Korean Journal of Radiology

KJR

Check for updates

Image-Based Generative Artificial Intelligence in Radiology: Comprehensive Updates

Ha Kyung Jung[1]*, Kiduk Kim[2]*, Ji Eun Park[1], Namkug Kim[1,2]

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[2]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

Generative artificial intelligence (AI) has been applied to images for image quality enhancement, domain transfer, and augmentation of training data for AI modeling in various medical fields. Image-generative AI can produce large amounts of unannotated imaging data, which facilitates multiple downstream deep-learning tasks. However, their evaluation methods and clinical utility have not been thoroughly reviewed. This article summarizes commonly used generative adversarial networks and diffusion models. In addition, it summarizes their utility in clinical tasks in the field of radiology, such as direct image utilization, lesion detection, segmentation, and diagnosis. This article aims to guide readers regarding radiology practice and research using image-generative AI by 1) reviewing basic theories of image-generative AI, 2) discussing the methods used to evaluate the generated images, 3) outlining the clinical and research utility of generated images, and 4) discussing the issue of hallucinations.
Keywords: Generative artificial intelligence; Generative adversarial networks; Diffusion models; Evaluation metrics; Medical imaging

INTRODUCTION

Generative artificial intelligence (AI), an umbrella term that uses deep learning (DL) and other machine learning techniques to extract features from underlying structures in source data (e.g., images) and generates artificial data using AI models, such as generative adversarial networks (GANs), diffusion models, or transformer-based models [1,2].

A combination of supervised and unsupervised learning methods is used in generative AI. The algorithm is trained on labeled datasets to predict new images of specific outputs in supervised learning. Unsupervised learning, which is used to determine distributions, clusters, and relationships in the data, is employed in scenarios wherein specific output has not been assigned to the input data. In any form, generative AI aims to generate diverse content as access to large annotated data is often challenging and requires time and labor.

Image-based generative AI (hereafter referred to as image generative AI), including GAN and diffusion models, is commonly used for image quality enhancement, domain transfer, and imputation as augmented input data in the field of medical imaging. Synthetic image data have enhanced lesion and image quality in ophthalmology [3], enabled denoising of electrocardiogram output [4], and augmented data with synthetic photos of skin lesion-enhanced lesion classifier performance [5]. In radiology, generative AI has been used to generate image data from radiographs, mammography, ultrasonography (US), computed tomography (CT), and magnetic resonance imaging (MRI). A paradigm shift has been observed with the application of

foundation models, trained on a large and broad amount of data, across multiple downstream tasks [2,6,7]. The potential of foundation models lies primarily in text data; however, their potential for performing tasks with images has garnered an increasing amount of attention. This may be attributed to importance of generating sufficient unannotated data as an important input for foundation models based on self-supervised learning. Therefore, achieving a good grasp of generative AI, particularly in image-based applications, and exploring its applications is necessary.

The diverse usage of generative (synthetic) images in the field of radiology can be categorized into two approaches (Fig. 1): Approaches 1 and 2. Approach 1 involves utilizing synthetic images without any modification, whereas Approach 2 involves leveraging synthetic images as training data for AI modeling to enhance performance. Approach 1 has been used in previous studies to generate missing sequences, generate images in less harmful ways, generate better-quality images, and ensure feature reproducibility depending on the type of image generation. Approach 2 has been used as training data for downstream DL tasks (supervised learning), including lesion detection, diagnosis, and segmentation, in previous studies. This review summarizes the basic theories of image-generative AI, the strengths and weaknesses of the evaluation metrics, and the diverse applications of image-generative AI in the field of radiology according to their
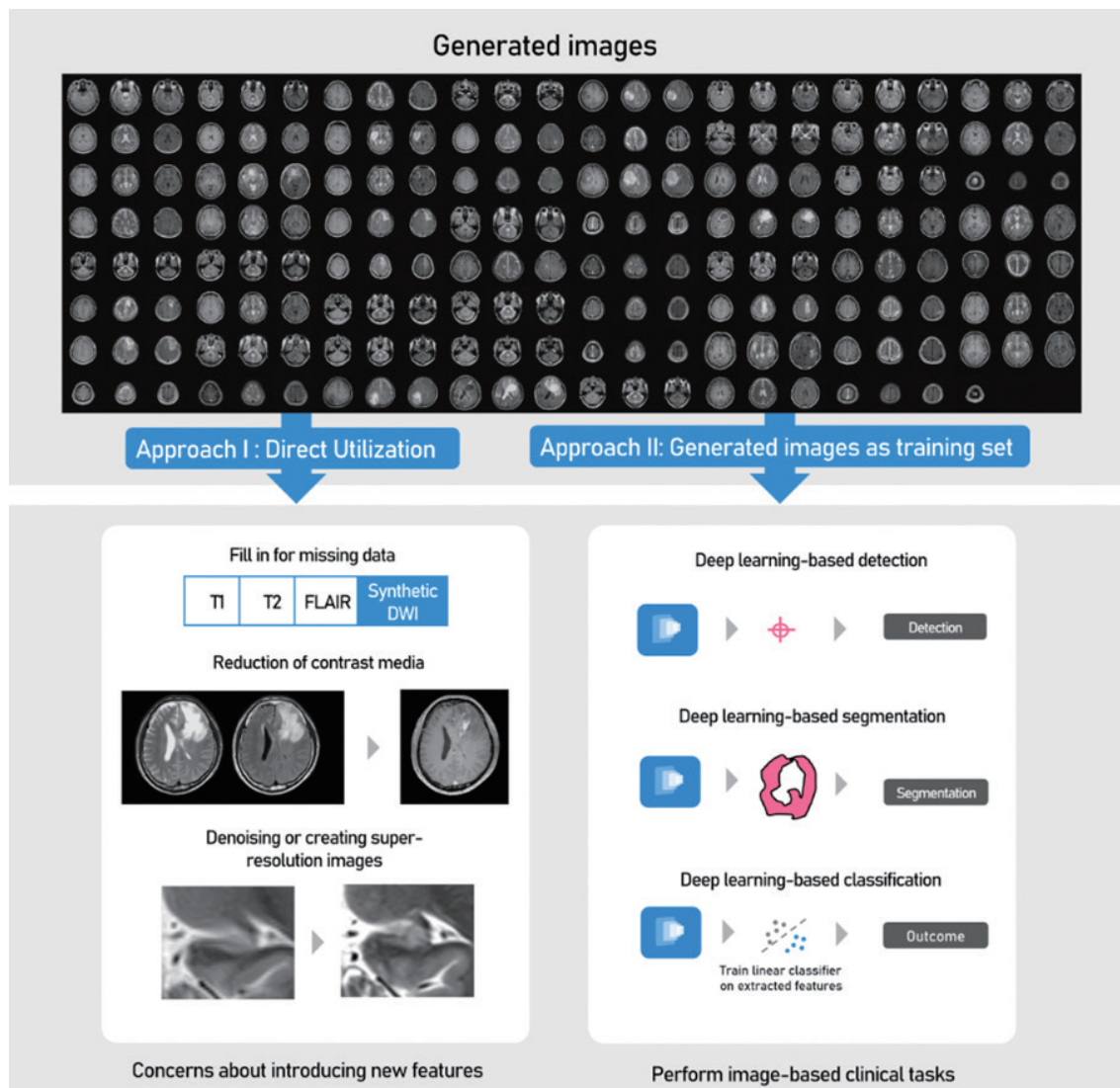
Fig. 1. Clinical utility of image generative artificial intelligence. Approach 1 utilizes synthetic images as themselves, while Approach 2 utilizes real and synthetic images as augmented input data to enhance performance in clinical tasks, such as lesion detection, segmentation, and diagnosis.

