# Synthesizing Realistic Knee MRI Images: A VAE-GAN Approach for Enhanced Medical Data Augmentation

Revathi S A[1], B Sathish Babu[2]
Dept. of Computer Science, RV College of Engineering, Bangalore, India[1]
Dept. of AI&ML, RV College of Engineering, Bangalore, India[2]

*Abstract*—**This study presents a novel approach for synthesizing knee MRI images by combining Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). By leveraging the strengths of VAEs for efficient latent space representation and GANs for their advanced image generation capabilities, we introduce a VAE-GAN hybrid model tailored specifically for medical imaging applications. This technique not only improves the realism of synthesized knee MRI images but also enriches training datasets, ultimately improving the outcome of machine learning models. We demonstrate significant improvements in synthetic image quality through a carefully designed architecture, which includes custom loss functions that strike a balance between reconstruction accuracy and generative quality. These improvements are validated using quantitative metrics, achieving a Mean Squared Error (MSE) of 0.0914 and a Fréchet Inception Distance (FID) of 1.4873. This work lays the groundwork for novel research guidelines in biomedical image study, providing a scalable solution to overcome dataset limitations while maintaining privacy standards, and pavement of reliable diagnostic tools.**

***Keywords—Custom loss function; decoder; discriminator; GAN; latent space; VAE***

## I. Introduction

Deep learning has renovated medical imaging, improving diagnostic accuracy, treatment preparation, and patient monitoring. Despite its potential, the field faces significant challenges, particularly in musculoskeletal imaging, where limited access to diverse, high-quality datasets impedes progress. This is exacerbated by privacy concerns and the cost-intensive nature of data collection and annotation. To address these barriers, this study proposes a novel VAE-GAN framework for synthesizing realistic and diverse knee MRI images. By integrating the latent space representation capabilities of Variational Autoencoders (VAEs) with the generative strength of Generative Adversarial Networks (GANs), the proposed method aims to overcome the shortcomings of existing augmentation techniques while maintaining clinical relevance.

The VAE-GAN framework balances the diversity of VAEs with the sharpness and realism provided by GANs. Specifically, the VAE's latent space ensures continuity, making it suitable for generating diverse samples, while the GAN component enhances image fidelity, addressing the critical need for high-quality training datasets in medical imaging. This synergy enables the generation of realistic knee MRI images that reflect the variations needed for machine learning model training.

However, existing generative methods present restrictions that make them less suitable for this task:

*1)* Standalone GANs are prone to mode collapse, which restricts the variety of generated images and reduces their applicability in representing diverse medical conditions.

*2)* VAEs often produce blurry outputs due to their reliance on reconstruction loss functions, which prioritize structural accuracy over fine details.

*3)* Other generative models, such as Diffusion Models, require significantly more computational resources, making them unrealistic for medical imaging applications with restricted data.

By addressing these limitations, the proposed VAE-GAN framework emerges as an effective solution to the problem of data scarcity in knee MRI imaging, paving the way for better augmentation techniques that enhance the performance of diagnostic tools.

The limited availability of high-quality, diverse, and privacy-compliant medical imaging datasets significantly hampers the progress of robust and generalizable deep learning techniques for diagnostic applications. Existing methods for data augmentation and synthesis often fail to achieve the required balance of realism, diversity, and fidelity, limiting their effectiveness in medical imaging contexts.

The research explores how a hybrid generative model can effectively synthesize realistic and diverse knee MRI images to address the problem of data scarcity in medical imaging. It investigates the optimal architectural and loss function designs needed to stabilize image quality with reconstruction accuracy in the synthesized data. Furthermore, the study examines how the proposed VAE-GAN framework compares to existing generative methods in terms of quality metrics and clinical applicability.

The objectives of this research are to implement a VAE-GAN-based framework capable of producing realistic and diverse knee MRI images and to design a custom loss function that integrates reconstruction loss, KL divergence, and adversarial loss to achieve high-fidelity image generation. Additionally, the study aims to validate the model's performance using quantitative metrics such as Mean Squared Error (MSE) and Fréchet Inception Distance (FID), as well as through qualitative assessments of the synthetic images' visual quality.

Magnetic Resonance Imaging (MRI), a critical tool for assessing musculoskeletal disorders, especially knee-related conditions, is a domain where deep learning can substantially enhance diagnostic accuracy and patient outcomes. However, the accomplishment of deep learning techniques in biomedical imaging is heavily dependent on the accessibility and image resolution of the datasets used. High-quality medical image data is often scarce and challenging to obtain due to patient privacy concerns. Furthermore, the processes of data collection and annotation are both costly and time-consuming, frequently resulting in datasets that lack the necessary diversity and volume to train robust and generalizable machine learning models. In knee MRI data, specific challenges include variations in pathology presentation and imaging protocols, complicating the training process further [1].

