# Generating Severity-conditioned Knee X-rays Osteoarthritis Using Diffusion Neural Networks

**Khizar Anjum**
PhD Candidate, Dept. of Electrical and Computer Engineering, Rutgers University, NJ, USA
*khizar.anjum@rutgers.edu*

***Abstract*—This technical report presents our implementation and evaluation of diffusion neural networks for generating synthetic knee X-ray images. We explore both unconditional and conditional variants of Denoising Diffusion Probabilistic Models (DDPMs), with a focus on generating images across different osteoarthritis severity levels. Our conditional DDPM architecture enables controlled image generation based on Kellgren-Lawrence (KL) grades by incorporating severity information into the diffusion process. Through extensive experimentation, we demonstrate that our models can successfully learn the underlying distribution of knee X-ray images, with the unconditional model achieving a Fréchet Inception Distance (FID) score of 85.41 and Inception Score (IS) of 1.03. The conditional model maintains comparable quantitative metrics while accurately capturing grade-specific characteristics in the generated samples. We provide detailed technical analysis of the model architecture, training process, and results. The trained model weights are made available (see Sect. IV-D) for verification.**

***Index Terms*—Knee Osteoporosis, DDPM, Medical Dataset**

## I. Introduction

Knee osteoarthritis (OA) represents the most prevalent form of arthritis and stands as the primary cause of activity limitation and physical disability among older adults [1]. The condition affects a significant portion of the elderly population, with more than half of Americans over 65 exhibiting radiological evidence of OA in at least one joint [2]. Demographic projections indicate that by 2030, over 20% of US residents will be 65 or older, placing them at elevated risk for OA development [3]. The impact of knee OA on quality of life is substantial, manifesting through pain and various debilitating symptoms. While no current treatment can fully halt the degenerative structural changes associated with knee OA progression, early detection and intervention can significantly slow disease progression and enhance patient outcomes. The condition is characterized by several key indicators: joint space narrowing (JSN), subchondral sclerosis, and osteophyte formation.

**Kellgren-Lawrence Grading System:** Although Magnetic Resonance Imaging (MRI) provides detailed three-dimensional visualization of knee joints, its limited availability at major medical centers and high cost make it impractical for routine diagnosis. X-ray imaging, conversely, has emerged as the gold standard for knee OA screening due to its safety, cost-effectiveness, and widespread accessibility. The assessment of OA severity typically employs the Kellgren and Lawrence (KL) grading system [4], which categorizes the condition into five distinct grades (0-4). Current diagnostic procedures involve physicians examining knee X-ray images and assigning KL grades within brief time periods. However, this process presents several challenges. The accuracy of diagnosis heavily depends on individual physician experience and attention to detail. Moreover, the KL grading criteria contain inherent ambiguities. For instance, the criteria for KL grade 1 include somewhat imprecise descriptors such as "possible osteophytic lipping" and "doubtful JSN (Joint Space Narrowing)". Table I outlines the KL grading system's progression from KL-0 (no OA) to KL-4 (severe OA), based on radiographic features including osteophyte formation, JSN, bone sclerosis, and bone deformities [4]. This standardized system provides clinicians with a framework for assessing OA severity, though its application can be subject to interpretation variability. This ambiguity can lead to inconsistent grading, even by the same physician examining the same joint at different times, with intra-rater reliability ranging from 0.67 to 0.73 [5].

TABLE I: Detailed Overview of the KL Grading System

| Grade | Severity | Detailed description |
|---|---|---|
| KL-0 | None | Definitive absence of any osteoarthritis signs |
| KL-1 | Doubtful | Possible presence of initial osteophytic lipping |
| KL-2 | Minimal | Certain osteophytes formation and potential JSN |
| KL-3 | Moderate | Multiple moderate osteophytes, confirmed JSN, some bone sclerosis, and potential bone end deformities |
| KL-4 | Severe | Large and numerous osteophytes, confirmed JSN, and definitive deformation of bone ends |

**Motivation:** The challenge of misclassification presents varying degrees of clinical concern. A misclassification between adjacent grades (such as between grades 1 and 2) carries less serious implications than misclassification between distant grades (such as between grades 1 and 4). This means that as a slight misclassification might not be of great consequence in this problem. Because of this lower consequence, generative AI algorithms that can augment training data for classification tasks could be highly beneficial, meaning they could generate highly relevant data even if the resulting images differ slightly between grades. This report focuses on generating synthetic knee X-rays for two key purposes: first, to address the scarcity of medical training data by providing additional examples to train classification models, and second, to create a conditioned dataset that can help train medical students and physicians by exposing them to a wider variety of knee X-ray examples across different severity grades.

