BMC Medical Imaging

RESEARCH

Open Access

Comparative evaluation of generative artificial intelligence models for synthetic knee radiograph augmentation in clinical research

Kwangho Chung[1†], Ji-Hoon Nam[2,3,4†], Arailym Dosset[2], Yong-Gon Koh[5], Jae Min Kim[6], Paul Shinil Kim[7], Jin Woo Lee[8], Kyoung-Mi Park[3], Hyuck Min Kwon[8*†] and Kyoung-Tak Kang[2,3,4*†]

Abstract

Background In this study, the capability of state-of-the-art generative models to synthesize realistic knee radiographs was evaluated to address dataset scarcity in osteoarthritis (OA) research.

Methods Three generative frameworks—Style Generative Adversarial Network3 (StyleGAN3), a stable diffusion + Cycle-consistent Generative Adversarial Network (CycleGAN) pipeline, and Deep Convolutional Generative Adversarial Network (DCGAN)—were trained on 10,042 real knee X-rays. Image quality was assessed using Fréchet Inception Distance (FID) while visual fidelity was evaluated via a Visual Turing Test conducted by two orthopedic surgeons and a musculoskeletal radiologist. Joint Line Convergence Angle (JLCA) was compared between real and synthetic images for anatomical fidelity. Inter- and intra-observer reliability for JLCA was measured using intraclass correlation coefficients (ICC).

Results StyleGAN3 achieved the best performance (FID 10.84), showing high visual and anatomical fidelity. Integrating Stable Diffusion with CycleGAN showed a moderate FID of 39.79, suggesting that adversarial enhancements improved the diffusion-based synthesis. DCGAN showed lower quality, achieving an FID of 74.15. Expert accuracy in distinguishing real from synthetic images ranged between 36% and 88%, confirming difficulty in visual differentiation. Furthermore, JLCA measurements showed no significant difference between real (4.19 ± 3.07°) and synthetic (3.36 ± 2.19°) images generated by DCGAN ($p = 0.12$). Similarly, Diffusion + CycleGAN (3.91 ± 2.59° vs. 3.72 ± 2.52°, $p = 1.00$) and StyleGAN3 (4.27 ± 3.01° vs. 3.60 ± 2.37°, $p = 0.25$) showed no statistically significant differences. These results indicate that all elevated generative models maintained high anatomical fidelity relative to

[†]Kwangho Chung and Ji-Hoon Nam contributed equally to this work and share first authorship.

[†]Hyuck Min Kwon and Kyoung-Tak Kang contributed equally to this work and share corresponding authorship.

*Correspondence:
Hyuck Min Kwon
HYUCK7777@yuhs.ac
Kyoung-Tak Kang
tagi1024@gmail.com

Full list of author information is available at the end of the article

real radiographs. Inter-observer agreement was strong, with ICC values ranging between 0.83 and 0.97. Intra-observer reliability was also excellent.

Conclusion StyleGAN3 generated the most realistic knee radiographs. Diffusion-based pipelines showed promising results when enhanced with adversarial networks. These findings underscore the potential of generative AI to mitigate data limitations in orthopedic research.

Keywords X-ray, DCGAN, StyleGAN3, CycleGAN, Knee

Background

Osteoarthritis (OA) is a major global health burden. In 2020, an estimated 7.6%, approximately 595 million individuals of the global population, were affected by OA, with over 240 million experiencing symptomatic activity-limiting disease [1, 2].

However, radiograph-based severity assessment using the Kellgren-Lawrence (KL) grading system can be inconsistent because of inter- and intra-observer variability, especially in borderline cases and in multicenter datasets with heterogeneous acquisition protocols [3–5]. These issues motivate the development of robust imaging algorithms, but such efforts are often constrained by limited access to large, diverse, and balanced datasets across disease severities [6].

Recent progress in deep learning has enabled automated analysis of knee radiographs [7], yet model performance and generalizability depend heavily on the quantity and diversity of training data [7, 8]. Curating large radiograph datasets is costly and time-consuming because expert annotation is required, and class imbalance is common. Generative modeling provides a complementary approach by synthesizing realistic images that can potentially enrich under-represented patterns and support stress-testing of downstream algorithms. Importantly, the clinical value of synthetic radiographs depends not only on visual realism but also on preservation of anatomical structures relevant to clinical measurements [9].

Overcoming these obstacles requires novel techniques, including generative adversarial networks (GANs), which can produce synthetic images that closely mimic real ones.

GANs for imaging generally use convolutional architectures and includes two neural networks trained in a competitive fashion through a minimax optimization process [8]. Previous studies used deep convolutional generative adversarial networks (DCGAN) to generate synthetic images for three distinct categories of liver lesions: cysts, metastases, and hemangiomas [10]. Recently, diffusion methods, such as Stable Diffusion, have used a pre-trained variational autoencoder to project images into a latent space, a transformer-based text encoder for semantic conditioning, and a UNet denoiser for high-fidelity reconstruction [11, 12]. However, these models can produce blurred edges or minor anatomical inconsistencies. Pairing them with Cycle-consistent Generative Adversarial Network (CycleGAN) unpaired image-to-image translation enforces cycle consistency between synthetic and real domains and results in sharper domain-adapted radiographs without requiring pixel-aligned training pairs [13, 14]. Additionally, Style Generative Adversarial Networks (StyleGAN3) alias-free architectural innovations deliver significant enhancements in image fidelity and spatial coherence, which are essential for preserving the anatomical integrity of medical images [15]. Recently, studies in medical imaging have confirmed StyleGAN3's ability to generate high-fidelity images with precise anatomical details, outperforming earlier GAN variants and reducing artifacts that can compromise clinical utility [16, 17]. These strengths show that StyleGAN3 is particularly well-suited for synthesizing realistic knee radiographs for OA research. However, no quantitative evaluation of how well these state-of-the-art deep-learning methods generate synthetic images of knee-joint OA was observed.

Therefore, this study aimed to evaluate and compare three generative frameworks–DCGAN, a Stable Diffusion + CycleGAN pipeline, and StyleGAN3—on a knee X-ray dataset. Visual realism was assessed using the FID, expert-based visual Turing tests, and anatomical accuracy through Joint Line Convergence Angle (JLCA) measurement and reliability analysis. We hypothesized that StyleGAN3 would produce the most realistic and anatomically accurate synthetic knee radiographs among the evaluated models.