To address these challenges, the proposed VAE-GAN (Variational Autoencoder-Generative Adversarial Network) framework combines the robust data encoding capabilities of VAEs with the powerful image generation capabilities of GANs. VAEs excel at compressing data into a latent space, enabling the generation of new data instances, but sometimes produce outputs that lack the sharpness and detail characteristic of high-quality MRI scans [2]. Conversely, GANs generate sharp, high-definition images through adversarial training but face challenges such as training instability and mode collapse—a condition where the variety of produced images is insufficient [3]. By integrating these models, the VAE-GAN framework achieves a balance between realistic detail and diversity in synthetic knee MRI images.

The introduction should also include a section listing the reasons for choosing the proposed VAE-GAN method, detailing why it is particularly appropriate for addressing the challenges outlined in the study. Additionally, it would be advantageous to point out which limitations of existing methods make them less suitable for this problem. Including such a rationale would provide greater clarity on the motivations behind the selected approach.

This paper explores the architecture and functionality of the VAE-GAN model, focusing on its application for synthesizing knee MRI images. A detailed examination of the encoder, decoder/generator, and discriminator is provided, along with the custom loss function that optimizes image reconstruction accuracy and resolution. Experimental results validate the model's effectiveness in creating authentic and diverse synthetic knee MRI images, demonstrating its potential to enrich medical imaging datasets and enhance the performance of machine learning models used in medical diagnostics.

Additionally, this study carefully addresses the ethical concerns associated with deploying generative methods for data augmentation in healthcare. The input images used for model training were thoroughly anonymized, and all procedures adhered strictly to ethical standards, ensuring patient confidentiality at every stage.

### A. Paper Structure

The remainder of the paper is organized as follows: Section II reviews related work in biomedical imaging and generative modeling. Section III presents the methodology, including the proposed VAE-GAN architecture and training procedures. Section IV discusses the experimental setup, results, and evaluation. Section V highlights limitations, ethical considerations, and future work. Finally, Section VI concludes the study and summarizes its contributions to the field.

## II. Literature Review

Bio Medical imaging plays a crucial part in current healthcare by facilitating the diagnosis and management of various health conditions. However, challenges such as limited data availability and concerns over patient privacy hinder the development of robust machine-learning techniques for image analysis. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have emerged as powerful generative models, offering solutions to these challenges. This section explores the implementation of VAE-GAN models in medical imaging, highlighting their methods, effectiveness, and potential drawbacks.

VAEs and GANs are distinct but complementary approaches in generative modelling. VAEs consist of an encoder-decoder structure, where the encoder compresses medical images into a latent space, retaining essential features. The decoder reconstructs the image from this compressed representation, with a tailored loss function ensuring that the output closely matches the original and that the latent space accurately reflects key medical attributes [13]. Conversely, GANs use a generator-discriminator framework. The generator generates images that captures those in the training set, whereas the discriminator evaluates their authenticity. This adversarial interaction drives the generator to create realistic images [3].

The synergy of VAE-GAN techniques lies in their capability to combine VAE's efficient latent space encoding with GAN's image refinement capabilities. The VAE encodes medical images into a meaningful latent space, and the GAN uses this encoding to generate realistic images, preserving essential characteristics while enhancing overall quality. VAE-GAN models have shown remarkable versatility, particularly in data augmentation. They can produce synthetic images that are almost indistinguishable from real ones, expanding the range of training datasets and improving machine learning models' performance in tasks namely image segmentation and classification. These models also excel in denoising, removing extraneous noise while preserving critical structures, making them perfect for applications like tumor detection and disease progression monitoring [3]. Additionally, VAE-GANs can generate disease-specific images, offering valuable resources for medical training and refining diagnostic systems.

Despite their potential, VAE-GAN models have some drawbacks. Their complex architecture requires substantial computational resources and careful hyperparameter tuning for optimal performance. One common issue is mode collapse, where the variety of generated images diminishes, limiting the model's applicability across diverse medical scenarios [3]. The latent space's interpretability in VAEs remains a challenge, potentially obscuring insights into how model decisions are made [4]. Additionally, biases in training datasets can propagate through VAE-GAN models, leading to biased outputs and skewed diagnostic results [5]. While VAE-GAN models represent significant developments in medical image processing, addressing these limitations is critical for their ethical and practical application in healthcare settings.

Recent advancements in generative models, particularly GANs and Diffusion Models (DMs), have expanded the scope of medical image synthesis. GANs have demonstrated proficiency in creating realistic images, but they face tests such as training unpredictability and mode breakdown [6, 7]. Conversely, VAEs produce a larger variety of outputs and are fewer prone to mode collapse, though the images they generate sometimes lack sharpness due to smoothing tendencies in their loss functions [8]. Diffusion Models (DMs) offer a promising solution by producing original and diverse outputs, though their adoption in clinical settings is hindered by high computational demands and extended processing times [9, 10]. This review explores the evolving landscape of these technologies, aiming to balance quality, speed, and diversity in medical image synthesis [11].