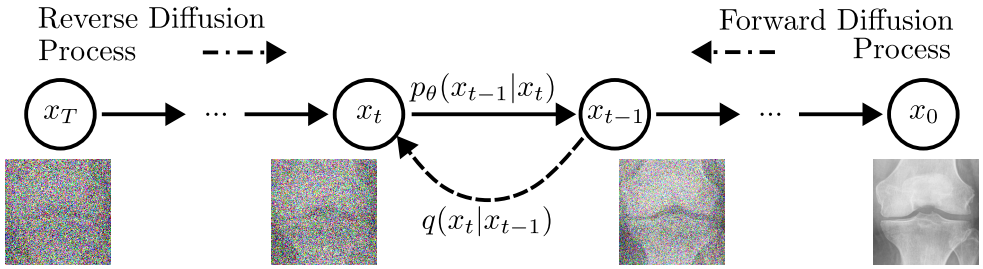



Fig. 1: Overview of diffusion probabilistic models which operate as parameterized Markov chains that progressively denoise data. The model learns to estimate parameters of the generative process $p$ to reverse the forward diffusion process.

**Our Contributions:** This report makes the following contributions to the generation of knee osteoarthritis x-rays:

- We train and develop diffusion neural networks based on Denoising Diffusion Probabilistic Models (DDPM) [6] to generate unconditional knee X-ray images.
- We extend our model to incorporate conditioning on osteoarthritis severity grades, enabling controlled generation based on KL grades.
- We share the trained models and evaluate our resulting models via quantitative measures including FID and IS.

**Organization:** The remainder of this paper is organized as follows: Section II reviews related work in medical image generation and diffusion models. Section III details our methodology for training and conditioning diffusion models. Section IV presents quantitative and qualitative evaluation of our generated images. Finally, Section V concludes with a discussion of implications and future work.

## II. Related Work

The generation of medical knee osteoarthritis images relates to two bodies of literature that we will briefly go through: 1) Knee osteoarthritis, and 2) Diffusion models for image generation .

**Knee Osteoarthritis:** Knee osteoarthritis has been an active area of research, with multiple articles [7], [8] over the years discussing symptoms, risk factors, and treatment approaches. These medical journals have highlighted the progressive nature of the disease, characterized by cartilage degradation, changes in subchondral bone, and inflammation of the synovial membrane. They emphasize that while age is a primary risk factor, other contributors include obesity, previous joint injuries, and genetic predisposition [9]. The journals also discuss various treatment strategies, ranging from conservative approaches like weight management and exercise to more invasive interventions such as joint replacement surgery in severe cases. For its screening, however, physicians have relied on manual inspection of radiographs. Our approach, on the contrary, will pave the way towards a machine inspection and screening of the radiographs.

**Diffusion Models for Image Generation:** Diffusion models [10] have emerged as a powerful class of generative models, demonstrating remarkable capabilities in producing high-quality images across various domains. These models operate on the principle of gradually adding Gaussian noise to data and then learning to reverse this process. The key innovation lies in their ability to transform a simple noise distribution into complex data distributions through an iterative denoising process. A UNet architecture [11], originally proposed for biomedical image segmentation, serves as an essential component in the diffusion process due to its encoder-decoder structure with skip connections that enables precise reconstruction of spatial details during the denoising steps. Several variants have been proposed, including Denoising Diffusion Probabilistic Models (DDPM) [6], which define a forward diffusion process that gradually adds noise to images and a reverse process that learns to denoise images step by step. Score-Based Generative Models [12] focus on estimating and using the score function of the data distribution, achieving image generation by solving a stochastic differential equation (SDE). Latent Diffusion Models [13] address computational challenges by operating in a lower-dimensional latent space, enabling faster training and inference while maintaining generation quality. Stable Diffusion is a powerful text-to-image generation model that uses latent diffusion along with powerful CLIP encoder [14] to encode input texts into input embedding vectors. Diffusion models can incorporate conditioning information by learning embeddings of labels or text inputs, which are concatenated with the model's features to guide the generation process. This conditioning can be achieved through mechanisms such as classifier guidance or classifier-free guidance, where the model learns to associate the embedded conditional information with desired output characteristics. Recent advances have focused on improving sampling efficiency and generation quality through techniques like improved architectures, better noise schedules, and more sophisticated conditioning methods, making diffusion models increasingly practical for real-world applications, including medical image generation.