Basic Theories of Image Generative AI

### Hierarchy of Various Modeling Technique

The core objective of a generative model is to model the probability distribution of data. However, deterministic modeling of the distribution is challenging owing to the complex and high dimensional nature of the true data distribution probability. The dataset is considered to follow a certain probability distribution in a generative model and trained accordingly. Thus, new samples are generated probabilistically, making generative models inherently random. Such models are called probabilistic or stochastic models [2].

Generative models can be categorized into explicit and implicit models based on the modeling of the data distribution probability. Explicit density models generate new samples based on p (x) by directly modeling the data distribution of p (x). Variational autoencoders (VAEs) [8-10], PixelCNNs [11-13], and some early diffusion model families [14,15] are explicit density generative models. Implicit density models use a generative process wherein sampling is performed via generative processes to implicitly capture the data distribution; thus, the data distribution p (x) is not explicitly defined. GAN families [16-19] and some diffusion models [20-22] employ implicit density generative models.

Explicit generative models can be categorized into tractable and non-tractable, depending on whether the probability distribution or likelihood function associated with the model can be solved (tractable). Autoregressive generative models such as PixelCNNs and Normalizing Flows [23-26] are examples of tractable models. Autoregressive models factorize the probability density function of all the pixels of a given image while normalizing flow models transform a probability distribution through a series of invertible mappings. Non-tractable models approximate the probability distribution using techniques such as Markov chain Monte Carlo (MCMC) and variational inference. VAEs, some explicit diffusion models, and energy-based [27-30] are non-tractable models. Figure 2 presents the taxonomy of image-generative AI.

GANs and diffusion models, which have become the most popular generative AIs in recent years, are the focus of this review. Figure 3 presents the structures of the GANs and diffusion models. Both methods generate samples by learning the data distribution; however, the sampling characteristics and image generation techniques vary. GAN is trained by comparing a generator with a discriminator; the trained generator generates a sample from a random latent vector [16]. In contrast, noise is iteratively added to an image in steps (forward process) by diffusion models to generate an image by denoising the noisy image (reverse process) [14]. Thus, diffusion models generate better-quality images than GANs [31]. Unlike diffusion models, which can learn more stably, GANs are susceptible to mode collapse [32,33], wherein the discriminator and generator fail to compete [20,31]. However, because diffusion models are computationally complex and require a longer duration to longer to generate as they use MCMC to train and generate images across many iterations of noise diffusion and denoising [21]. Furthermore, unlike GANs, diffusion models
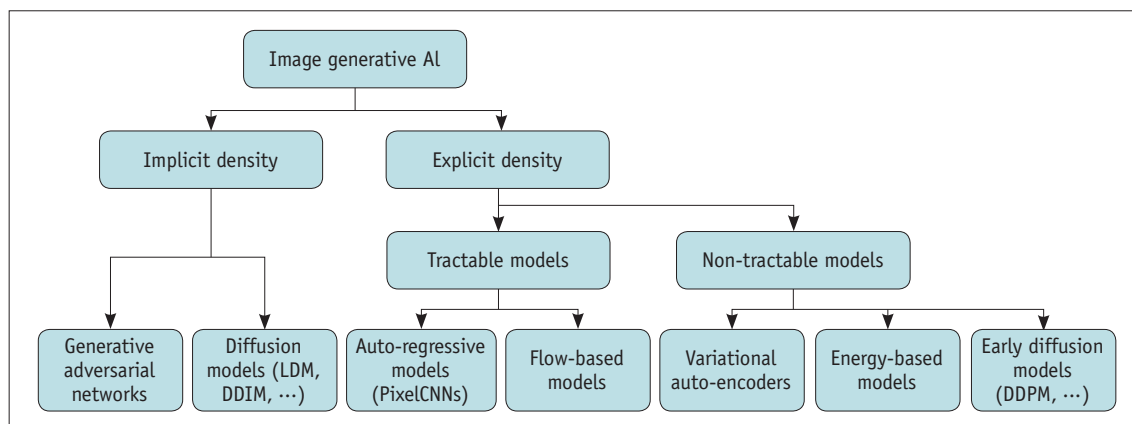
Fig. 2. Hierarchy of image generative AI. Image generative AI can be categorized as implicit and explicit density models. Explicit density models can be further categorized as tractable and non-tractable models. AI = artificial intelligence, LDM = latent diffusion model, DDIM = denoising diffusion implicit model, CNN = convolutional neural network, DDPM = denoising diffusion probabilistic model
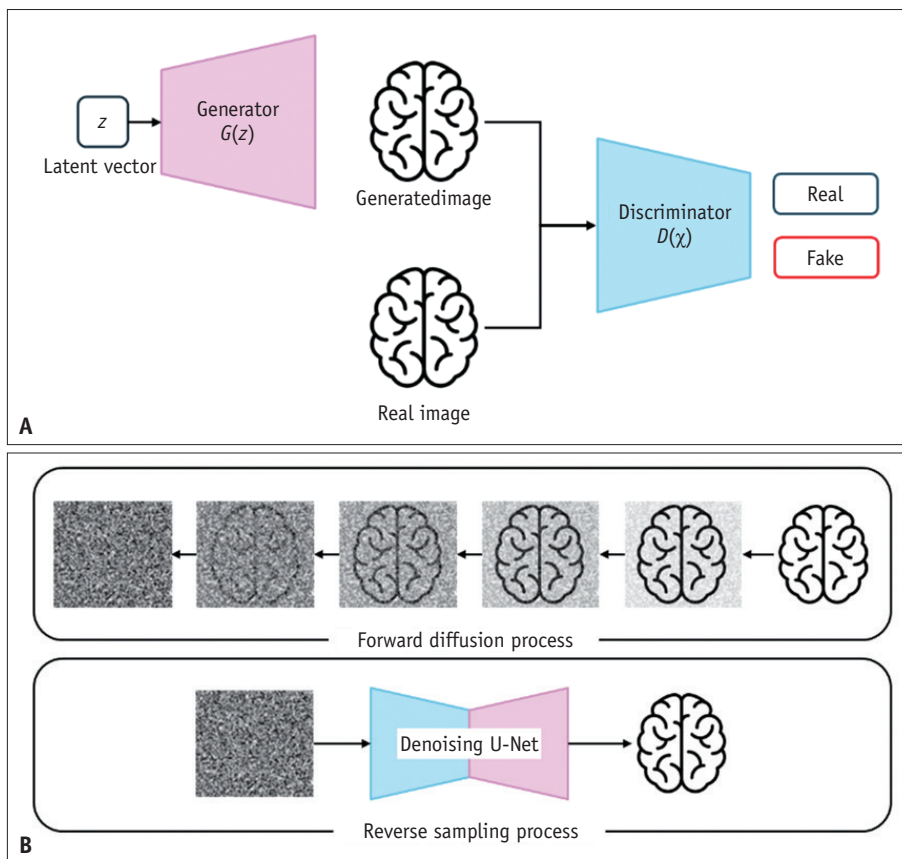
Fig. 3. Schema of the GAN **(A)** and diffusion models **(B)**. GAN uses a generator and discriminator, whereas diffusion models exhibit gradual denoising and image generation steps. GAN = generative adversarial network