Methods

In this section, the datasets and methods used to generate high-resolution knee images are described. The flow chart of the proposed model is illustrated in Fig. 1.

Dataset

Our Institutional Review Board approved this study before it was conducted. Radiographic images were obtained from 10,042 consecutive patients who underwent total knee arthroplasty (TKA) at our institution. The dataset consisted of 10,042 patients (Females: $n = 7{,}843$, Males: $n = 2{,}199$) with a mean age of 70.3 ± 6.60 years. Radiographically, the cohort predominantly consisted of

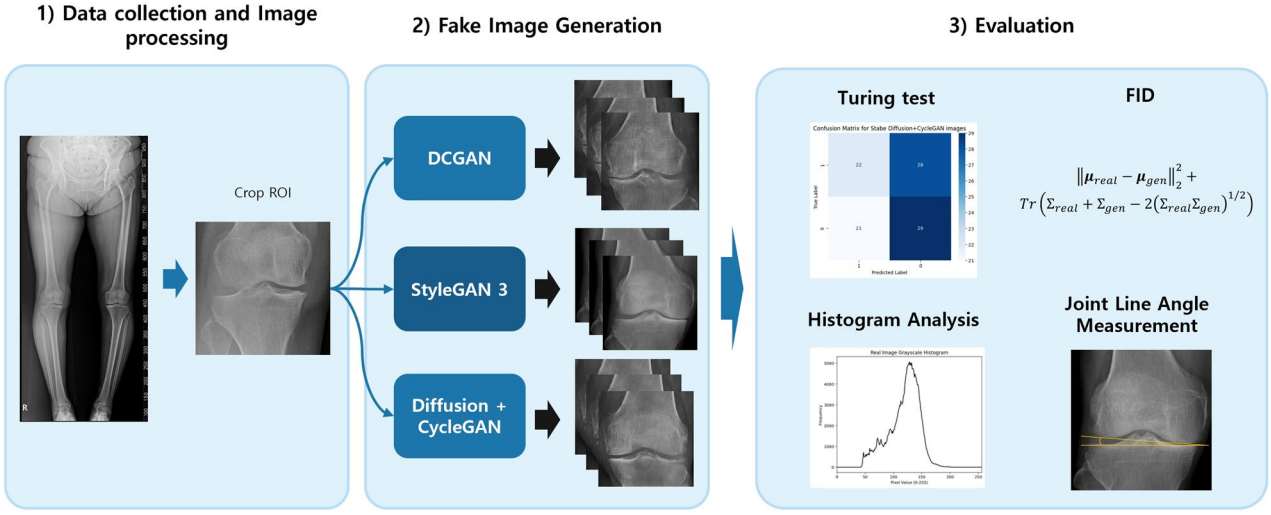

Fig. 1 Overview of the proposed workflow for data preprocessing, image synthesis, and evaluation

advanced osteoarthritis cases classified as Kellgren-Lawrence (KL) grade 3 or 4. Patients with a history of bony surgery or joint replacement, which could have altered the femoral or tibial geometry, were excluded. The dataset included two types of radiographic images obtained from different imaging systems. First, full-leg plain radiographs of the recruited patients were acquired using a conventional radiography system. Second, a tomosynthesis-enabled X-ray system (Sonialvision G4; Shimadzu, Kyoto, Japan) reconstructed a single panoramic view by sequentially acquiring approximately 5 cm-wide narrow-slit projections with a flat-panel detector moving parallel to the X-ray tube. To ensure the model focused on relevant anatomical features, all X-ray images were manually cropped to isolate the knee joint region, removing unnecessary background and artifacts. These cropped images were then resized to 512 × 512 pixels. Apart from this cropping and resizing, no further image preprocessing or intensity normalization techniques were applied to preserve the original radiographic characteristics.

Generative adversarial networks

The GAN framework involves a creative learning process in which two models are simultaneously trained. The key design concept of a GAN is that the generator creates images that mimic real knee X-rays, whereas the discriminator identifies whether the sample comes from the training data or the generator. The generator continuously improved its ability to produce convincing images by learning from the feedback provided by the discriminator.

Meanwhile, the discriminator sharpens its detection skills to better distinguish between genuine and synthetic images. This interaction is considered a sophisticated guessing game, where both opponents strive to outsmart each other. The generator aims to maximize the discriminator's error rate, whereas the discriminator attempts to minimize its errors.

For research purposes, we used three distinct GAN frameworks to generate knee radiographs: DCGAN, StyleGAN3, and the integration of Stable Diffusion and CycleGAN models. DCGAN was chosen as a standard baseline to represent early adversarial architectures. StyleGAN3 was selected to evaluate the limits of state-of-the-art unconditional GANs which generate images from random noise. Finally, the Stable Diffusion + CycleGAN model was included to investigate the potential of text-guided latent diffusion models, contrasting their prompt-based generation capability against the noise-based approaches of traditional GANs.

Deep convolutional generative adversarial networks

The DCGAN showed considerable success in the GAN framework, which was specifically designed for image synthesis tasks. The DCGAN architecture replaces traditional fully connected dense layers with convolutional and transposed-convolutional layers, implements batch normalization in the generator and discriminator, and utilizes leaky ReLU activation to prevent gradient vanishing during training. The input to the generator network is random noise sampled from a standard normal distribution, and the output generated mimics images.

Style generative adversarial networks 3

StyleGAN3 introduces an alias-free architecture that considerably improves image quality and spatial consistency, an important requirement in the medical field, where the structural accuracy of the generated images is essential [18].

The Generator comprises two main components: a mapping and a synthesis network. The mapping network allows the transformation of a random latent vector z into an intermediate latent vector w. This vector w modifies each layer of the synthesis network, enabling multiscale control over image features, including texture, shape, and fine details. The synthesis network progressively generates images from low to high resolution using a series of convolutional blocks. Each block integrates nonlinear activation and upsampling processes. The primary innovation of StyleGAN3 is replacing traditional upsampling with a filter-based approach, as it helps in maintaining anatomical accuracy, preventing distortions, and improving overall image quality. StyleGAN3 showed a strong performance in generating high-fidelity images in recent medical studies [15, 17], making it ideal for healthcare applications.

Stable diffusion

Owing to advancements in generative models, diffusion models are considered robust and reliable alternatives to GAN, showing promising results in medical imaging. Recent studies showed the effectiveness of these models in synthesizing medical images [7, 19].