GANs have become essential in advancing realistic image synthesis, particularly for enhancing medical imaging datasets. While GANs are adept at producing images that thoroughly resemble actual medical data, they can still suffer from mode collapse, leading to a limited variety of generated images. To address this, more advanced versions of GANs, such as Wasserstein GAN (WGAN) [12] and Conditional GAN (CGAN) [13], have been developed, significantly improving image quality and offering greater control over image generation. Deep generative models are revolutionizing biomedical imaging by improving diagnostic and treatment processes. VAEs are praised for their straightforward training and ability to capture complex data distributions. This ability makes them irreplaceable in medical imaging, where a rich representation of data is crucial. However, the occasional blurriness of VAE-generated images has led to the advancement of hybrid models like VAE-GANs, which combine the encoding power of VAEs with the image refinement abilities of GANs, resulting in clearer and more diverse outputs [18].

Diffusion Models (DMs) represent a breakthrough in generative modelling, excelling in capturing detailed, high-dimensional data distributions typically found in medical images. Their ability to replicate intricate features suggests they may outperform VAEs and GANs in relationships with realism and quality [19]. However, the high computational demands and lengthy processing times required by DMs limit their practicality, especially in real-time applications. Innovations such as Progressive Distillation [20] and the Fast Diffusion Probabilistic Model (FastDPM) [21] aim to optimize the sampling process while maintaining high-quality outputs. The Denoising Diffusion Implicit Model (DDIM) [22] balances speed with output fidelity, underscoring the importance of refining generative models.

This review categorizes deep generative models into three main types: GANs, VAEs, and DMs. GANs have made a significant impact on medical image augmentation, with iterations such as Wasserstein GAN (WGAN) [14] and Deep Convolutional GAN (DCGAN) [16] proving essential in generating realistic 2D MRI sequences. Studies by Han et al. [23] validate WGAN's superior performance in producing images that are indistinguishable from original medical scans. Likewise, Progressive Growing of GANs (PGGAN) [17], when merged with traditional data augmentation methods, has directed to improved classification algorithm performance. Conditional GANs [15] have been especially effective in generating targeted, realistic synthetic images, as demonstrated by Frid-Adar et al. [24] and Guibas et al. [25].

VAEs also show a critical role in diversifying medical imaging datasets. Research by Zhuang et al. [26] highlights Conditional VAEs (CVAEs) and Conditional WGANs' capability to produce high-quality brain images, enhancing classification model accuracy. The outline of Independently Conditional VAE (ICVAE) by Pesteie et al. [28] marks a noteworthy development in VAE technology, improving classification and segmentation tasks through increased diversity. Recent studies recognize the probability of Diffusion Models in generating high-resolution 3D MRI images. Research by Pinaya et al. [27] emphasizes DMs' superiority in producing authentic MRI samples. Moreover, the brain SPADE model introduced by Fernandez et al. [29] integrates DMs with VAE-GANs to create labeled MRI images, crucial for training segmentation models. Furthermore, explorations by Lyu and Wang [30] into DMs' image translation capabilities show that they can convert MRI to CT scans with greater accuracy than outdated methods, demonstrating the significant impact of these models on medical imaging.

In conclusion, deep generative methods, namely GANs, VAEs, and DMs, are transforming medical imaging. These methods are pivotal in tackling issues like data scarcity and augmenting datasets for improved diagnostic accuracy. Their ability to synthesize high-quality images, translate between imaging modalities, and enhance training datasets underscores their potential to advance diagnostic practices and medical research. Future developments promise to further expand their efficacy, making them invaluable tools in healthcare.

## III. Methodology

The methodology deployed for constructing and training a VAE-GAN model for synthesizing knee MRI images is composed of several intricate stages, from initial data preprocessing to the sophisticated dynamics of model training. This model capitalizes on the compressive data encoding capabilities of Variational Autoencoders (VAEs) and the image generation prowess of Generative Adversarial Networks (GANs). Herein, we elucidate the encoder's latent space utilization for image synthesis, the beta hyperparameter's role in the VAE loss function, the chosen architecture's pertinence to knee MRI imaging, the enhanced model diagrams for augmented clarity, and the tailoring of the custom loss function to medical imaging.

The VAE-GAN framework was selected because it combines the strengths of two advanced generative models. VAEs excel at compressing data into a continuous latent space, ensuring diversity in synthesized outputs, while GANs are well-known for generating sharp and realistic images through adversarial training. By integrating these models, the VAE-GAN framework addresses limitations of standalone methods:

*1)* VAEs sometimes produce blurry outputs due to their reconstruction-focused loss functions.

*2)* GANs are prone to training instability and mode collapse, reducing output diversity.

This hybrid approach effectively balances the sharpness and realism of GAN-generated images with the structural fidelity and diversity provided by VAEs, making it particularly suitable for the complex requirements of knee MRI image synthesis. Fig. 1 shows model architecture flowchart.