do not use techniques for training on small datasets [34] and require larger datasets for training [35].

Controllable Image Generation and Manipulation

Generative models can be used to generate samples; however, they are often used in combination with other techniques such as conditional or controllable generation, to produce the desired outcomes. Strict divisions have not been made; however, common techniques include conditional image synthesis, which generates images of a desired class; image editing and manipulation techniques, which transform an input image into the desired shape; and image-to-image translation (and/or style transfer), which transforms an input image into the desired style. The Supplement describes the details of conditional image synthesis, image editing and manipulation, and image-to-image translation.

Towards Multimodal Generative AI

Generative AI for medical imaging driven by computer vision has made significant strides; however, its practical application remains limited. Medical images must be

analyzed in conjunction with electronic medical records (EMRs) and patient demographics in clinical settings. Multimodal learning, such as vision-language modeling (VLM), offers a solution [36].

Multimodal learning employs images generated using various methods to process diverse clinical information [2]. These models accept and output data in visual and textual formats. For instance, a VLM can generate structured reports from chest radiographs (CXRs) and create CXR images from brief reports [37]. Text-guided image-to-image translation [38,39] has been used for cross-modality medical image translation [40,41] or anomaly detection via counterfactual image generation [42,43]. Multimodal generative AI, including VLMs, can overcome the limitations of image-only models, representing a significant advancement.

Metrics and Methods for Evaluating Generative Images

Notably, methods using the human eye, such as the visual Turing test and qualitative evaluation, have been used to

evaluate image-generative AI [44]. However, these methods are highly subjective; thus, quantitative metrics that can objectively evaluate generative performance must be developed. Generative models capable of creating realistic samples that reflect real-world data and are sufficiently diverse to reflect data distributions must be developed. Therefore, two crucial aspects of the generated images, fidelity (image quality) and diversity (image variety), are assessed by the technical quantitative metrics of image-generative models. Most metrics are evaluated on a dataset basis as the training of a generative model aims to model real-world data distributions.

A major challenge associated with the use of generative models is the identification of a universal gold standard quantitative metric [45,46]. Therefore, technical and quantitative metrics commonly used for evaluating image-generative AI are summarized in this review. Table 1 summarizes the evaluation metrics of the generated images.

Traditional Metrics

The performance of image-generative AI is quantitatively evaluated using traditional metrics such as the peak signal-to-noise ratio (PSNR) [47] and structural similarity index measure (SSIM) [48]. PSNR is defined as ratio of the logarithmic value of the maximum possible pixel value in an image to the mean squared error of the pixel values between the original and compared images. SSIM can capture the visual characteristics of human perception. It uses the mean, standard deviation, and covariance of two images to evaluate the luminance, contrast, and structure of the transformed images. The root mean square error between the original and transformed images has also been used for evaluation. However, traditional metrics evaluate information loss during image compression and require original images for comparison, thereby limiting their ability to evaluate diversity, which is an important feature of generative AI. Thus, these metrics are of limited use in tasks such as supervised image-to-image translation, wherein absolute labels are available and image super-resolution.

Data Distribution Evaluation Metrics

The inception score (IS) [49] and Fréchet Inception distance (FID) [50] are modern quantitative metrics that are widely used to evaluate the capability of generative AI using the distribution of generated and real images in the latent space. A pretrained DL model (e.g., the ImageNet-pretrained Inception v3 model) has been used to model the conditional label distribution of the generated images [51] to measure the IS. However, IS can also be measured using the Kullback–Leibler divergence and conditional label distribution [52]. A higher IS indicates a better performance. A pretrained Inception v3 model, which compares real images with generated images, has been used to determine FID. FID is calculated by comparing the mean and covariance of all activation vectors obtained using the inception network on the real and generated datasets, with a lower FID score indicating better performance.

Precision and recall have also been used to evaluate generative models in terms of the distribution of real and generated datasets in latent space [53]. Precision is defined as the proportion of true positives among those classified as positive in diagnostic metrics, whereas recall is defined as the proportion of true positives among the total number of real positives. In terms of the distribution of the real and generated datasets, precision can be expressed as the generated data within the real data distribution, whereas recall can be expressed as the real data within the generated data distribution in the latent space. Thus, precision and recall are correlated with the fidelity and diversity of the generated images, respectively. A pretrained network that embeds all data as latent vectors in the latent space [50] and uses the k-means clustering algorithm [54] has been used to compare the distributions of the datasets as direct estimation of the distribution of real and generated datasets is challenging. Studies have been conducted to improve the precision and recall, including the density and coverage [54,55]. Figure 4 presents the difference between the diagnostic and generative precision recall.

Metrics and Methods for Evaluating Downstream Applications

Image generative AI enables the use of generated images as training AI data for clinical or research purposes. Table 2 summarizes the common evaluation metrics and methods used for downstream applications. Previous studies have detailed AI evaluation metrics and methods [56-58]. The visual Turing test has been used to verify the realism of images before applying AI-generated images as data for model training or educational materials. Overlap or distance measures, such as Intersection over Union and Dice similarity coefficient (DSC), have been used to assess segmentation tasks. A confusion matrix and receiver operating characteristic (ROC) curve analysis have been used to evaluate classification tasks, with the F1 score serving as a harmonic mean of positive

