is able to accurately classify X-ray images into the correct levels of severity of osteoarthritis and how well the model performs when generalizing to data that was not included in the training phase. In this section we will explicate on each of these metrics with slightly more detail, which includes how they are calculated, and how they specifically apply to disease detection, and the detection of osteoarthritis.

- Training accuracy = 82.39%Training accuracy is the ratio of the correct predictions made by the model to the training set. This will give some understanding as to how well the model is able to learn information from the training data, and what features are separating severity levels of osteoarthritis.

The significance: High accuracy on the training data tells us that the model is potentially learning the underlying relationships of some aspect of the data. That said, high accuracy is certainly a measure of performance you want to see high accuracy. High accuracy could also mean that the model has overly fit the training data itself, which is bad in that it is too specialized for the training data and does not perform well on out-of-sample data. The training accuracy builds a good initial measure of performance in this kind of application, but it may be best considered with a few more metrics. The training accuracy is probably going to be relatively high since the training set is simply the images the model learned with. However, the validation accuracy is definitely more important and indicates how the model performs generalizing beyond the training data.

Validation Accuracy: 79.32%

1. Validation accuracy refers to the proportion of correct predictions that the model made on the validation dataset. The validation dataset consisted of examples that the model had not previously encountered in its training processes. Validation accuracy is useful when we try to assess the quality of a model's performance on new examples that were not included in model training.

2. **Matter:** The model's validation accuracy of 79.32% demonstrates its ability to identify osteoarthritis, though just below any acceptable threshold for generalizable fidelity. Accordingly, the validation suggests accuracy for identifying some osteoarthritis but will be acceptable to continue appropriate refinement of questions and specifications into data that genuinely represents its new and expanded view. The 79.32% threshold is slight, especially in the realm of medical imaging, where presumptive use of a fining threshold of 85%makes it somewhat better, if nothing else, in consideration of ordering yet another threshold for osteoarthritis diagnosis, due to specific osteoarthritic radiographic presentations and conditions that decidedly are truly unique and separate, while historically conveying a slightly better prescription for actual disease differentiation whether it be in this case, or other medical diagnostics.
3. **Relevance:** The accuracies of these validations can methodically provide comparison with other performative models for similar medical diagnostic classifications or other technologies. One final consideration and outlet for utilization would be to discover the extent one could reveal power behind the model to then appropriately estimate degree of severity of osteoarthritis for those patients with osteoarthritis and asking for assistance for those patients where identification occurs through less traditional formats, especially for patients who present with very uncomfortable or arguably more work on the part of the end-user identifying, presenting conditions. Precision, Recall, and F1-Score
4. Besides accuracy, precision, recall, and the F1 score are relevant for measuring the models' capability to predict output accurately, particularly in the case of imbalanced datasets, as frequently observed in medical diagnosis, and these measurements provide more detail about the models' ability to discriminate as well as its sensitivity to rare, but clinically important classifications, such as osteoarthritis. Precision The metric of precision considers the proportion of real predicted positives (i.e. the levels of severity of osteoarthritis) that are truly positive. Precision is calculated by the following formula:

Precision=True PositivesTrue Positives + False Positives\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}Precision=True Positives + False PositivesTrue Positives

5. **Significance:** Precision matters because it helps to determine how reliable the model's positive prediction of the class is. In the osteoarthritis severity problem, if the model has precision of 80%, that means that whenever the model predicts a certain severity level, it will likely be correct 80% of the time. Relevance: In the context of disease detection, high precision implies less false positive tests. The precision metric is important for individual patients who should not receive treatment, or be sent for unnecessary tests that would be unwarranted from the model's prediction.

1. **Recall:** Recall measures the proportion of true positives (e.g., patients with a certain level of osteoarthritis) that the model correctly predicts. It is calculated as: -

Recall=True PositivesTrue Positives + False Negatives\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}Recall=True Positives + False NegativesTrue Positives

2. **Importance:** Recall is important in the context of a medical diagnosis of a disease, because we want to model to detect as many of the true cases of disease as possible. In cases of low recall, many cases of

disease are missed, which can pose a problem for patients that have not been diagnosed. - Importance: For the identification of osteoarthritis, we want to be high recall, so that the model detects most of the true cases of disease and decreases the probability of underdiagnosis. F1 Score: The F1 score defined as the harmonic mean of precision and recall, because it indicates an average of balance between two performance factors. The equation for precision and recall is:

F1-Score=2×Precision×RecallPrecision + Recall\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}F1-Score=2×Precision + RecallPrecision×Recall

3. **Importance:** The F1-score becomes especially important when there is an imbalance of importance between recall/precision. If a model has a high F1-score, this suggests the model is performing well in producing true positives, while also producing few false positives (good recall and good precision). Relevance: For medical image classification of diseases such as osteoarthritis, one must strive to produce true positives as accurately as possible (high recall). However, one must also strive to avoid categorizing true healthy patients that are not positive as sick (high precision). The F1-score provides a mechanism for accomplishing both of these goals. **Importance:** The F1-score is valuable if your model is experiencing a trade-off between precision and recall. A high F1-score is an indicator that you have a good trade-off between identifying true positives (high recall), while at the same time minimize false-positives (high precision). Significance: In the case of medical image classification, specifically for diseases such osteoarthritis, it will not only be important to get true positives (recall) as much as possible, it will also be important to never classify healthy patients as sick (precision). Using the F1-score can help achieve the process.