In this study, we fine-tuned a Stable Diffusion v1.5 model to generate high-resolution knee radiographs from text prompts. Stable Diffusion is a latent diffusion model that uses a pre-trained VAE to encode an image into a lower-dimensional latent space. It uses a CLIP-based transformer to embed text prompts, and a UNet architecture to iteratively denoise latent representations guided by these text prompts. We adapted and optimized the UNet component while keeping the text encoder fixed during training to enable high-quality synthesis.

Cycle generative adversarial networks

CycleGAN has become popular owing to its ability to perform unpaired image-to-image translation, which is beneficial in the medical imaging field [20]. It utilizes a "round-trip" process that learns meaningful transformations between source and target domains without requiring paired data. CycleGAN uses two Generators and Discriminators: one generator translates images from domain A to domain B, while the other performs the reverse mapping. Each generator is paired with a discriminator that attempts to differentiate real images in each domain from the generated images, facilitating adversarial learning.

Image synthesis

We trained the three different models using 10,042 images. First, the DCGAN model is trained for 550 epochs. The hyperparameters of this model included a batch size of 32, a latent vector z of 256, a number of feature maps for the generator of 64, a discriminator of 32, a learning rate of 0.002, and a beta1 hyperparameter for the Adam optimizer of 0.5. Second, StyleGAN3 was trained utilizing the StyleGAN3-r configuration with a batch size of 16. The training process continued until 1,000k image and image mirroring was disabled. To monitor progress, model snapshots and sample outputs were saved every 10k image. The model was optimized using a non-saturating logistic adversarial loss, where the generator learns to produce images that the discriminator classifies as real, while the discriminator learns to distinguish real images from generated samples. To stabilize training and reduce discriminator overfitting, R1 gradient penalty regularization was applied to real images with a regularization weight of $\gamma = 8.2$. Both the generator and discriminator were optimized using the Adam optimizer following the official StyleGAN3 implementation. The third approach utilizes a two-stage generative framework that incorporates latent diffusion models and cycle-consistent adversarial networks. In the first step, we used a pre-trained Stable Diffusion v1-5 architecture from the HuggingFace libraries. The training process was conducted using the AdamW optimizer with a learning rate of 1e-5, utilizing mixed-precision FP16 acceleration to improve the computational performance. The model was trained for 10 epochs. The optimal checkpoint was selected from the 8th epoch based on a rigorous qualitative visual assessment. We prioritized the anatomical integrity of bone structures and the absence of generation artifacts, identifying this epoch as yielding the most clinically realistic radiographic features. Text prompts were used in the training and image generation processes. During the training stage, the text prompt was "a high-quality X-ray image of a knee," while in the inference stage, prompts helped guide the generation of X-ray images. In the second stage, we use CycleGAN, an unpaired image-to-image translation framework. We refer to Domain A as the generated images from the fine-tuned Stable Diffusion model, whereas Domain B is our original dataset. The purpose of CycleGAN is to convert the images from Domain A into those of Domain B, thereby improving the quality of the generated images. To provide a comparative baseline, we also calculated the FID for the images generated by the Stable Diffusion model alone, prior to the CycleGAN refinement stage.

Visual turing test

To assess the realism of the synthesized images, a Visual Turing Test was conducted involving two expert orthopedic surgeons and one musculoskeletal radiologist. For each of the three generative models, each expert was presented with a randomly selected set of 100 knee radiographs, consisting of 50 real images and 50 synthetic images generated by the respective model. The

participants were informed beforehand that the set contained an equal number of real and synthetic images. This prior disclosure was intended to minimize prevalence bias, ensuring that evaluators focused strictly on visual quality and anatomical details rather than relying on probability guessing in uncertain scenarios. Their task was to classify each image as real (Class 1) or generated (Class 0). We recorded their evaluations and analyzed the results to determine whether medical professionals could reliably distinguish between authentic and AI-generated samples.

Fréchet inception distance

The Fréchet Inception Distance (FID) was used to compare the distributions of the real and synthesized images and evaluate the realism of the generated images. FID has shown a strong alignment with human perceptual judgments. In this study, we calculated the FID between 10,042 real knee radiographs and 10,042 generated images for each of the generative models.

Joint line convergence angle measurement

To assess the anatomical consistency and clinical use of the synthesized knee radiographs, we conducted JLCA measurements (Fig. 2). The JLCA is an important radiographic parameter for evaluating limb alignment. Accurate replication of this angle in the synthetic images shows that the generative models preserve critical anatomical landmarks. For this analysis, 100 real and 100 synthetic images were randomly selected for each generative model. Their JLCAs were measured. This comparison allowed us to compare the structural accuracy of the generated images with real radiographs and evaluate the potential of the synthetic data for use in clinical decision-making.

Statistical analysis was performed to compare the JLCA values between the real and synthetic groups for each model. The normality of the data distribution was first assessed using the Kolmogorov–Smirnov test. Based on the distribution, group comparisons were conducted using the independent samples t-test or the Mann–Whitney U test, as appropriate. To account for multiple comparisons across the three generative models (Real vs.