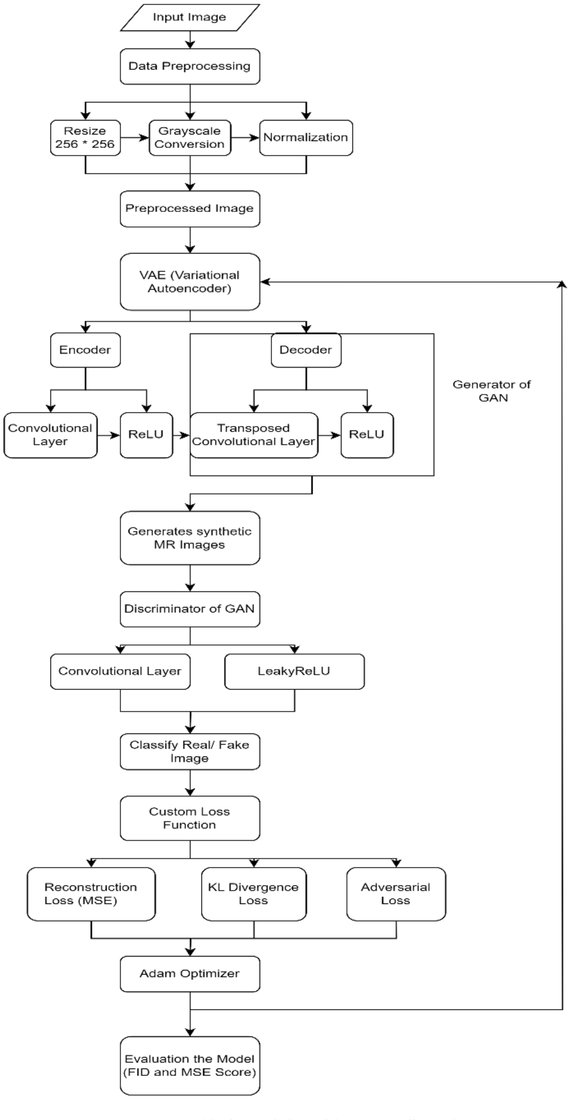


Fig. 1. Detailed model architecture flowchart.

## A. Data Collection and Sample Selection

The dataset used in this training was sourced from the Osteoarthritis Initiative (OAI), an openly accessible repository of knee MRI data. The images were selected based on the following criteria:

*1)* Availability of high-resolution knee MRIs.

*2)* Representation of four severity levels of osteoarthritis: normal, mild, moderate, and severe.

*3)* Ensuring diversity across patient demographics, imaging protocols, and pathology presentations.

From the OAI dataset, 150 knee MRI images were selected, including: 10 normal images, 50 mild images, 65 moderate images, 34 severe images.

Efforts were made to ensure that the dataset represented a wide spectrum of knee osteoarthritis conditions, thus enhancing the generalizability of the model.

## B. Data Preprocessing

*1) Initial data input:* The dataset consists of 150 knee MRI images, each resized to 256x256 pixels. These images correspond to four distinct severity levels of knee osteoarthritis (OA): normal, mild, moderate, and severe. This provides an adequate range for capturing pathological variations across OA stages.

*a) Grayscale conversion:* Knee MRIs were initially in color, but for this task, the images were converted into grayscale to simplify the data. Grayscale imaging emphasizes intensity values, which are critical in medical imaging to capture structural details. This conversion reduces the number of channels from 3 to 1, making computations more efficient without losing the essential information.

*b) Normalization:* Min-max normalization was applied to scale pixel values to a [0, 1] range. This normalization aids in stabilizing numerical computation during model training by ensuring consistent data scales, which speeds up convergence and reduces training instability.

*c) Resizing images:* Each image was resized to a fixed 256x256 resolution using bicubic interpolation. This method preserves image quality while ensuring uniformity in input data dimensions, which is necessary for training convolutional neural networks (CNNs) and other deep models (see Fig. 2).

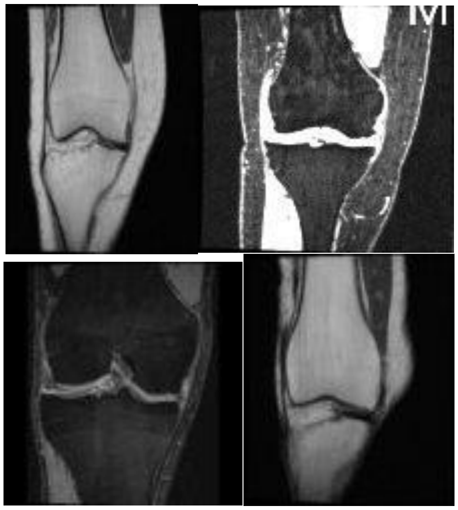


Fig. 2. Input MRI images of knee OA.

## C. Encoder and Decoder Model

After these steps, the data is cleaned, standardized, and ready for input into the VAE-GAN model, providing an optimal starting point for further processing and generation.

*1) Variational Autoencoder (VAE):* The encoder compresses the input MRI images into a latent space representation. It comprises of a sequence of convolutional layers followed by ReLU activations. Key architectural choices include:

*a) Convolutional layers:* Four convolutional layers are used, with increasing filter sizes (32, 64, 128, and 256) at each layer, designed to progressively capture image features from basic edges to complex patterns.