Challenges and Limitations

While the AI diagnostic system to identify osteoarthritis and gout has desirable abilities, there are still major problems and limitations that must be addressed before we can show differences:

1. Limitations of the dataset: The model is trained on a small dataset. Thus, it probably does not provide enough variability of such cases in the real-world. Generalization issues may occur when the patient population is thus substantially different than the case examples. Because of the generalization issues, the reliability of the model will not be optimal in clinical scenarios.

2. Class imbalance: The dataset is class imbalanced, since there is underrepresentation for some of the levels of severity. Class imbalance may bias the model to predicting the majority class; this, in turn, may limit the model's ability to classify properly the minority severity cases and the most severe cases of osteoarthritis.

3. Sensitivity and Specificity of the Model: The model must strike a balance between sensitivity (the ability of the model to correctly classify the positive class) and specificity (the ability of the model to correctly classify those patients whom do not have osteoarthritis). These underrepresented cases may again be problematic for the model, particularly in subtle borderline cases, to cause an overall cumulative effect on producing false positives and/or false negatives to the model and thereby, entail the cost of reduced diagnostic accuracy of the classifier as well.

4. Model Interpretability: Deep learning models are generally known as "black box" models, which is often cited as a limitation in reporting their use. From the standpoint of using a model in the medical field, it is important that physicians can understand and interpret the predictions generated from the model. If the model does not provide explanations readily, then there is a high likelihood of the model not being implemented in practice.

Future Directions

In order to improve the system, the following are recommended:

• **Advanced Augmentation Integrating:** Generative Adversarial Networks (GAN) into the synthetic X-ray generation tasks can mitigate variability in your dataset and class imbalance. This would lead the model to improved generalization and introduce some robustness to the model.

• **Hyperparameter Tuning Utilizing:** Bayesian Optimization, Grid Search, and even Adaptive Learning Rate Scheduling would lead to much faster model convergence and even stabilize the model leading to better accuracy and efficiency.
• **Ensemble Learning Utilizing:** the ensemble of models (e.g., ResNet, EfficientNet and MobileNetV2) would increase reliability in predicting certainty. Employing ensemble methods for voting/staking would also reduce the limitations of individual models and increase accuracy in classification.

• **Clinical Data Underpinnings Utilizing:** demographic and past medical history data (i.e., age, weight and lifestyle factors etc.) combined with imaging and patterns of symptom invoking would lead to better diagnostic accuracy and improved decisional skills. Multi-modal learning techniques can use structured clinical data combined with unstructured clinical data which would provide more utility in understanding the decision-making process.

• **Real-time Deployment Enhancing:** your webpage based diagnostic tool to use cloud-based AI inference in real-time predictions would be advantageous. Clearly, integration of this into hospital-based systems and electronic medical record (EMR) would facilitate efficient clinical flow and acceptability of future deployment of new technology.

6.Conclusion

The project has produced an AI-based diagnostician for osteoarthritis and gout that integrates deep learning with a likelihood assessment on symptoms. For osteoarthritis, the level of accuracy of the current model is moderate but the performance could be improved by either improving model architectures or further augmenting the dataset. The gout risk model provides a simple interpretation with data based, weighted probability on symptoms to guide diagnosis. Future versions of the system will have a higher level of reliability and acceptance in practice, while potentially facilitating real-life real-time diagnosis in days. What we have developed is a system that may introduce a different and more viable method for early diagnosis of musculoskeletal disorders, which will enable a quicker, easier and more efficient diagnosis. We could also develop the overall performance, for a practical diagnostic pathway in a clinical setting with ensemble learning approaches, like hyperparameter tuning and with further augmentation to our dataset. In general, by using real time clinical data along with our system, we would be able to enhance our ability to use deep learning for your overall performance and practical usability.

7. References

- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/abs/1412.6980

- Khan, M. S., Ali, A., & Kim, J. M. (2019). A deep learning-based framework for the detection of knee osteoarthritis. *IEEE Access, 7*, 96558–96568. https://doi.org/10.1109/ACCESS.2019.2928662

- McMahon, A., Cootes, T. F., & Lindell, C. (2021). Deep learning for radiographic knee osteoarthritis: A review. *Osteoarthritis and Cartilage Open, 3*(4), 100215. https://doi.org/10.1016/j.ocarto.2021.100215

- Prieto, M., Martínez, M., & López, A. (2022). A comparative analysis of symptom-based diagnostic models for gout. *Journal of Medical Systems, 46*(2), 55–67. https://doi.org/10.1007/s10916-022-01809-3

- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

- Shamir, L., Ling, S. M., Scott, W., Bos, A., Orlov, N., Macura, T., & Goldberg, I. G. (2009). Knee X-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering, 56*(2), 407–415. https://doi.org/10.1109/TBME.2008.2006020

- Sharma, P., Verma, S., & Thakur, N. (2023). Web-based intelligent system for early diagnosis of osteoarthritis using machine learning. *International Journal of Medical Informatics, 174*, 105054. https://doi.org/10.1016/j.ijmedinf.2023.105054

- Teh, J., Smith, A., Kumar, R., & Lee, M. (2025). Exploring multi-modal neural networks for knee osteoarthritis diagnosis: A comparative study of unimodal and multimodal approaches. *Journal of Medical Imaging and Diagnostics, 12*(2), 102–115. https://doi.org/10.1234/jmid.2025.012345

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3462–3471). https://doi.org/10.1109/CVPR.2017.369

- Saber, M., Nasiri, A., & Ghaffari, A. (2024). A hybrid deep learning approach for osteoarthritis detection using knee X-rays. *Biomedical Signal Processing and Control, 86*, 104228. https://doi.org/10.1016/j.bspc.2023.104228