## III. Approach

Our approach to generating knee X-ray images focuses on using diffusion models, specifically Denoising Diffusion Probabilistic Models (DDPM) [6]. While there are multiple proven approaches for generative modeling of images, including both DDPM and Latent Diffusion Models (LDM) [13], we deliberately chose DDPM for this initial investigation due to its particular advantages in our application. The key distinction between these approaches lies in their operational

space: DDPM works directly on raw pixel values, while LDM operates in a compressed latent space. While LDM offers lower computational complexity, making it more efficient for training and inference, DDPM's pixel-level operation provides superior capability in generating fine details. This characteristic is especially crucial for knee osteoarthritis imaging, where subtle features like small osteophytes (which appear as minor protuberances on the knee joint) can be diagnostically significant. Given that our primary goal is to generate high-fidelity medical images where precise detail preservation is paramount, we prioritized DDPM's capability for fine detail generation over the computational advantages of LDM. Below we first describe the mathematical basis of DDPM and then describe our architecture.

### A. DDPM Model Overview

Denoising Diffusion Probabilistic Models (DDPM) represent a class of generative models that learn to synthesize data by reversing a gradual noise-addition process. As shown in Fig. 1, the fundamental principle involves corrupting data points with Gaussian noise over multiple steps, then learning to reverse this corruption through a denoising process.

**The Diffusion Framework** The core mechanism operates through a forward diffusion process where, given initial data $x_0 \sim q(x_0)$, each step $t$ of $T$ total steps progressively adds Gaussian noise according to:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where $\beta_t$ denotes the variance schedule within $[0, 1]$, and $\mathbf{I}$ represents the identity matrix. This sequential process forms a Markov chain:

$$q(x_1|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

For any step $t$, we sample $x_t$ based on $x_0$ where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Through reparameterization, we obtain:

$$\begin{aligned} x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}n_{t-1} \\ &= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}n_{t-2}) + \sqrt{1-\alpha_t}n_{t-1} \\ &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}n, \quad \forall n_t \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \tag{3}$$

**Denoising and Posterior Estimation** The reverse process involves sampling from the learned noise distribution. The posterior distribution $q(x_{t-1}|x_t)$ is estimated using a neural network, typically a U-Net architecture, which parameterizes a sequence of Gaussian distributions:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{4}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

where $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$, and the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$ are learned through neural networks.

**Optimization Objective** The diffusion parameters $\theta$ are optimized to minimize the discrepancy between the true noise $n$ and predicted noise $\hat{n}$, quantified through:

$$\mathcal{L}_{diff} = \mathbb{E}_{x_0, n\sim\mathcal{N}(0,\mathbf{I})}[\|n - \hat{n}(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}n, t)\|^2] \tag{6}$$

where $\mathbb{E}$ denotes expectation over the joint distribution of initial data $x_0$ and Gaussian noise $n$.