*b) Filter size:* The kernel sizes used were 3x3 for the initial layers, transitioning to 2x2 in deeper layers to down sample spatial dimensions efficiently.

*c) ReLU activation:* Applied after each convolutional layer to introduce non-linearity, enabling the model to learn more complex representations.

*d) Batch normalization:* Introduced after each convolutional block to stabilize and accelerate training, preventing vanishing/exploding gradient problems.

*e) Latent space representation:* The final dense layer of the encoder outputs two variables, mean (μ) and variance ($\sigma^2$), representing the latent distribution. The dimensionality of the latent space was set to 128 dimensions, balancing between representational capacity and computational efficiency.

*2) VAE Decoder / GAN Generator:* The decoder (also serving as the GAN generator) reconstructs images from the latent space representation:

*a) Transposed convolutional layers:* Four layers of transposed convolutions are applied to up sample the image back to its original size (256x256). This "unpooling" mechanism helps restore the resolution lost during encoding.

*b) Activation functions:* ReLU activation is applied throughout, except in the final layer, where a sigmoid activation is used to ensure pixel values remain between [0,1].

*c) Batch normalization:* Continues to be applied in the decoder to avoid overfitting and stabilize the gradient flow.

*3) Generative Adversarial Network (GAN) discriminator:* The discriminator serves to distinguish between actual and synthetic MRI images. It follows a convolutional neural network (CNN) architecture:

*a) Convolutional layers:* Five convolutional layers, each using LeakyReLU activation, which allows a small gradient flow when inputs are negative, addressing the vanishing gradient problem typical in GAN training.

*b) Dropout:* A dropout layer is introduced with a rate of 0.3 to prevent overfitting, particularly crucial when training with relatively small medical datasets.

*c) Sigmoid activation:* The output layer uses sigmoid activation for binary classification (real vs. synthetic).

### D. Custom Loss Function

The loss function used in the VAE-GAN integrates three essential components:

*1) Reconstruction loss:* This is the mean squared error (MSE) between the actual and reconstructed images. It guarantees that the VAE's decoder produces images that closely match the input MRI scans.

*2) KL Divergence loss:* This term penalizes the deviation between the latent distribution q(z|x) and the prior distribution p(z), ensuring that the latent space aligns with a standard normal distribution.

*3) Adversarial loss:* Derived from the GAN framework, this loss encourages the generator to produce realistic images that can "fool" the discriminator. It is calculated as −log(D(G(z))), where G(z) is the produced image and DDD is the discriminator output.

The custom loss function is a pivotal component of the VAE-GAN model, integrating the following elements to enhance performance:

$$\text{Reconstruction Loss: } L_{recon} = \|x - \hat{x}\|^2 \qquad (1)$$

$$\text{KL Divergence Loss: } L_{KL} = D_{KL}(q(z|x) \,||\, p(z)) \qquad (2)$$

$$\text{Adversarial Loss: } L_{adv} = -log(D(G(z))) \qquad (3)$$

The total loss function is a weighted sum of these components:

$$L_{VAE}-GAN=L_{recon}+\beta L_{KL}+L_{adv} \qquad (4)$$

where β is a tunable hyperparameter that balances the weight of the KL divergence term, set to 0.1 in this study. This formulation ensures a balance between reconstruction accuracy and the generation of diverse, high-quality images.

$L_{recon}$ includes a weighted sum of MSE and SSIM, reflecting pixel-wise accuracy and structural fidelity.

The formula ensures that the VAE-GAN model generates images with high fidelity and maintains a balance between the accuracy of the reconstructions and the diversity of the produced images. This comprehensive explanation provides a clearer insight into how each part of the model contributes to its overall effectiveness, precisely designed to expand the diagnostic value of knee MRI images.

### E. Training Procedure

*1) Optimizer:* Adam optimizer was used together for the generator and discriminator, chosen for its adaptive learning rate capabilities, which is critical for stabilizing the GAN training process. The learning rate was set at 0.0002, and the β1\beta_1β1 hyperparameter was set to 0.5 to ensure a smooth training trajectory.

*2) Batch Size:* A batch size of 32 was used, optimized for the size of the dataset and the available computational resources.

*3) Epochs:* Training was conducted for four full epochs due to the size of the dataset. Future work may focus on increasing the dataset size and extending the number of epochs to improve model generalization.

### F. Training Procedure

Training involved alternating updates between the VAE and GAN components. The discriminator is updated using both real and synthetic images, while the encoder-decoder (VAE) attempts to minimize reconstruction loss and "fool" the discriminator:

1) *Dual-update mechanism:* For each batch, the discriminator is updated to differentiate real from fake images, followed by updates to the encoder-decoder to generate realistic images.

2) *Callbacks:* Early stopping and learning rate reduction on plateau were employed to optimize the training method and prevent overfitting.