**Table 1.** Metrics for evaluating generated images

Metrics	Details	Tasks	Meaning
Traditional method			
PSNR	1. Ratio between the maximum pixel value and the distorting effect of reconstruction 2. Higher PSNR indicates better image quality 3. Does not always correlate with human perception	Supervised image-to-image translation	Fidelity
SSIM	1. Measure similarity between two images using luminance, contrast, and structure 2. Higher SSIM indicates better image quality	Supervised image-to-image translation	Fidelity
Data distribution method			
IS	1. A score that measures the fidelity and diversity of generated images at once 2. Calculate KL divergence between the marginal and conditional distribution of generated images using pretrained classification model 3. Higher IS indicates better fidelity and diversity 4. It can be counterintuitive and fooled based on the pretrained classification model 5. Require approximately 5000 images of generated images	1. Controllable image generation 2. Image editing and image manipulation 3. Supervised and unsupervised image-to-image translation	Fidelity and diversity
FID	1. A score that measures the fidelity and diversity of generated images at once compared to real images 2. Calculate multivariate Gaussian distribution between real and generated images using pretrained classification model 3. Lower FID indicates a more similar distribution between generated and real images 4. Result can be biased according to the pretrained classification model 5. Require approximately 50000 samples	1. Controllable image generation 2. Image editing and image manipulation 3. Supervised and unsupervised image-to-image translation	Fidelity and diversity
Precision & Recall	1. Precision measures for fidelity while recall measures for diversity 2. Precision is calculated by the valid generated samples over all the generated samples in the latent space, while recall is calculated by the valid generated samples over all the real samples in the latent space 3. Achieving both high precision and high recall is desirable, but there is a trade-off between precision and recall 4. There exist some variants in precision and recall, such as density and coverage	1. Controllable image generation 2. Image editing and image manipulation	Diversity

PSNR = peak signal-to-noise ratio, SSIM = structural similarity index measurement, IS = inception score, KL = Kullback–Leibler, FID = Fréchet inception distance

predictive value and sensitivity. Obtaining false negative and true negative rates for detection tasks is challenging; therefore, detection sensitivity and false positive rates are typically used. ROC analysis involves the calculation of localization ROC, free-response ROC (FROC), or alternative FROC (AFROC) based on lesion-level sensitivity and false positive rates. Table 2 describes each of these measures. However, the details of each measure are beyond the scope of this article. Previous studies have assessed the methods used for clinical evaluation of AI algorithms for medical diagnosis [56] and performance metrics of machine learning [57,58].

Clinical and Research Usage: Approach 1–Direct Utilization of Synthetic Images

Approach 1 utilizes synthetic images, whereas Approach 2 uses synthetic images as augmented input data along with real images to enhance performance in clinical tasks (Fig. 1). GAN and diffusion models, two commonly used image-generative AI techniques in medical imaging, are illustrated and discussed here. Table 3 summarizes the published articles using Approach 1 [59-75].
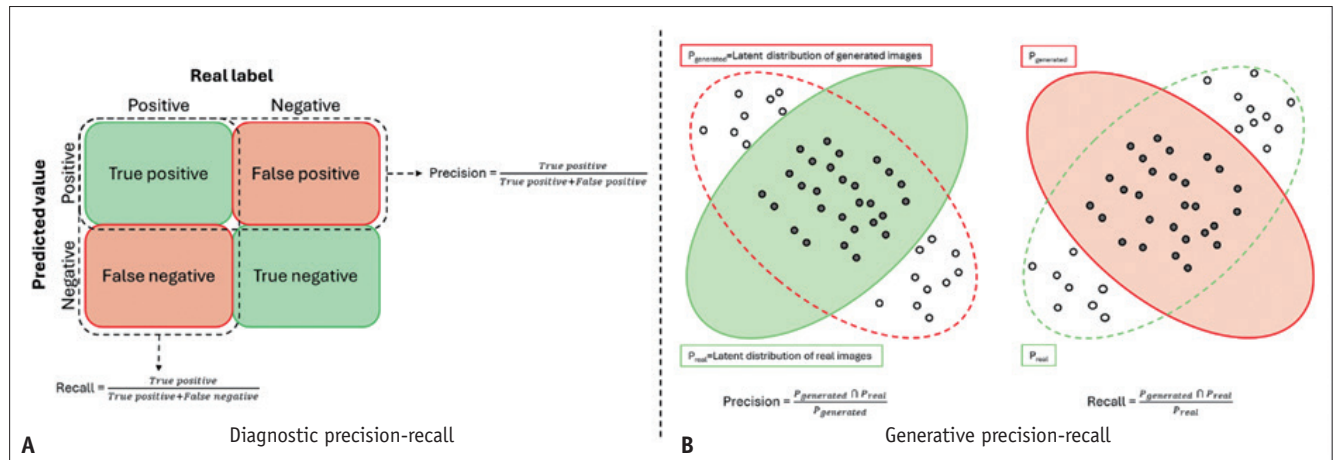
Fig. 4. Explanation of diagnostic **(A)** and generative precision-recalls **(B)**. Diagnostic precision is identical to positive predictive value, whereas diagnostic recall is identical to sensitivity. Generative precision-recall overlaps with the latent distribution of real data. Generative precision indicates the overlap area divided by the latent distribution of generated data, whereas generative recall indicates overlap area divided by the latent distribution of real data.

Table 2. Common metrics and methods for downstream applications of AI-generated images

Metrics	Details	Tasks
Visual Turing's test	1. Create mixed dataset of generated and real samples for human raters to distinguish between them 2. Highly dependent on the human raters 3. Realism and 3D continuity of generated samples can be evaluated by human raters using Likert scale	Substitution of real images (e.g., data for model training, educational use)
IoU	1. Measures overlap between the predicted object contour and ground truth contour 2. Ranging from 0 (no overlap) to 1 (complete overlap)	Segmentation
DSC	Twice the overlapped area between the ground truth and predicted bounding boxes divided by sum of their area	
ROC	Evaluate performance by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold levels	Classification
Confusion matrix (accuracy, sensitivity, specificity, F1 score)	1. Accuracy: fraction of correct predictions 2. Sensitivity: fraction of positives correctly predicted 3. Specificity: fraction of negatives correctly predicted 4. F1 score: harmonic mean of positive predictive value and sensitivity	
Detection sensitivity, false positive rates	1. In the detection task setting, false negative and true negative is unknown 2. Sensitivity: fraction of positives correctly predicted 3. False positives: fraction of negatives incorrectly predicted as positive	Detection

LROC, FROC, AFROC	1. LROC
	  - X-axis: false positive rate (1 - specificity) at the case level
	  - Y-axis: probability of correct localization at the case level or the fraction of true positives that correctly hits annotated lesion
	2. FROC
	  - X-axis: average number of false positives per case
	  - Y-axis: lesion localization fraction or lesion level sensitivity
	3. AFROC
	  - X-axis: false positive rate (1 - specificity) at the case level
	  - Y-axis: lesion localization fraction or lesion level sensitivity	

AI = artificial intelligence, 3D = three-dimensional, IoU = intersection over union, DSC = Dice similarity coefficient, ROC = receiver operating characteristic, LROC = localization ROC, FROC = free-response ROC, AFROC = alternative FROC