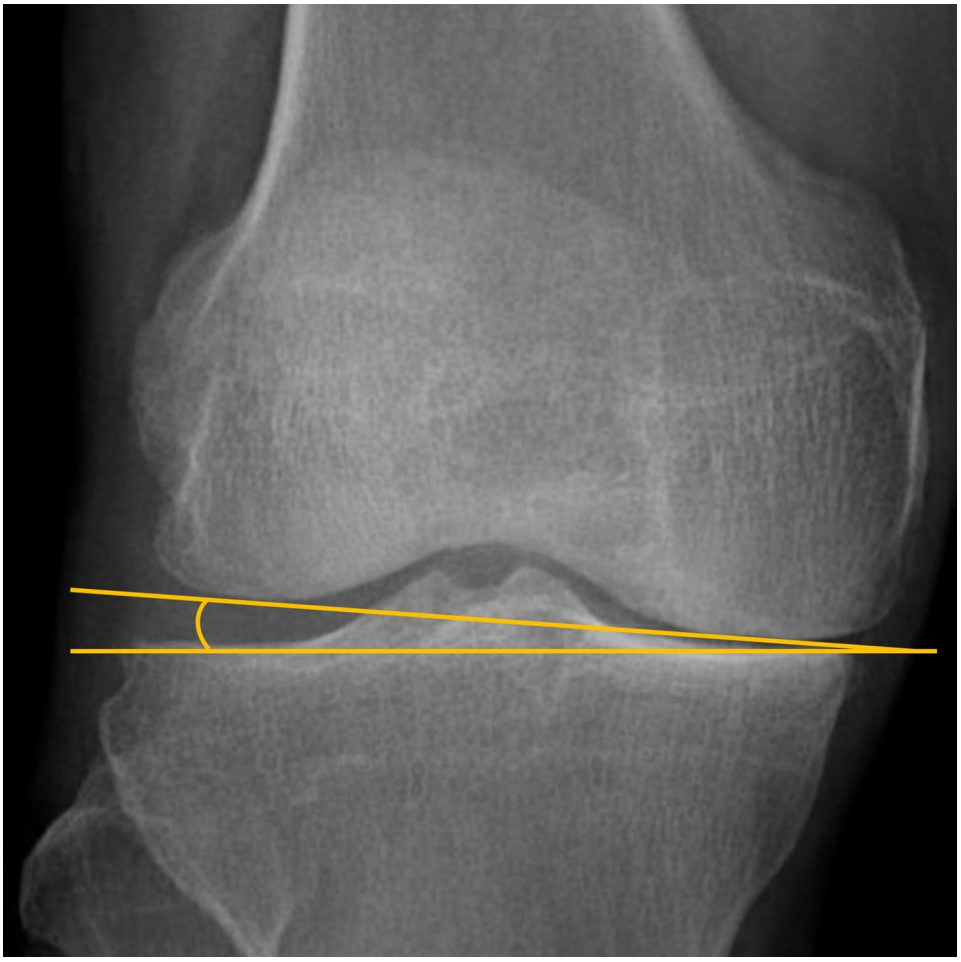

Fig. 2 Measurement method for the joint-line convergence angle on a standing anteroposterior knee radiograph

DCGAN, Real vs. StyleGAN3, Real vs. Diffusion + CycleGAN), a Bonferroni correction was applied to adjust the significance level, ensuring robust statistical validity. A p-value of < 0.05 (after correction) was considered statistically significant.

Intraclass correlation coefficient

To evaluate the consistency of the JLCA measurements, we assessed intra- and inter-observer agreements using the Intraclass Correlation Coefficient (ICC). Intra-observer reliability was determined by two radiologists who assessed the image sets twice, with a four-week interval between sessions. Inter-observer reliability was evaluated by comparing measurements independently conducted by two raters on the same image sets. According to standard interpretation, ICC values > 0.75 indicate good reliability, whereas values > 0.90 indicate excellent agreement.

Histogram analysis

To assess the intensity distribution and visual consistency between the real and generated knee radiographs, we conducted a histogram analysis. We computed the grayscale histograms of the real and synthetic images across the three generative models. These histograms represent the frequency of the pixel intensities across the image. By comparing the histograms of the real and generated images, we evaluated how closely the synthetic images resembled the distribution of real radiographs. To quantify the similarity of pixel intensity distributions, we computed the average histograms for both the entire real ($N = 10{,}042$) and synthetic datasets. We then calculated the Pearson Correlation Coefficient between these two average distributions to assess their statistical similarity. A close resemblance between the histograms of the generated and real images shows a higher visual similarity to the real images.

Results

In this section, the evaluation parameters and experimental results of the images generated by different generative models were discussed. The images generated by each model are presented in Fig. 3.

Visual turing test

The performance of the Visual Turing Test for each generative model is shown as a confusion matrix in Fig. 4. This matrix classifies images as fake or real, based on evaluations by expert surgeons and radiologists. The key components of the confusion matrix include True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Classification accuracy is calculated as the ratio of correct predictions to the total number of predictions. The classification accuracies for StyleGAN3 between the two surgeons were 36% and 41%, respectively. The radiologist showed 54% accuracy, while the DCGAN showed accuracies of 54%, 65%, and 88%, respectively. The integration of Stable Diffusion and CycleGAN yielded accuracies of 51%, 48%, and 63%, respectively. Notably, these evaluations showed that the two expert surgeons had difficulty consistently distinguishing between fake and real images.

Fréchet inception distance

Table 1 shows the number of real and fake images and the FID values for each model. StyleGAN3 achieved the lowest FID score of 10.84, indicating superior visual quality. In contrast, DCGAN had the highest FID of 74.15, signifying the lowest performance among the tested models. The Stable Diffusion model, when evaluated alone, established a baseline FID score of 46.37. The two-stage pipeline, which incorporates a CycleGAN to refine the initial diffusion outputs, demonstrated a notable improvement by reducing the FID to 39.79.

Optimal input number

It is crucial to indicate the optimal number of input datasets for training the StyleGAN3 model. The FID metric was used to evaluate the model's performance across different input dataset sizes and levels of augmentation. The FID score for larger datasets 10,042 images achieved the best results at an augmentation level of 1000k, with a value of 10.59. This internal training value is consistent with the score of 10.84 reported in Table 1, where the sample size was strictly matched to $N = 10{,}042$ for fair comparison across all models. The FID values for smaller datasets showed moderate improvement as augmentation levels increased.

Joint line angle measurement

Table 2 shows the JLCA measurements for real versus synthetic ("fake") images, reporting mean ± SD, min–max ranges, and p-values from the t-test. DCGAN-generated images measured $3.36 \pm 2.19°$ (range 0.12–9.29°) versus $4.19 \pm 3.07°$ (0.01–13.00°) for real images ($p = 0.12$). Diffusion+CycleGAN showed $3.72 \pm 2.52°$ (0.18–9.78°) versus $3.91 \pm 2.59°$ (0.14–10.12°; $p = 1.00$), and StyleGAN3 produced $3.60 \pm 2.37°$ (0.03–10.81°) versus $4.27 \pm 3.01°$ (0.02–17.30°; $p = 0.25$). Consequently, no statistically significant differences were observed between the real and synthetic images across all three generative models ($p > 0.05$), indicating high anatomical fidelity.