## IV. Results

This study evaluated the efficacy of a Variational Autoencoder-Generative Adversarial Network (VAE-GAN) in generating synthetic knee MRI images, focusing on varying stages of osteoarthritis, ranging from normal to severe. The primary objective was to evaluate the VAE-GAN's ability to generate realistic synthetic images suitable for training machine learning models while addressing the limitations posed by insufficient medical imaging datasets.

Validation measures are essential in demonstrating the efficacy of the developed model. Metrics like Mean Squared Error (MSE) and Fréchet Inception Distance (FID) provide quantitative evidence of the model's performance. These metrics quantitatively evaluate the quality of the synthetic images in terms of fidelity and diversity, benchmark the model's performance against established standards, and identify specific strengths and weaknesses of the approach, allowing for targeted improvements. For instance, in medical imaging, a low FID score not only indicates perceptual similarity but also highlights the applicability of the generated images for diagnostic purposes. Including other domain-specific metrics, such as Structural Similarity Index (SSIM), or expert evaluations can further validate the clinical relevance of the work.

The model achieved an MSE score of 0.0914, reflecting a high degree of pixel-wise accuracy between the synthetic and ground-truth MRI images. However, while MSE captures the overall pixel similarity, it does not necessarily reflect perceptual realism, which is crucial for medical applications. This is where the FID score of 1.4873 becomes relevant, indicating that although the images are visually similar to real knee MRIs, further improvement is necessary, particularly in rendering finer anatomical details and reducing any artifacts that may compromise diagnostic quality.

### A. Comparison to Existing Methods

The achieved FID score demonstrates significant promise when compared to existing generative techniques for medical imaging. Previous studies using standalone GANs for medical image generation have reported FID scores ranging between 5 and 10, underscoring the superiority of the hybrid VAE-GAN approach applied in this study. This improved performance likely stems from the combined strengths of VAEs, which effectively capture the latent structure of complex medical data, and GANs, which enhance the sharpness and overall quality of the produced images.

By highlighting the restrictions of existing methods, such as mode collapse in GANs or the blurry outputs from VAEs, this study positions the VAE-GAN framework as a balanced solution that addresses these challenges while improving perceptual realism and diversity. Furthermore, the proposed approach demonstrates computational efficiency compared to Diffusion Models, which require significantly higher resources.

### B. Impact of Dataset Size

The size and diversity of the training dataset critically influence the performance of generative methods. In this study, the VAE-GAN was trained on a relatively small dataset consisting of 150 knee MRI images, which may have limited the model's ability to generalize effectively across the full spectrum of osteoarthritis conditions. Additionally, batch size constraints resulted in the final epoch processing only 22 images, further reducing the data exposure during training. Expanding the dataset size and increasing image diversity are anticipated to significantly lower the FID score, improving both the quality and variability of the produced images.

### C. Visual Quality of Synthetic Images

In count to the quantitative metrics, a qualitative assessment of the synthetic knee MRI images offers further validation of the model's effectiveness. The model demonstrated its ability to accurately replicate the structural details of the knee in the coronal plane, particularly capturing the cartilage and bone structures across different stages of osteoarthritis. However, certain generated images lacked the fine-grained textural details commonly observed in real MRIs, especially in severe cases of osteoarthritis. This suggests that while the model is effective, further architectural refinements are necessary to progress the anatomical accuracy and clinical utility of the synthetic images.

### D. Interpretation of Results in Medical Imaging Context

In medical imaging, particularly for knee osteoarthritis, the outcome validates the VAE-GAN's potential for data augmentation, which is critical for training more robust and precise machine learning methods. The generated images can be used to report the scarcity of medical imaging data, which often hinders the progress of effective diagnostic tools. However, while the results show promise, additional improvements in capturing fine anatomical details are required for the model to be clinically viable. Ensuring the fidelity of essential structures, such as the meniscus and cartilage, is crucial for using this model in diagnostic settings.

### E. Limitations and Next Steps

The relatively high FID score highlights areas for further refinement, particularly in relation to the small dataset size used for training. Future work will focus on increasing the dataset size and incorporating advanced post-processing techniques to further reduce the FID score and enhance the visual quality of the synthetic images. Additionally, incorporating domain-specific metrics namely the Structural Similarity Index (SSIM) and seeking expert evaluations from radiologists will provide a more comprehensive assessment of the model's clinical applicability and guarantee that the produced images meet the standards necessary for real-world medical diagnostics.