Table 3. Direct utilization of synthetic images

Clinical utility	Authors	Network	Image data	Study purpose	Number of patients (training/validation/test set)	Evaluation methods
Filling in missing sequences	Ozbey et al. [69]	Adversarial diffusion model	Brain MRI T1, T2, PD/pelvic MRI/pelvic CT	Generate images among brain T1, T2, PD sequences, and pelvic MR to CT	IXI 25/5/10 BRATS 25/10/20 Pelvic 9/2/4	PSNR and SSIM
	Conte et al. [61]	GAN	Brain MRI T1, T2, FLAIR, and contrast-enhanced T1	1. Generate T1w from contrast-enhanced T1 and generative FLAIR from T2 2. Tumor segmentation	BRATS 135/33/42	Mean squared error, SSIM, and DSC
	Schlaeger et al. [71]	GAN	Brain MRI FLAIR	Generate double inversion recovery from FLAIR	74 (longitudinal total n = 214)/NA/NA	Counts of lesion detection in multiple sclerosis
Reducing radiation	Xia et al. [73]	Denoising diffusion probabilistic models	Breast cone-beam CT	Generate low-dose CT reconstruction to reduce radiation	84/0/21	Ability to reconstruct half-dose and one-third-dose
	Emami et al. [62]	GAN	Brain MRI contrast-enhanced T1 to CT	Generate simulated CT from contrast-enhanced T1	12/3/NA	MAE, PSNR, and SSIM
	Xie et al. [75]	Joint probability distribution of diffusion model	Brain MRI T1 to PET	Generate PET images from brain MRI images (3T, 5T, and 7T)	ADNI (split is unclear)	PSNR
	Lin et al. [66]	3D-reversible GAN	Brain MRI T1 to PET/PET to brain MRI T1	Generate brain MRI to PET and PET to MRI	ADNI (split is unclear)	RMSE, PSNR, and SSIM
Reducing contrast media	Lyu et al. [67]	GAN	Neck CT angiography	1. Generate contrast-enhanced CT angiography from non-contrast CT angiography 2. Compare diagnostic accuracy	1137/400/212 (internal) and 42 (external)	PSNR, SSIM, and visual quality, diagnostic accuracy
	Preetha et al. [70]	Cycle GAN	Brain MRI T1, T2 and FLAIR	1. Generate contrast-enhanced T1 from pre-T1, T2, and FLAIR 2. Calculate volumetry and estimate association with time to progression	1540 (775 + 260 + 505)/NA/521	SSIM, MAE, DSC for volumetry, and hazard ratio

Table 3. Direct utilization of synthetic images (continued)

Clinical utility	Authors	Network	Image data	Study purpose	Number of patients (training/validation/test set)	Evaluation methods
Generating better-quality images	Xiao et al. [74]	Hybrid GAN + UNet	Pelvic MRI T2/ brain MRI T1, T2, FLAIR/ abdomen CT	Generate higher-quality images of pelvic and brain MR and abdomen contrast-enhanced CT images	200/66/66	PSNR and SSIM
	Chung & Ye [60]	Score-based diffusion model	Knee MRI	1. Generate high-resolution images from a fast MRI dataset 2. Test detection performance using generated data	973 (volumes) (split is unclear)	PSNR and SSIM
	Wicaksono et al. [72]	Conditional GAN	Brain MR angiography	1. Generate high-resolution (1 mm) time-of-flight MR angiography from low-resolution MR angiography 2. Test diagnostic performance of stenosis and aneurysm	50/0/130	SSIM, visual score, and 2 by 2 diagnostic test accuracy
	Küstner et al. [64]	GAN	Coronary MR angiography	Generate high-resolution coronary MR angiography	50/0/16	Normalized MSE and SSIM
	Bae et al. [59]	GAN	Chest X-ray	1. Generating bone-suppressed Chest X-rays 2. Compare pulmonary nodule detection performance of human readers	348/0/111	Area under the AFROC curve
Improving feature reproducibility	Marcadent et al. [68]	Cycle GAN	Chest X-ray from different vendors	Translate images among different manufacturers and reduce inter-manufacturer variability of radiomics features	6528/0/914 (test 1) and 200 (test 2)	Concordance correlation coefficient of radiomics features
	Lee et al. [65]	GAN	Low-dose abdomen CT	Generate denoised images and improve reproducibility of radiomics features from generated images	2016-NIG-AAPM-Mayo Clinic Challenge Data 2254 slices/0/224 slices	PSNR, SSIM, and concordance correlation coefficient
	Hwang et al. [63]	RouteGAN	Chest CT from different vendors with different radiation doses and kernels	1. Convert images into target CT style 2. Improve accuracy and lesion quantifying function of deep learning-based software	5009 slices/50 slices/ 150 slices	DSC, recall, precision

MRI = magnetic resonance imaging, PD = proton density, CT = computed tomography, MR = magnetic resonance, PSNR = peak signal-to-noise ratio, SSIM = structural similarity index measurement, GAN = generative adversarial network, FLAIR = fluid-attenuated inversion recovery, NA = not available, T1w = T1-weighted, DSC = Dice similarity coefficient, MAE = mean absolute error, PET = positron emission tomography, ADNI = Alzheimer's Disease Neuroimaging Initiative, RMSE = root mean squared error, AFROC = alternative free-response receiver operating characteristic

Generation of Missing Sequences

In the big data era, many DL applications are data-hungry when training a model. When establishing a DL model for CT or MRI, limitations may arise during training due to missing data caused by artifacts, variations in image acquisition parameters, and time constraints. Generative AI synthesizes and fills in the missing data to resolve this limitation. An adversarial diffusion model that can generate missing images has been used to perform multi-contrast MRI and MRI-CT translations [69]. Pelvic MRI to CT image translation exhibited superior quantitative and qualitative image quality compared with that of conventional diffusion models and GANs when applied to MRI sequences, including T1, T2, and PD sequences.

A multicenter study on brain tumor imaging demonstrated the feasibility of DL-based tumor segmentation by filling in of missing sequences [61]. Fluid-attenuated inversion recovery (FLAIR) and contrast-enhanced T1-weighted (T1w) images play crucial roles in brain tumor segmentation models. Synthetic images were obtained using the two GANs to generate T1w and FLAIR images from the contrast-enhanced T1w and T2w images, respectively (Fig. 5). Transfer learning (discussed later) of the DL segmentation was applied using these synthesized images. The segmentation results were compared with the original data segmentation results without missing data. The DSCs for lesion segmentation utilizing the generated images were comparable with those of the original scans. The median DSCs for the segmentation of the whole lesion, FLAIR hyperintensities, and contrast-enhanced areas using the generated scans were 0.82, 0.71, and 0.92, respectively, on replacing T1w and FLAIR images; 0.84, 0.74, and 0.97, respectively, on replacing the FLAIR images only; and 0.97, 0.95, and 0.92, respectively, on replacing the T1w images only.