Intraclass correlation coefficient

To compute the ICC for inter- and intra-observer agreements, 30 real and 30 fake images were randomly selected, resulting in a total of 60 images for each model. The inter-observer agreement was strong, with ICC

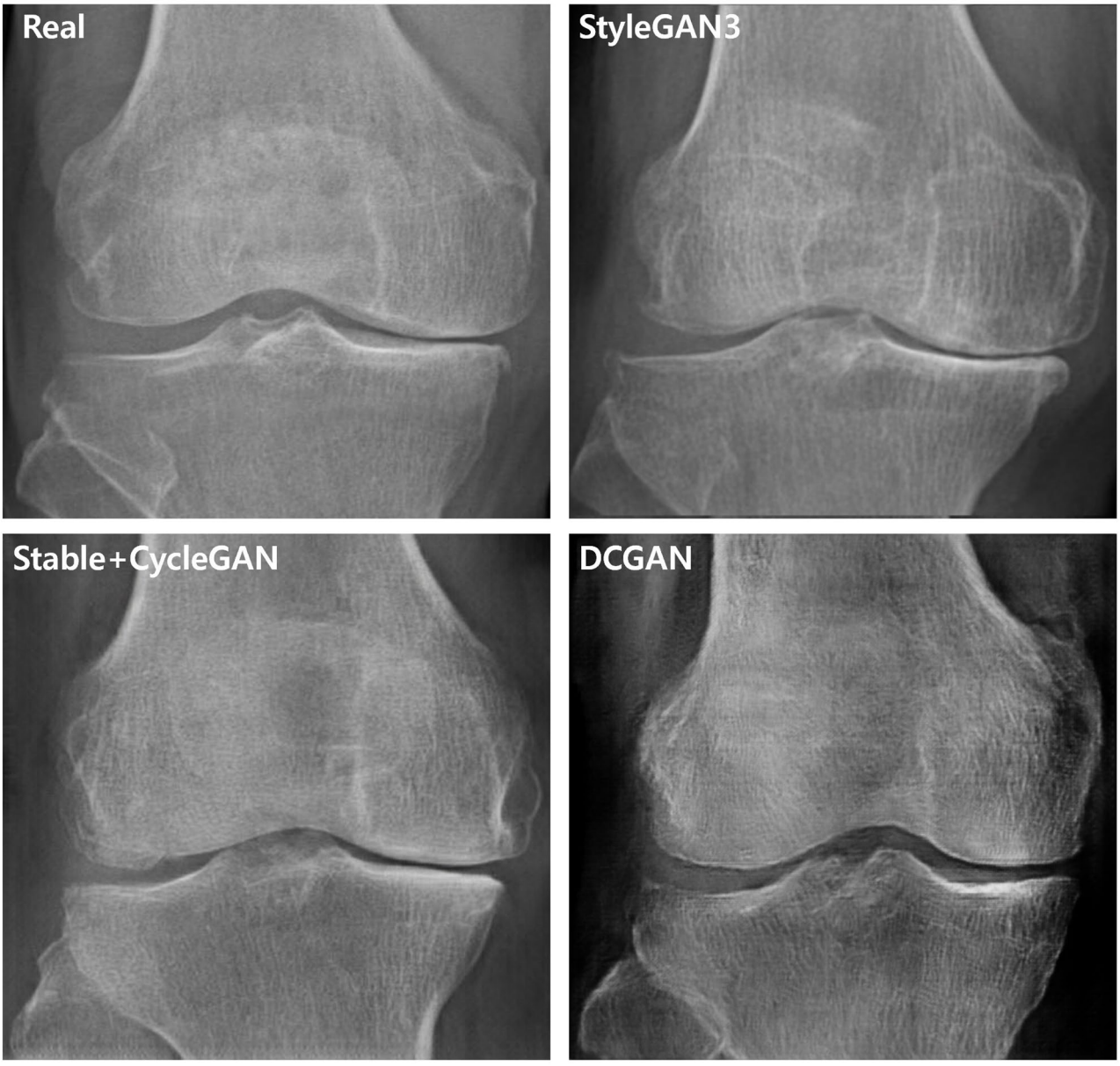

Fig. 3 Representative real and synthetic knee radiographs generated by DCGAN, StyleGAN3, and Diffusion + CycleGAN

values ranging between 0.83 and 0.97. For intra-observer reliability, both raters showed excellent repeatability. Rater 1 showed ICC values between 0.89 and 0.99 (confidence limits 0.77–0.99), while Rater 2 ranged from 0.94 to 0.99 (confidence limits 0.88-1.00).

Histogram analysis

The comparative histogram analysis (Fig. 5) revealed that the pixel intensity distributions of the synthetic images closely aligned with those of the real dataset ($N = 10{,}042$). Quantitative evaluation confirmed this high degree of similarity. StyleGAN3 demonstrated the highest fidelity with a Pearson correlation coefficient of 0.9984, followed closely by DCGAN (0.9976) and Diffusion + CycleGAN (0.9815). These near-perfect correlation values indicate that all three generative models successfully replicated the global intensity patterns of the original knee radiographs, ensuring visual and structural consistency.

Discussion

The most important finding of this study is that StyleGAN3 achieved the highest visual realism and deceived experts at rates as low as 36% in a Visual Turing Test, demonstrating its superiority in high-resolution synthetic knee radiograph generation. Additionally, StyleGAN3 produced anatomically consistent images, as shown by nonsignificant differences in JLCA measurements and strong ICC reliability. These results underscore its