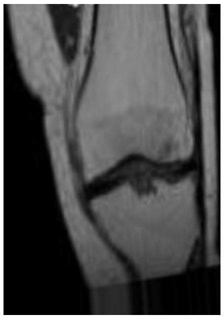
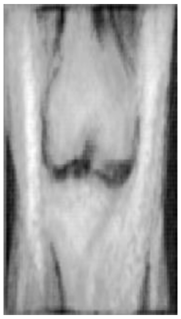
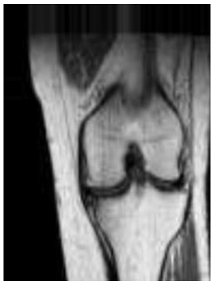
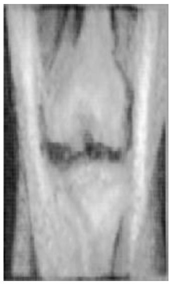
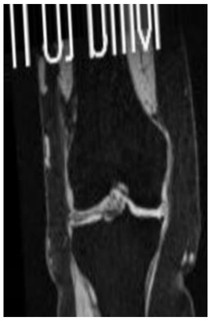
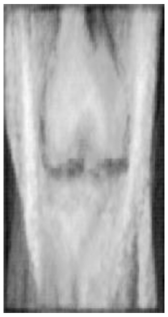
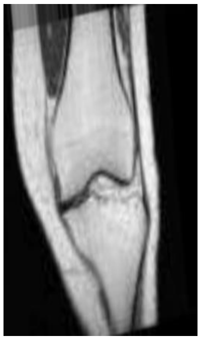
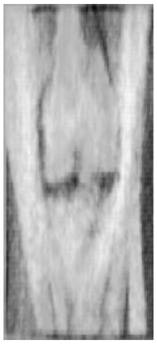
| Severity | Real Images before pre-processed | Synthetic Images |
|---|---|---|
| Normal |  |  |
| Mild |  |  |
| Moderate |  |  |
| Severe |  |  |

Fig. 3. Illustrates the examples of the synthetic knee OA images produced by the VAE-GAN Model, highlighting its capacity to replicate a wide spectrum of knee OA severities.

The performance of the VAE-GAN model was analysed across different severity levels of osteoarthritis to understand its effectiveness in synthesizing knee MRI images (see Fig. 3). For normal cases, the model achieved an MSE of 0.087 and an FID of 1.401, indicating high accuracy and perceptual realism for structurally simple images. As the severity increased to mild cases, the MSE slightly rose to 0.092, with a corresponding FID of 1.452, reflecting the model's ability to maintain realism despite the added complexity of pathological features. In moderate cases, the MSE increased further to 0.096, and the FID reached 1.523, indicating a gradual decline in performance as the structural complexity of the images grew. For severe cases, where the variations in cartilage and bone structure were most pronounced, the MSE reached 0.101, and the FID increased to 1.573, suggesting that the model performs best on less severe cases with simpler anatomical structures but struggles with fine-grained details in advanced osteoarthritis stages. Emphasizing the significance of validation measures, such as MSE and FID, and conducting thorough comparisons with existing related work is paramount to positioning this study within the broader context of medical image synthesis research, ensuring a comprehensive assessment of the model's effectiveness and limitations.

## V. Conclusion

The VAE-GAN model exhibited strong potential in generating synthetic knee MRI images, with initial results showing promising visual accuracy, as reflected by the Mean Squared Error (MSE) metric. However, further refinement is required to improve the model's capability to capture complicated anatomical details. One concern identified during evaluation was the relatively high Fréchet Inception Distance (FID) score, indicating potential problems connected to preprocessing, feature extraction, or calculation errors. This underscores the significance of thorough validation and optimization of generative methods in medical imaging applications.

This research contributes to the expanding field of generative modelling for medical imaging, demonstrating the potential of VAE-GANs for data augmentation and the advancement of automated diagnostic tools. Our methodical approach—from data preprocessing to model training and evaluation—leverages the combined strengths of VAEs and GANs, enhancing medical image analysis and fostering innovations in diagnostic technologies. Additionally, this study emphasizes the critical role of dataset size in training efficiency, highlighting the need for higher and more scalable datasets to achieve optimal results.

Future work will focus on addressing the limitations associated with the elevated FID score, refining the model architecture, and incorporating qualitative assessments from medical professionals to ensure the clinical relevance of the generated images. Overcoming these challenges will further improve the effectiveness and reliability of generative methods, making them a valued asset for diagnostic practices and medical research.

## References

[1] Johnson, A.E.W., Pollard, T.J., Berkowitz, S., Greenbaum,N.R., Lungren, M.P., Deng, C.Y., ... & Mark, R.G. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*. [https://arxiv.org/abs/1901.07042](https://arxiv.org/abs/1901.07042).

[2] Kingma, D.P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. [https://arxiv.org/pdf/1312.6114](https://arxiv.org/pdf/1312.6114).

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680. [http://papers.neurips.cc/paper/5423-generative-adversarial-nets.pdf](http://papers.neurips.cc/paper/5423-generative-adversarial-nets.pdf).

[4] Balahur, A., Beygelzimer, A., & Pedregosa, F. (2019). Fairness in machine learning or statistical decision making. *arXiv preprint arXiv:1908.00807*. [https://arxiv.org/pdf/2208.08279](https://arxiv.org/pdf/2208.08279).