The double inversion recovery (DIR) sequence is the most sensitive imaging tool for detecting demyelinating plaques in multiple sclerosis. However, the acquisition time of DIR sequence is longer than that of routine FLAIR owing to the presence of two different inversion pulses. DIR images can be synthesized from FLAIR in patients with multiple sclerosis. Two neuroradiologists demonstrated the improved sensitivity of MS plaques in synthetic DIR sequences compared with that in FLAIR sequences ($P < 0.001$) [71].

Acquisition of Images in Less Harmful Way

Patients undergoing imaging studies are exposed to radiation and iodinated or gadolinium-based contrast agents (GBCAs). Generative AI can generate contrast-enhanced images from non-contrast or CT images using MRI images without subjecting patients to radiation exposure, thereby
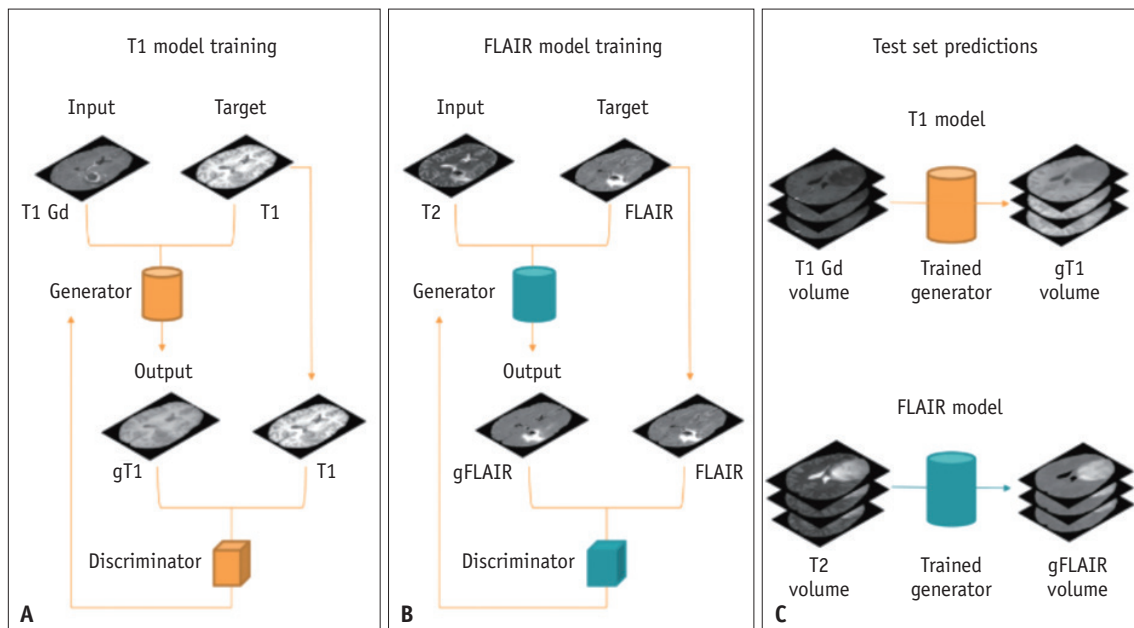
Fig. 5. Training process of GAN to generate synthetic T1w **(A)** and FLAIR **(B)** images. Synthetic images are obtained using two GANs, one for generating T1w images from contrast-enhanced T1w images **(A)** and another for generating FLAIR images from T2w images **(B)**. Full volumes of synthetic T1w and FLAIR images are obtained from the test set using trained generators **(C)**. Adapted from Conte et al. *Radiology* 2021;299:313-323, with permission of Radiological Society of North America [61]. GAN = generative adversarial network, T1w = T1-weighted, FLAIR = fluid-attenuated inversion recovery, T2w = T2-weighted

reducing patient risk.

Reducing the Risk of Radiation

Image-to-image translation involves translating a source image into a target image such that certain visual properties of the original images are preserved. MRI is non-invasive and free of ionizing radiation; thus, previous studies have explored image translation from MRI to CT or positron emission tomography (PET) imaging and avenues for reducing radiation dose from standard to low-dose CT.

Dedicated breast CT provides images of higher quality than mammography and tomosynthesis; however, it subjects patients to higher radiation exposure [73]. Cone beam breast CT reconstruction involves adapting denoising diffusion probabilistic models into a parallel framework. The reconstructed images exhibit competitive quality; furthermore, the radiation dose is reduced to half or one-third of the standard radiation dose.

Brain MRI and CT are performed within a short interval while planning radiation therapy. MRI provides better soft-tissue contrast, whereas CT is required for geometrical registration. Synthetic brain CT images generated from contrast-enhanced T1w MR image inputs using a GAN (Fig. 6) [62] exhibit robustness in preserving details and accurately presenting the abnormal anatomy. These results indicate the potential of using near-real-time MR-only treatment planning in the brain.

MRI and PET provide complementary structural and functional information on neurodegenerative diseases. A joint probability distribution of a diffusion model (JPDDM) was developed using the public Alzheimer's Disease Neuroimaging Initiative (ADNI) public dataset and validated to generate synthetic brain PET images from MRI inputs [75].

Notably, JPDDM exhibited a higher PSNR and a more accurate recovery of PET images from MRI compared with those of CycleGAN and score-based diffusion models. A bidirectional model using three-dimensional (3D) PET and MR images was developed [66] to reconstruct PET images. This framework effectively mapped the structural and functional information of the brain tissue, and the synthesized images that were almost identical to real images.

Reducing the Risk of Contrast Media

Iodinated contrast agents (CAs) can cause adverse effects such as iodine allergy and pose risks to patients with renal insufficiency or serious illnesses. CT angiography, a widely used vascular imaging technique, and low-dose CAs have been used to mitigate iodinated CA-related adverse effects. A GAN-based CT angiography model was developed by training GAN to synthesize non-contrast CT images using 1137 CT angiography images [67]. The visual quality of synthetic CT was comparable with that of real CT, with synthetic images exhibiting good diagnostic accuracy for vascular diseases in the internal (accuracy 94%) and external test sets (accuracy 86%).