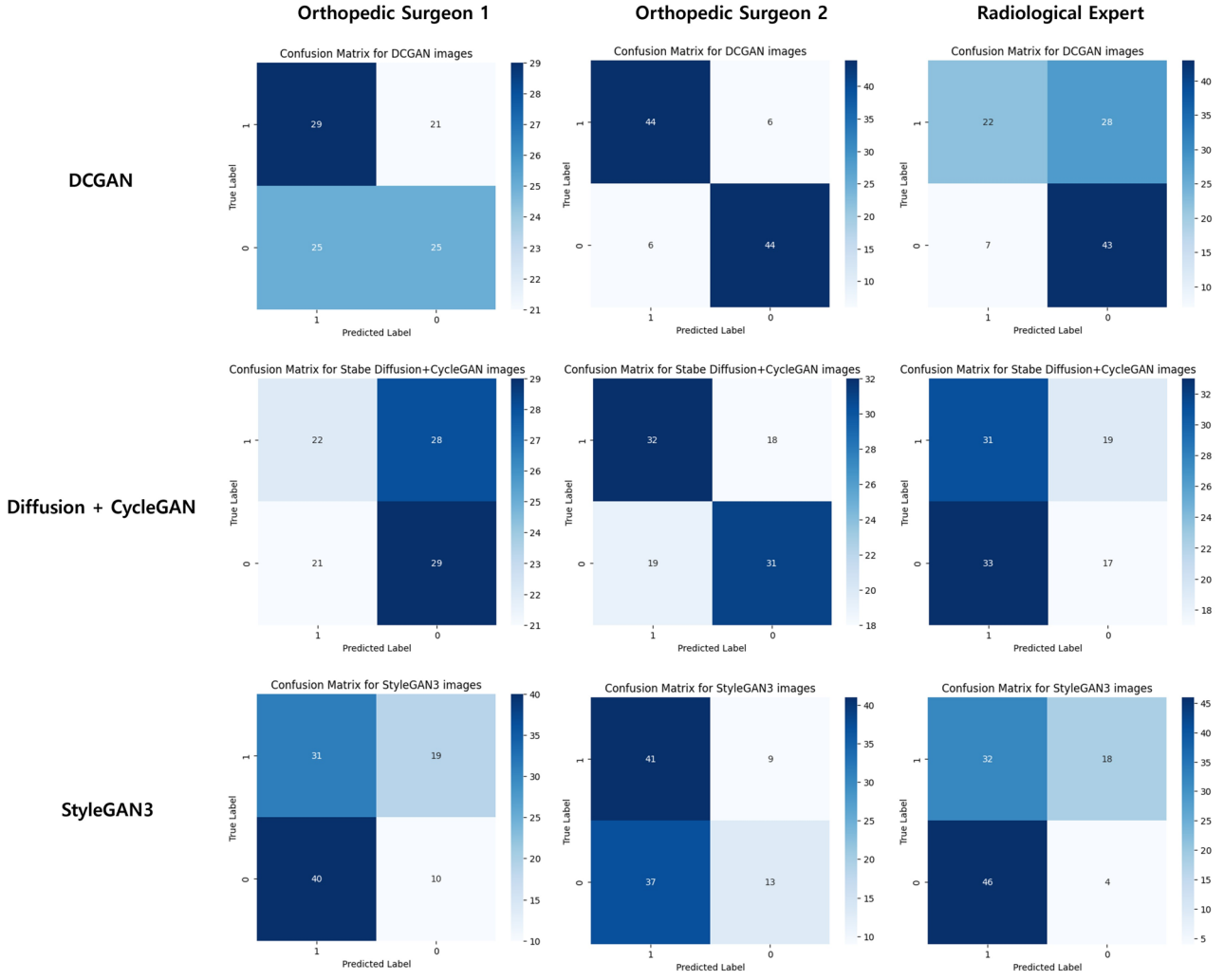

Fig. 4 Turing-test results: confusion matrices distinguishing real from synthetic images for each generative model

Table 1 Fréchet Inception Distance (FID) scores of each generative model

	Dataset size		FID
	Real	Fake	
DCGAN	10,042	10,042	74.15
Diffusion + CycleGAN	10,042	10,042	39.79
StyleGAN3	10,042	10,042	10.84

Table 2 Joint line convergence angle between models

	Mean ± std (max, min)		*p*-value
	Real	Fake	
DCGAN	4.19 ± 3.07 (0.01, 13.00)	3.36 ± 2.19 (0.12, 9.29)	0.12
Diffusion + CycleGAN	3.91 ± 2.59 (0.14, 10.12)	3.72 ± 2.52 (0.18, 9.78)	1.00
StyleGAN3	4.27 ± 3.01 (0.02, 17.30)	3.60 ± 2.37 (0.03, 10.81)	0.25

potential for clinical data augmentation and AI-aided diagnosis.

In this study, we introduce three advanced models designed to generate high-resolution synthetic knee radiographs. The primary goal of this study was to mitigate the scarcity of data in medical research by leveraging robust and reliable generative models for high-resolution knee radiographic images. We conducted a comprehensive analysis using quantitative metrics alongside expert-based visual assessment.

Recent advancements in medical image synthesis have utilized generative models to enhance data augmentation. Prezja et al. [21] introduced CycleGAN to simulate OA progression and regression, whereas another study [8] reported that DeepFake knee X-rays could effectively deceive medical experts and enhance classification performance under limited real data conditions. However, both studies lacked comprehensive anatomical validation and quantitative metrics, including FID. We used StyleGAN3 because of its ability to generate the

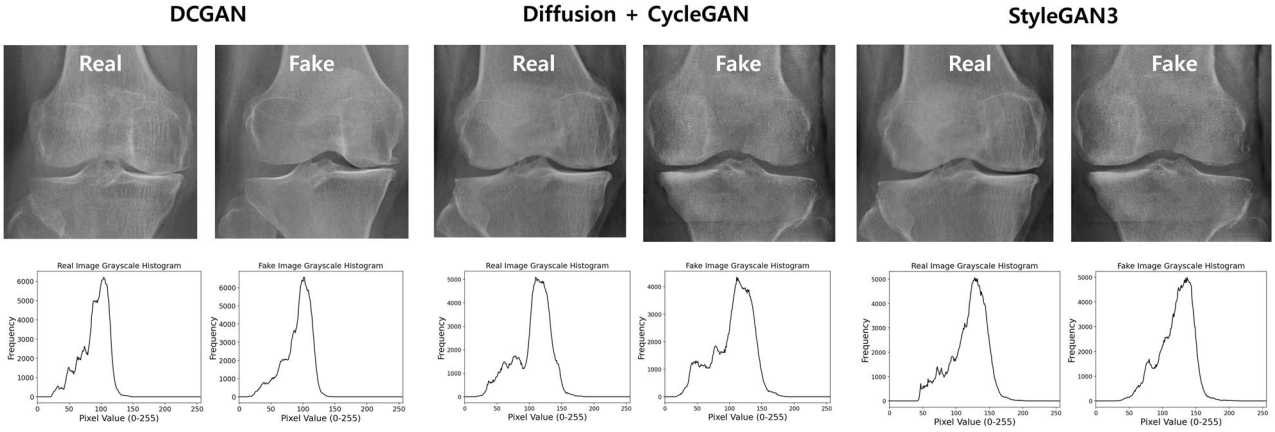

Fig. 5 Comparative grayscale-intensity histograms of real versus model-generated knee radiographs