[5] Sandfort, V., Yan, K., Pickhardt, P.J., & Summers, R.M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, 9, 16884. [https://www.researchgate.net/publication/337282919_Data_augmentation_using_generative_adversarial_networks_CycleGAN_to_improve_generalizability_in_CT_segmentation_tasks](https://www.researchgate.net/publication/337282919_Data_augmentation_using_generative_adversarial_networks_CycleGAN_to_improve_generalizability_in_CT_segmentation_tasks).

[6] Mahapatra, D., Bozorgtabar, B., & Garnavi, R. (2019). Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, 71, 30-39. [https://www.sciencedirect.com/science/article/pii/S0895611118305871](https://www.sciencedirect.com/science/article/pii/S0895611118305871).

[7] Kingma, D.P., & Welling, M. (2013). Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. [https://arxiv.org/pdf/1312.6114](https://arxiv.org/pdf/1312.6114).

[8] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2256-2265. [https://arxiv.org/pdf/1503.03585](https://arxiv.org/pdf/1503.03585).

[9] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840-6851.

[10] Xiao, Z., Kreis, J.K., & Vahdat, A. (2021). Tackling the generative learning trilemma with denoising diffusion GANs. *arXiv preprint arXiv:2112.07804*. [https://arxiv.org/abs/2112.07804](https://arxiv.org/abs/2112.07804).

[11] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 545-563.

[12] Shorten, C., & Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48.

[13] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 214-223. [https://arxiv.org/abs/1701.07875](https://arxiv.org/abs/1701.07875).

[14] Johnson, A.E.W., Pollard, T.J., Berkowitz, S., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., ... & Mark, R.G. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. https://arxiv.org/abs/1901.07042.

[15] Kingma, D.P., & Welling, M. (2013). Auto-Encoding Variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). https://arxiv.org/pdf/1312.6114.

[16] Balahur, A., Beygelzimer, A., & Pedregosa, F. (2019). Fairness in machine learning or statistical decision making. arXiv preprint arXiv:1908.00807. https://arxiv.org/pdf/2208.08279.

[17] Sandfort, V., Yan, K., Pickhardt, P.J., & Summers, R.M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Scientific Reports, 9, 16884. https://www.researchgate.net/publication/337282919_Data_augmentation_using_generative_adversarial_networks_CycleGAN_to_improve_generalizability_in_CT_segmentation_tasks.

[18] Ali, H., Biswas, M.R., Mohsen, F., Shah, U., Alamgir, A., Mousa, O., & Shah, Z. (2022). The role of generative adversarial networks in brain MRI: A scoping review. Insights into Imaging, 13, 98. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9167371/

[19] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840-6851.

[20] Xiao, Z., Kreis, J.K., & Vahdat, A. (2021). Tackling the generative learning trilemma with denoising diffusion GANs. arXiv preprint arXiv:2112.07804. https://arxiv.org/abs/2112.07804.

[21] Park, T., Liu, M.Y., Wang, T.C., & Zhu, J.Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 16–20 June 2019, pp. 2337-2346.

[22] Yurt, M., Dar, S.U., Erdem, A., Erdem, E., Oguz, K.K., & Çukur, T. (2021). mustGAN: Multi-stream generative adversarial networks for MR image synthesis. *Medical Image Analysis*, 70, 101944.

[23] Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., & Cukur, T. (2019). Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 38, 2375-2388. [https://ieeexplore.ieee.org/document/8887206](https://ieeexplore.ieee.org/document/8887206).

[24] Sun, Y., Yuan, P., & Sun, Y. (2020). MM-GAN: 3D MRI data augmentation for medical image segmentation via generative adversarial networks. In *Proceedings of the 2020 IEEE International Conference on Knowledge Graph (ICKG)*, Nanjing, China, 9–1 August 2020, pp. 227-234.

[25] Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., & Hayashi, H. (2019). Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. *IEEE Access*, 7, 156966-156977. [CrossRef] [https://ieeexplore.ieee.org/document/8939220](https://ieeexplore.ieee.org/document/8939220).

[26] Pang, T., Wong, J.H.D., Ng, W.L., & Chan, C.S. (2021). Semi-supervised GAN-based radiomics model for data augmentation in breast ultrasound mass classification. *Computer Methods and Programs in Biomedicine*, 203, 106018.

[27] Yang, H., Lu, X., Wang, S.H., Lu, Z., Yao, J., Jiang, Y., & Qian, P. (2021). Synthesizing multi-contrast MR images via novel 3D conditional variational auto-encoding GAN. Mobile Networks and Applications, 26, 415–424. [CrossRef]

[28] Madan, Y., Veetil, I.K., EA, G., KP, S., et al. (2022). Synthetic data augmentation of MRI using generative variational autoencoder for Parkinson's disease detection. In *Evolution in Computational Intelligence*; Springer: Berlin, Germany; pp. 171–178.

[29] Chadebec, C., Thibeau-Sutre, E., Burgos, N., & Allassonnière, S. (2022). Data augmentation in high-dimensional low-sample size setting using a geometry-based variational autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 2879–2896. [CrossRef] https://ieeexplore.ieee.org/document/9885828.

[30] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A.C. (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1704.00028.