Gadolinium deposition has been observed in the brain after repeated administration of GBCA. Therefore, the potential risks of GBCAs must be weighed against their clinical benefits and diagnostic value. The feasibility of synthetic GBCA images was tested using a conditional GAN (cGAN) that performed brain multimodal imaging of T1w, T2w, T2*w, diffusion weigted imaging (DWI), and arterial spin labeling (ASL) images to create contrast-enhanced T1w images. This GAN used two attention U-Net blocks as generators [76]. Synthetic images demonstrate contrast enhancement in the small vessels and lesions.
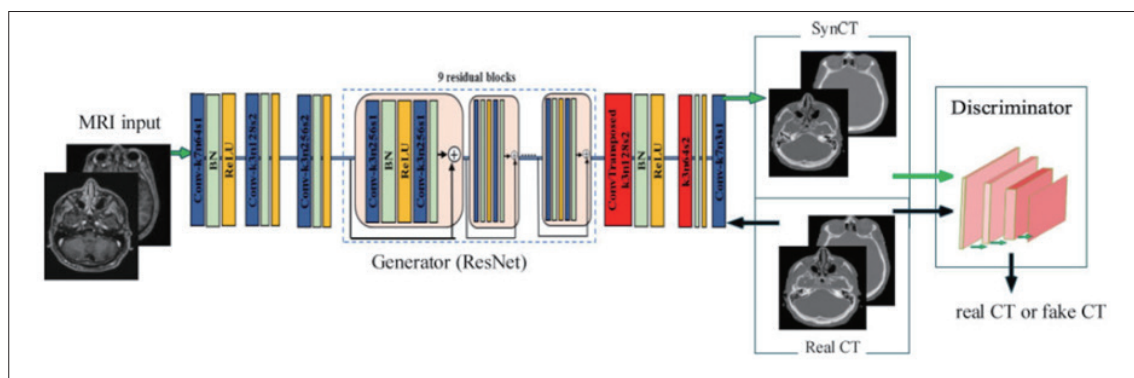
Fig. 6. Generated brain three-dimensional CT image from brain MRI using generative adversarial network. Simulation CT is mandatory for patients with brain metastasis undergoing radiosurgery. Generated brain CT can mitigate radiation dose risk. Adapted from Emami et al. *Med Phys* 2018;45:3627-3636, with permission of John Wiley and Sons [62]. CT = computed tomography, MRI = magnetic resonance imaging

Generation of Better-Quality Images

Challenges such as patient motion, prolonged imaging acquisition times, and increased radiation dose have posed difficulties in acquiring high-resolution CT and MRI images. Generative AI research aims to overcome the trade-off between spatial and temporal resolution [60,64,72,74,77]. High quality images can also be obtained by generating bone-suppressed images from the original images [59].

Application of super-resolution to low-resolution MR or CT images has demonstrated improvements in image quality and time efficiency. Consequently, this field has been combined with supervised learning (not image-generative AI), including CNN and U-Net. Hybrid models combining supervised and unsupervised learning have been studied [74]; for instance, a hybrid super-resolution reconstruction model based on GANs and U-net was developed by combining the frequency domain and perceptual loss functions in a previous study. Application to bladder MR, brain MR, and abdominal CT images resulted in better image quality, PSNR, and SSIM of the reconstructed images across the different datasets.

Score-based diffusion models [60,77] trained on public knee fast MRI data for MRI reconstruction tasks have demonstrated the feasibility of achieving higher resolution. These model used various sampling methods and body parts, not just training data, to generate images from the subsampling data, indicating practicality. The PSNR and SSIM of the synthetic images were superior to those of the images generated using other DL-based image reconstruction methods. Transfer learning was applied to pathology detection using YOLOv5 after image generation [77]. Notably, synthetic super-resolution images performed better than U-net-based synthetic images or ground-truth high-resolution images.

Super-resolution images were considered to depict small vascular structures. High-resolution (1 mm) brain time-of-flight (TOF) MR angiography (MRA) was generated using a GAN-based model [72] with low-resolution (4 mm) TOF-MRA image inputs. Synthetic MRA exhibited a significant improvement in terms of image quality. Moreover, the vessel visibility scale rated by radiologists was higher ($P < 0.001$) than that of low-resolution input images. Synthetic MRA images have exhibited sensitivities and specificities comparable with those of routine high-resolution MRA images in terms of diagnostic performance for lesion detection. Similarly, a GAN has been used to generate high-resolution synthetic images for motion-compensated isotropic 3D

coronary MRA (CMRA) [64]. Significant improvement in vessel sharpness has been observed with the use of high-resolution synthetic MRA ($34.1\% \pm 12.3\%$). Specifically, a 16-fold increase in spatial resolution (isotropic 0.9 mm^3 or 1.2 mm^3 from anisotropic 0.9 x 3.6 x 3.6 mm^3 or 1.2 x 4.8 x 4.8 mm^3) compared with that of low-resolution CMRA was observed. The reconstructed images exhibited a 16-fold increase in spatial resolution when using low-resolution CMRA, and the image quality was comparable with that obtained using high-resolution CMRA.

Clinically useful and higher-quality images can be obtained by subtracting the bone from CXR images, mimicking the dual-energy technique [59]. This study aimed to validate the effectiveness of GAN-based bone suppression and the dual energy technique in pulmonary nodule detection compared with that of standard CXRs and compare the performance of the two techniques. Compared with that of standard CXR, the area under the AFROC curve (AUAFROC) of GAN-based bone suppression was significantly higher for readers (GAN vs. standard, AUAFROC of 0.981 vs. 0.907 [reader 1] and 0.958 vs. 0.808 [reader 2], $P < 0.01$). Furthermore, this method demonstrated a performance comparable with that of the dual-energy technique.

Reduction of Inter-Vendor Variation by Style Transfer

Applying radiomics to DL models aids in diagnosis and prediction using noninvasive tissue phenotyping. However, radiomic features may not be reproducible if the imaging acquisition parameters, manufacturers, and reconstruction algorithms vary [78-82].

The potential utility of generative AI in enhancing reproducibility has been suggested through image conversion across different manufacturers, CT protocols, and reconstruction algorithms. A study [63] included CT images using scanners from four manufacturers, standard- or low-radiation doses, and sharp or medium kernels, and classified them into groups 1–7 according to acquisition conditions. CT images were converted into the target CT style (Group 1: standard dose and sharp kernel) using a RouteGAN. CT conversion improved the radiologists' scores for fibrosis, honeycombing, and reticulation and made the scores less variable compared with the original images.

Quantitative radiomics features of texture features in CXRs were compared between original images from two manufacturers and GAN converted images (Fig. 7) [68]. The reproducibility, evaluated using the concordance correlation coefficient (CCC), increased to 72.8% and 79.3%
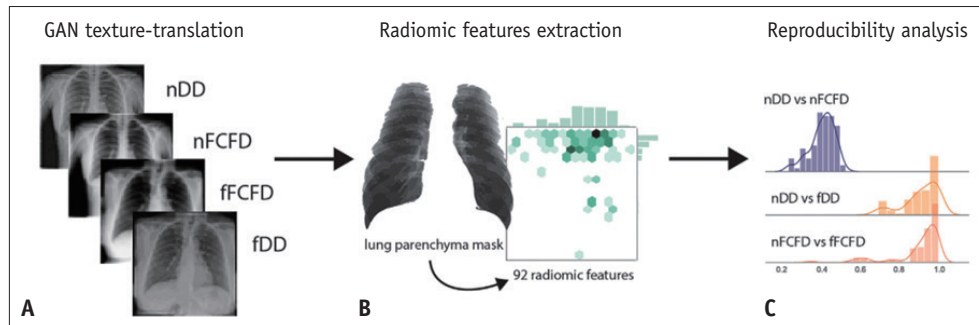
Fig. 7. Improved reproducibility of radiomics features obtained from chest radiographs. **A:** Style transfer among different manufacturers was performed to generate vendor-translated images using GAN. **B:** Radiomic features were extracted. **C:** The feature reproducibility improved significantly when calculating concordance correlation coefficient. Adapted from Marcadent et al. *Radiol Artif Intell* 2020;2:e190035, with permission of Radiological Society of North America [68]. Philips DD and Siemens FCFD. GAN = generative adversarial network, DD = DigitalDiagnost, FCFD = Fluorospot Compact FD, n = native, f = fake

(with a CCC threshold of 0.80) in each manufacturer after texture translation. Additionally, a classifier trained on translated CXRs showed increased accuracy, from 55.0% to 64.5%, in discriminating congestive heart failure (CHF) from non-CHF CXRs.