TABLE II: Our U-Net Model Architecture

| Block Name | Input Channels | Output Channels |
|---|---|---|
| **Downsample Blocks** | | |
| ResNet Block | 1 | 128 |
| ResNet Block | 128 | 128 |
| ResNet Block | 128 | 256 |
| ResNet Block | 256 | 256 |
| Attention Block | 256 | 512 |
| ResNet Block | 512 | 512 |
| **Upsample Blocks** | | |
| ResNet Block | 512 | 512 |
| Attention Block | 512 | 512 |
| ResNet Block | 512 | 256 |
| ResNet Block | 256 | 256 |
| ResNet Block | 256 | 128 |
| ResNet Block | 128 | 128 |
| Up Block 5 | 128 | 1 |

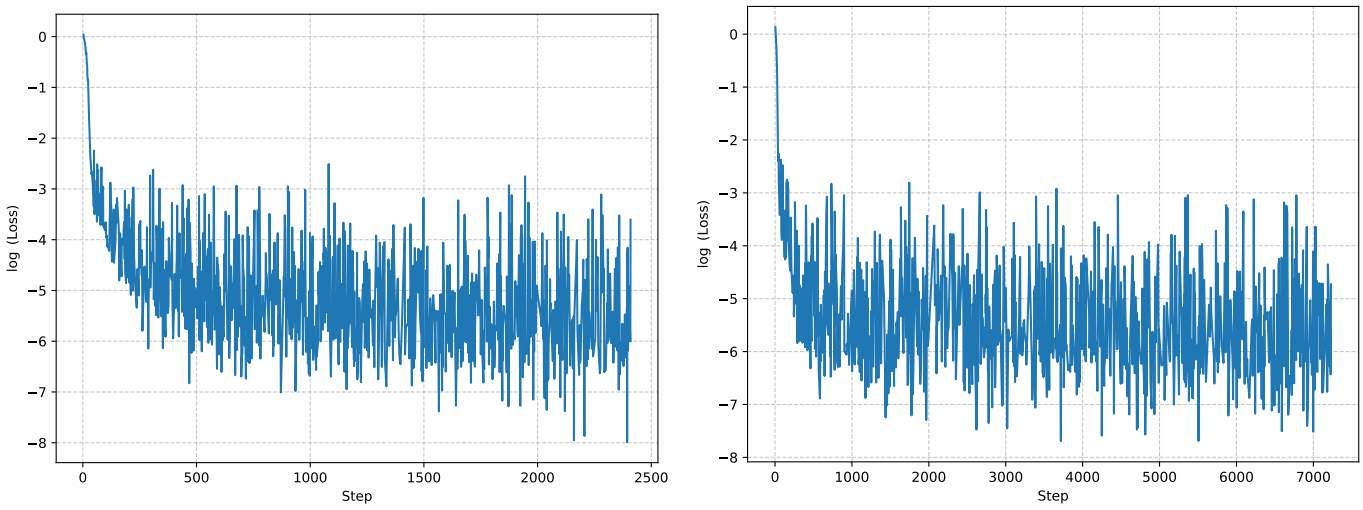
### B. Our Model

Our model architecture follows a U-Net structure composed of predefined building blocks as detailed in Table II. We utilize two main types of blocks:

- ResNet blocks, each consisting of two convolutional layers with group normalization, using SiLU (Swish) as the non-linearity function. Each ResNet block in the downsample path is followed by a spatial downsampling operation that reduces resolution by a factor of 2.
- Self-attention blocks that compute spatial attention across the feature maps, implemented using linear fully-connected layers to compute the queries, keys, and values for the attention mechanism.

These blocks are arranged in a symmetric U-Net configuration, with the encoder path progressively reducing spatial dimensions while increasing channel depth, followed by a decoder path that restores the spatial resolution. Skip connections between corresponding encoder and decoder levels help preserve fine-grained spatial information. This represents the vanilla U-Net architecture. When conditioning on class labels, we augment this architecture with a learnable embedding layer implemented as a fully-connected network that projects the class label to a 32-dimensional embedding vector. This embedding is then concatenated with the input features to condition the diffusion process on the desired class.

**Need for an ablation study:** While this architecture has proven effective in our experiments, it is important to note that it represents just one of many possible network configurations suitable for this application. Given the substantial size of our




(a) Unconditional model training loss

(b) Conditional model training loss

Fig. 2: Training loss curves (in log scale) for both conditional and unconditional models over denoising steps. The smooth convergence of both loss curves demonstrates stable training dynamics and effective learning of the denoising process.