highest-fidelity images from a limited amount of data. In our study, the generative models were trained on 10,042 images and achieved the best performance, showing the robustness and data efficiency of StyleGAN3. StyleGAN3 achieved the lowest FID value of 10.84, demonstrating an excellent visual similarity to real images. In comparison, Littlefield et al. reported higher FID scores of 26.33 for KL grading and 22.54 for bilateral views using a few-shot augmentation pipeline, indicating lower image fidelity than that of our study, but still acceptable for medical use [7]. Moreover, our integrated two-stage Stable Diffusion and CycleGAN pipeline achieved an FID score of 39.79, indicating acceptable visual and anatomical fidelity. This suggests that integrating CycleGAN improves the performance of the diffusion models by addressing some of their limitations. Owing to the architectural limitations of DCGAN, it demonstrated the highest FID value of 74.15, indicating noticeably lower image quality; nonetheless, it still showed the capacity to generate structurally coherent outputs, especially when trained on a sufficient number of images. Additionally, this trend closely resembles that of previous research [22]. In an earlier study, the DCGAN achieved an FID score of 117, which was higher than those of other models [22]. The findings of this study showed that StyleGAN3 is the most effective model for generating high-fidelity synthetic images, whereas diffusion-based pipelines can be significantly improved through adversarial enhancement.

From a quality-analysis perspective, StyleGAN3 showed the highest performance by generating realistic synthetic images. The Visual Turing Test showed that experienced surgeons and radiologists faced challenges in indicating whether real and generated images were present, highlighting the model's effectiveness and achieving the lowest classification accuracies (36–54%). Conversely, the performance of the DCGAN was comparatively lower in terms of visual fidelity and realism. The Stable Diffusion + CycleGAN pipeline shows moderate image quality, indicating that adversarial refinement can address the limitations of diffusion models. These findings align with those of other studies that found that images generated by the Deepfake model misled over 70% of experts. This highlights the deceptive realism of GAN-generated images and the challenges they pose for expert visual assessments [8].

Additionally, JLCA measurements between the real and generated images were compared across the three models to evaluate the anatomical fidelity of the synthesized images. The analysis revealed no statistically significant differences in JLCA between the real and synthetic groups for any of the generative models. Specifically, the DCGAN model showed a p-value of 0.12, while the Diffusion + CycleGAN and StyleGAN3 models yielded p-values of 1.00 and 0.25, respectively. These results demonstrate that all evaluated frameworks successfully preserved critical anatomical structures, showing high anatomical consistency between real and generated radiographs. Ultimately, these findings indicate that the investigated models are highly effective for medical image generation and hold significant potential for data augmentation and training in clinical research.

The ability to generate high-fidelity synthetic knee radiographs, as shown in this study, has great potential for advancing diagnosis, treatment, and research on OA. Given the critical role of the KL grading system in clinical decision making, the proposed generative model provides an effective environment for training and validating KL-based AI diagnostic models, consequently supporting appropriate treatment selection. As OA is a progressive disease, the model may also be leveraged to simulate the progression of disease, create longitudinal datasets, and support prognostic predictions. Furthermore, the high accuracy observed in quantitative anatomical parameters, including JLCA, shows that these synthetic images can be reliably used for alignment analysis and simulation-based surgical planning. Furthermore, the model

offers valuable educational opportunities by allowing trainees to study OA's key radiographic features, such as joint space narrowing, osteophytes, and sclerosis, even before encountering real clinical cases. Ultimately, this approach may contribute to advancing personalized precision medicine in orthopedic care.

However, this study has some limitations. Previous investigations of synthetic knee radiographs for OA have predominantly relied on homogeneous anteroposterior (AP) X-ray images. In contrast, our generative models were trained on heterogeneous full-leg panoramic radiographs acquired using two different imaging systems: a conventional X-ray unit and a tomosynthesis-enabled X-ray system. The tomosynthesis system reconstructs a continuous panoramic image by sequentially capturing narrow (~ 5 cm wide) slit projections using a flat-panel detector that moves parallel to the X-ray tube. Tomosynthesis mitigates stitching-related distortions and measurement errors; however, it has not been widely adopted in clinical practice. By combining images from both modalities, we increased the heterogeneity in resolution, contrast, and noise characteristics. Consequently, it is highly likely that we observed a higher FID value compared with earlier StyleGAN2-ADA studies that used only homogeneous AP images and reported lower FID values [22]. The need for a model to reconcile the two distinct acquisition modalities further amplifies the distributional complexity because each modality introduces unique imaging artifacts. Both image types were deliberately included to preserve clinical relevance and generalizability instead of limiting our analysis to tomosynthesis-only data. Images from patients with advanced-stage OA showed pronounced osteophyte formation, joint space narrowing, and subchondral sclerosis, rendering them visually more complex than mixed-grade datasets. Even advanced models may struggle to fully capture these intricate anatomical details without a larger sample size, extended training epochs, or more aggressive regularization and data augmentation strategies.

Another limitation is that our evaluation focused on visual realism and anatomical plausibility (e.g., JLCA) rather than direct utility in a downstream clinical task (e.g., KL grading or OA severity prediction). We did not perform downstream-task evaluation because KL annotations were not available for the entire cohort, and assigning KL grades to more than 10,000 radiographs would require substantial expert effort and adjudication beyond the scope of this work. Moreover, because our cohort comprised arthroplasty candidates, KL grades were likely skewed toward advanced disease (predominantly grades 3–4), which would limit the interpretability of a KL-based experiment and reduce generalizability to earlier OA stages. Future downstream-task studies that vary the proportion of synthetic data are needed to quantify practical benefit and to rule out amplification of label noise.

Despite these limitations, StyleGAN3 showed outstanding performance in generating high-fidelity knee radiographs. This was confirmed through a histogram analysis of pixel intensities, JLCA measurements, and a Visual Turing Test in which experts found it difficult to distinguish real from synthetic images.

Conclusion

The high-resolution knee radiograph synthesis technique developed in this study has the potential to address data scarcity in clinical settings and improve the accuracy of AI-based diagnostic and treatment planning support systems. Specifically, using synthetic images for training deep learning model complements real patient data and enhances the generalizability of predictive models. In future studies, we plan to validate the model's generalizability and robustness using a large-scale multicenter dataset that included diverse imaging protocols.

Abbreviations

AP	Anteroposterior
AI	Artificial intelligence
CycleGAN	Cycle-consistent Generative Adversarial Network
DCGAN	Deep Convolutional Generative Adversarial Network
FN	False Negatives
FP	False Positives
FID	Fréchet Inception Distance
GANs	Generative Adversarial Networks
ICC	Intraclass correlation coefficients
JLCA	Joint Line Convergence Angle
KL	Kellgren–Lawrence
OA	Osteoarthritis
StyleGAN3	Style Generative Adversarial Network 3
TKA	Total knee arthroplasty
TN	True Negatives
TP	True Positives

Acknowledgements

Not applicable.