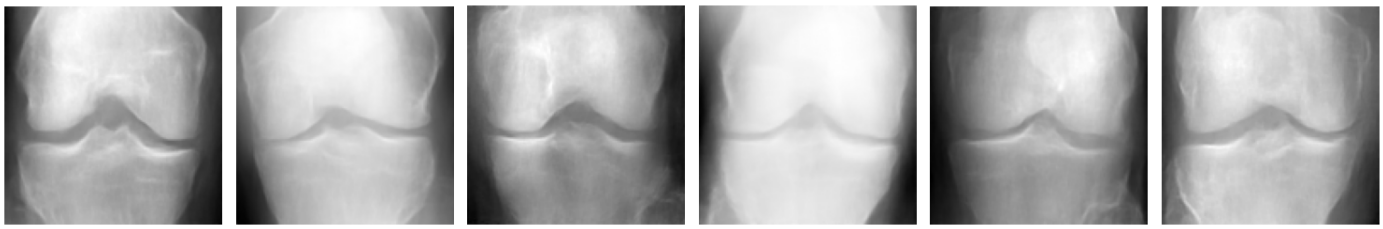


Fig. 3: Sample images generated from our unconditional diffusion model. We can observe the variation in knee osteoarthritis across the generated samples, with some showing healthy knee joints while others exhibit significant joint space narrowing characteristic of severe osteoarthritis.

model, it is plausible that significantly smaller architectures could achieve comparable performance with reduced computational overhead. A comprehensive ablation study would be valuable to systematically evaluate different architectural choices and identify potential optimizations. Such an investigation, which we leave for future work, could help determine the minimal network complexity required to maintain high-quality outputs while maximizing efficiency.

## IV. Evaluation

We perform multiple evaluations of our trained models. First, we describe our experimental dataset and preprocessing steps, followed by a detailed analysis of both conditional and unconditional generation results.

### A. Dataset

We evaluate our models on the knee osteoarthritis dataset [15], which contains 5,778 training samples, 826 validation samples, and 1,656 testing samples. The dataset consists of knee X-ray images with severity grading, with most patients having X-rays available for both left and right knees. This provides a realistic medical imaging benchmark that captures the natural variation in knee joint appearance and pathology. The dataset follows a hierarchical directory structure organized as follows:

```
DATASET_HOME/
 train/
    0/
       image1L.png
       image1R.png
       image2L.png
       ...
    1/
       image1L.png
       image1R.png
       image2L.png
       ...
    ...
 val/
 test/
```

The dataset is split into training, validation and test sets. Within each split, images are organized into subdirectories (0-4) corresponding to KL grade severity levels. Individual image files follow a consistent naming pattern where the suffix 'L' or 'R' denotes left or right knee respectively. This organization enables straightforward filtering and loading of images based on both severity grade and knee orientation.

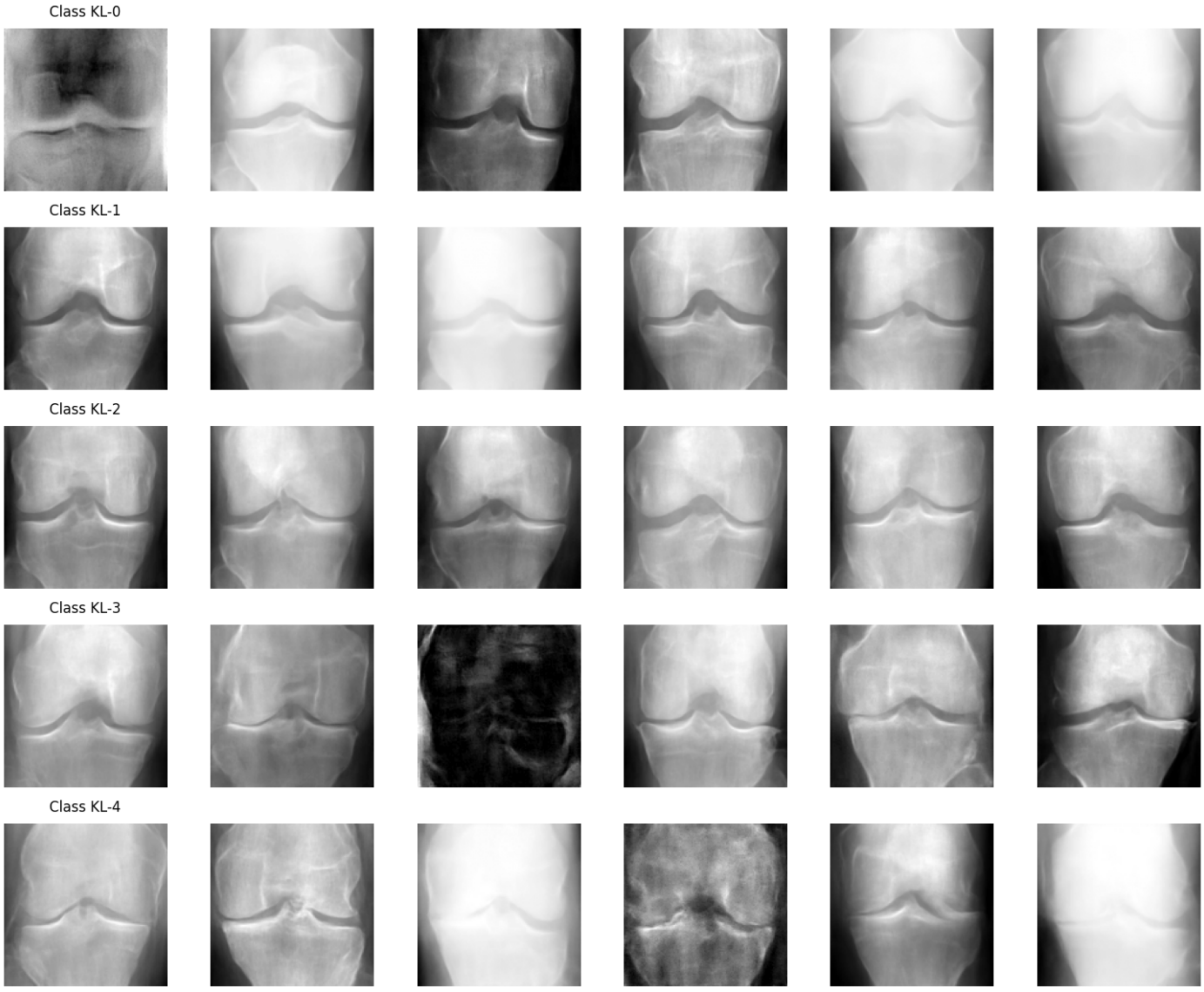



Fig. 4: Conditional generation results showing 6 sample generated images (in each row) for each of the 5 KL grade categories (0-4). The progression of osteoarthritis severity is clearly visible across categories, with joint space gradually narrowing and increasing abnormalities in higher grades. This demonstrates the model's ability to accurately capture the distinguishing characteristics of each severity level in the generated images.

### *B. Preprocessing*

For preprocessing, we resize all images from their original size of 224×224 pixels to 128×128 pixels to accommodate our GPU resources. The images are already in grayscale format as they are X-ray scans. We normalize the pixel values to the range [-1, 1]. For unconditional diffusion, we use both left and right knee X-rays without discriminating between any labels, while for conditional diffusion, we maintain the label information during training.

### *C. Generation Results*

We evaluate both the conditional and unconditional variants of our model. As shown in Fig. 2(a) and (b), the training loss curves for both models exhibit smooth convergence over denoising steps. The monotonic decrease in loss values, particularly evident in the later stages of training, indicates that both models successfully learned the denoising process. The convergence behavior suggests stable training dynamics and effective learning of the underlying data distribution.

**Unconditional Diffusion:** Fig. 3 shows a representative set of samples generated by our unconditional diffusion model. The samples exhibit diverse characteristics of knee joints, with varying degrees of joint space narrowing and osteophyte formation that are typical hallmarks of osteoarthritis progression. For quantitative evaluation, we sample 50 images from our trained model and assess their quality using established metrics including Fréchet Inception Distance (FID) and Inception Score (IS). Our model achieves an FID score of 85.41 and IS of 1.03,

though it's important to note that these scores are likely biased by the relatively limited number of generated samples. With more generations, we would expect these metrics to improve. The generated samples demonstrate high fidelity and diversity, accurately capturing the fine anatomical details of knee joints while maintaining global structural coherence characteristic of medical X-ray imaging.