Author contributions

Kwangho Chung: Writing. Arailym Dosset: Methodology. Ji-Hoon Nam: Formal analysis. Yong-Gon Koh: Data curation. Jae Min Kim: Data curation. Paul Shinil Kim: Editing. Jin Woo Lee: Review and edit. Kyoung-Mi Park: Editing. Hyuck Min Kwon: Supervision. Kyoung-Tak Kang: Supervision. H.M.K. and K.-T.K. supervised the study and served as corresponding authors. All authors reviewed and approved the final manuscript.

Funding

Not applicable.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Ethics approval for this study was obtained from the Institutional Review Board of Yonsei Sarang Hospital (YSSR IRB No. 2025-04-001). This study was conducted in accordance with the Declaration of Helsinki. The requirement for informed consent was waived because the study used fully anonymized

retrospective medical records in accordance with institutional and regulatory guidelines.

Consent for publication
This manuscript does not contain any individual person's data in any form (including images or videos) that would require consent for publication.

Competing interests
KTK is the Chief Executive Officer of Skyve and Clevion Co. Ltd., but this did not influence the results of our research. JHN is an employee of Skyve and Clevion Co. Ltd., but this also did not affect the results of our research. YGK is a shareholder of Skyve. The remaining authors declare no competing interests.

Author details
[1]Department of Orthopedic Surgery, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin 16995, Republic of Korea
[2]Clevion Co., Ltd., Seoul 07217, Republic of Korea
[3]Skyve R&D LAB, Skyve Co., Ltd., Seoul 07217, Republic of Korea
[4]Department of Mechanical Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
[5]Joint Reconstruction Center, Department of Orthopaedic Surgery, Yonsei Sarang Hospital, Seoul 06702, Republic of Korea
[6]Joint Reconstruction Center, Department of Radiology, Yonsei Sarang Hospital, Seoul 06702, Republic of Korea
[7]Department of Orthopaedic Surgery, The Bone Hospital, 67, Dongjak-daero, Dongjak-gu, Seoul 07014, Republic of Korea
[8]Department of Orthopaedic Surgery, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

Received: 18 November 2025 / Accepted: 18 February 2026
Published online: 03 March 2026

References

1. Courties A, Kouki I, Soliman N, Mathieu S, Sellam J. Osteoarthritis year in review 2024: Epidemiology and therapy. Osteoarthritis Cartilage. 2024;32(11):1397–404.
2. Global B, Lyons R, Gabbe B, Kemp A. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386(9995):743.
3. Muraki S, Oka H, Akune T, Mabuchi A, En-Yo Y, Yoshida M, Saika A, Suzuki T, Yoshida H, Ishibashi H. Prevalence of radiographic knee osteoarthritis and its association with knee pain in the elderly of Japanese population-based cohorts: the ROAD study. Osteoarthritis Cartilage. 2009;17(9):1137–43.
4. Kinds M, Welsing P, Vignon E, Bijlsma J, Viergever M, Marijnissen A, Lafeber F. A systematic review of the association between radiographic and clinical osteoarthritis of hip and knee. Osteoarthritis Cartilage. 2011;19(7):768–78.
5. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine-Promise, Progress, and Challenges. JAMA Intern Med. 2019;179(3):293–4.
6. Almajalid R, Zhang M, Shan J. Fully automatic knee bone detection and segmentation on three-dimensional MRI. Diagnostics (Basel). 2022;12(1).
7. Littlefield N, Amirian S, Biehl J, Andrews EG, Kann M, Myers N, Reid L, Yates AJ Jr, McGrory BJ, Parmanto B. Generative AI in orthopedics: an explainable deep few-shot image augmentation pipeline for plain knee radiographs and Kellgren-Lawrence grading. J Am Med Inform Assoc. 2024;31(11):2668–78.
8. Prezja F, Paloneva J, Pölönen I, Niinimäki E, Äyrämö S. DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. Sci Rep. 2022;12(1):18573.
9. McNulty JR, Kho L, Case AL, Slater D, Abzug JM, Russell SA. Synthetic medical imaging generation with generative adversarial networks for plain radiographs. Appl Sci. 2024;14(15):6831.
10. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 2018;321:321–31.
11. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst. 2020;33:6840–51.
12. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022;2022:10684–10695.
13. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. 2017;2017:2223–2232.
14. Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. IEEE Trans Med Imaging. 2017;36(12):2536–45.
15. Hong S, Marinescu R, Dalca AV, Bonkhoff AK, Bretzner M, Rost NS, Golland P. 3D-StyleGAN: a style-based generative adversarial network for generative modeling of three-dimensional medical images. In: Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM-4MICCAI 2021, and First Workshop, DALI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1: 2021: Springer; 2021. p. 24–34.
16. Che Azemin MZ, Mohd Tamrin MI, Hilmi MR, Mohd Kamal K. Assessing the efficacy of StyleGAN 3 in generating realistic medical images with limited data availability. In: Proceedings of the 2024 13th International Conference on Software and Computer Applications. 2024;2024:192-197.
17. Das S, Walia PJ. Enhancing early diabetic retinopathy detection through synthetic DR1 image generation: a StyleGAN3 approach. 2025.
18. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, Aila T. Alias-free generative adversarial networks. Adv Neural Inf Process Syst. 2021;34:852–63.
19. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, Merhof DJ. Diffusion models in medical imaging: a comprehensive survey. Med Image Anal. 2023;88:102846.
20. Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA, Išgum I. Deep MR to CT synthesis using unpaired data. In: Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, held in conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2: 2017: Springer; 2017. p. 14–23.
21. Prezja F, Annala L, Kiiskinen S, Lahtinen S, Ojala T, Nieminen PJ. Generating synthetic past and future states of knee osteoarthritis radiographs using cycle-consistent generative adversarial neural networks. Comput Biol Med. 2025;187:109785.
22. Ahn G, Choi BS, Ko S, Jo C, Han HS, Lee MC, Ro DH. High-resolution knee plain radiography image synthesis using style generative adversarial network adaptive discriminator augmentation. J Orthop Res. 2023;41(1):84–93.

Publisher's note