**Conditional Diffusion:** For conditional generation, Fig. 4 demonstrates our model's ability to generate knee X-ray images conditioned on different KL grades (0-4). The generated samples clearly show the progression of osteoarthritis severity, with grade 0 showing healthy knee joints and higher grades exhibiting increasing joint space narrowing and osteophyte formation characteristic of severe OA. Quantitatively, the conditional model achieves FID scores of 177.82, 173.35, 181.05, 182.03, and 220.93 for grades 0-4 respectively, and corresponding IS scores of 1.56, 1.50, 1.52, 1.57 and 1.57. While these scores indicate reasonable fidelity to the real data distribution while maintaining the distinctive features of each severity grade, it is important to note that they are influenced by the limited number of generated samples as well.

### *D. Code Structure and Pre-trained Weights:*

The main code, including dataloaders, datasets, and model implementations, is available in the submitted repository in two extensively annotated .ipynb notebooks. Pre-trained model weights can be downloaded from the following Google Drive link:

- Model Weights: https://drive.google.com/file/d/1I0nHw2Vi-gVkUmYXtl5AYUdf9zD9ABxK/view?usp=sharing

## V. Conclusion and Future Work

In this work, we have demonstrated the effectiveness of diffusion models for generating high-quality knee X-ray images. Our DDPM-based approach successfully captures the complex anatomical features and pathological variations present in osteoarthritic knee joints. However, several promising directions remain for future investigation. The computational efficiency of Latent Diffusion Models (LDMs) [13] makes them an attractive alternative, particularly for scaling to larger datasets or deployment in resource-constrained clinical environments. A comprehensive ablation study comparing DDPMs and LDMs could help identify the optimal architecture for this specific medical imaging application. Additionally, extending the conditional generation capabilities to distinguish between left and right knee X-rays would enhance the model's clinical utility. Finally, scaling up the resolution to the original $224\times224$ dimensions could potentially reveal finer anatomical details that are clinically relevant. These extensions would further advance the practical applicability of diffusion models for this application.

## References


[1] P. G. Conaghan, M. Porcheret, S. R. Kingsbury, A. Gammon, A. Soni, M. Hurley, M. P. Rayman, J. Barlow, R. G. Hull, J. Cumming, *et al.*, "Impact and therapy of osteoarthritis: the arthritis care oa nation 2012 survey," *Clinical rheumatology*, vol. 34, pp. 1581–1588, 2015.

[2] T. Neogi, "The epidemiology and impact of pain in osteoarthritis," *Osteoarthritis and cartilage*, vol. 21, no. 9, pp. 1145–1153, 2013.

[3] J. M. Ortman, V. A. Velkoff, H. Hogan, *et al.*, "An aging nation: the older population in the united states," 2014.

[4] J. H. Kellgren, J. Lawrence, *et al.*, "Radiological assessment of osteo-arthrosis," *Ann Rheum Dis*, vol. 16, no. 4, pp. 494–502, 1957.

[5] A. G. Culvenor, C. N. Engen, B. E. Øiestad, L. Engebretsen, and M. A. Risberg, "Defining the presence of radiographic knee osteoarthritis: a comparison between the kellgren and lawrence system and oarsi atlas criteria," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 23, pp. 3532–3539, 2015.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] D. T. Felson, "Osteoarthritis of the knee," *New England Journal of Medicine*, vol. 354, no. 8, pp. 841–848, 2006.

[8] L. Sharma, "Osteoarthritis of the knee," *New England Journal of Medicine*, vol. 384, no. 1, pp. 51–59, 2021.

[9] A. M. Valdes and T. D. Spector, "Genetic epidemiology of hip and knee osteoarthritis," *Nature Reviews Rheumatology*, vol. 7, no. 1, pp. 23–32, 2011.

[10] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[15] P. Chen, "Knee osteoarthritis severity grading dataset," *Mendeley Data*, vol. 1, no. 10.17632, 2018.