Another study developed an image conversion model to reproduce radiomic features across CT protocols and reconstruction kernels using an abdominal phantom with liver nodules [65]. For region of interest (ROI)-based analysis, they obtained 96 pairs of CCC from three categories of radiomics features, eight protocols, and 4 ROIs; similarly, for radiomic feature-based analysis, 6192 pairs of CCC were generated from 774 radiomics features and eight protocols. The results showed increased CCC in synthetic image pairs compared to the original images, specifically 83.3% (80/96) and 62.0% (3838/6192) in the ROI- and radiomic feature-based analysis, respectively.

Concerns With Direct Utilization of Generated/Synthetic Images and Potential Strategies to Mitigate Concerns

Image generative AI transforms images from one domain to another with "unpaired" images. This distinguishes image generative AI from CNN or U-net, which requires paired images for image translation. Therefore, in the clinical realm, caution is needed. The generated images may look like real images, but the specific images of an actual patient are not assumed to be similar to the generated images [83].

A recent survey of hallucinations in large foundation models [84] broadly classified them into four types: text, image, video, and audio. The presence of a sufficient number of positive pairs and ample variation mitigates hallucinations in image-generative AI. GANs can create artifacts or unnatural images [85]; notably, cycle-consistent
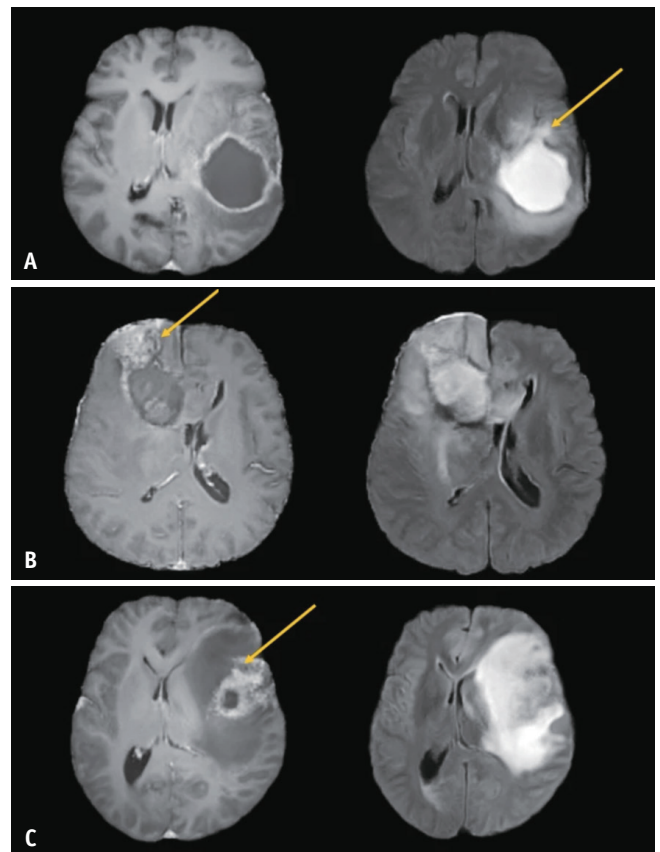
Fig. 8. Image hallucination and introduction of new features by generative adversarial network. **A:** T2/FLAIR image shows the open rim-like structure (arrow), which is semantically unnatural. **B:** Checkerboard-like artifact (arrow) is seen on the contrast-enhanced image. **C:** Bizarre enhancement (arrow) is seen at the anterior aspect.

GANs trained with unpaired data are particularly susceptible to these risks. Figure 8 presents the creation of unnatural images using the GAN. Sandfort et al. [83] reported the generation of unnatural images while generating virtual

non-contrast images from contrast-enhanced images using a cycleGAN. The output images synthesized from virtual non-contrast exhibited a metal clip that did not exist or a hyperdense aorta as if contrast media were injected, leading to the introduction of new features. Moreover, unlike "real" virtual non-contrast images, which enable physical/mathematical modeling of radiation absorption, synthesized images cannot be used as a tool for physically measuring tissues. Thus, it was concluded that image generation AI is not a magical tool but a sophisticated type of "style transfer." Synthetic non-contrast CT images can strengthen data augmentation methods, but are unsuitable for actual measurements or diagnostic purposes.

Diffusion models, in contrast, are a class of likelihood-based models that produce high-quality images [31,86-88] and possess advantages such as excellent distribution coverage, step-by-step training, and easy scalability. A comparative study between the GAN and diffusion models [31] revealed that the diffusion models exhibited reduced trade-offs between fidelity and diversity. However, it is unclear whether latent representations from diffusion models are semantically meaningful or have fewer false latent representations (new features). Nevertheless, the high fidelity and easy scalability of diffusion models reduces artifacts and prevents the creation of unnatural images (hallucinations).

Hallucinations are mitigated using the same rule for all generative AI models [89]. First, the input data quality is improved, and the models are trained on diverse and balanced datasets as the generative AI models feed on the data. Second, radiologists with domain expertise evaluate the image quality, semantic features of the disease, and the presence of artifacts obtained from generative images thereby incorporating human-in-the-loop validation. Third, rigorous testing, continuous iterations, and model refinement are performed to test the clinical scenarios.

Clinical and Research Usage: Approach 2–Generated Images to Augment Data for AI Training

Obtaining sufficient data for developing DL models is challenging owing to the less frequent occurrence of abnormal cases. A new dataset is generated by applying rotations, flipping, and translations to existing data to overcome data sparsity; however, this approach can result in overfitting. Generative AI can generate infinite images for use as completely new input data, thereby enhancing the model

performance in terms of lesion detection, segmentation, diagnosis, and classification. Table 4 summarizes the articles published using Approach 2 [43,70,83,90-96].

Improvement of Lesion Detection by Generating Pathologic Conditions or Pseudo-Healthy Images

GAN can be used to generate images of pathological areas or regions of interest. Lung nodules were synthesized on 3D chest CT images using a cGAN in a previous study [92]. The peripheral edge was smoothened to create realistic and natural-looking nodules, and the lung nodules were distributed from the center to the periphery. Notably, compared with that of previous conventional data augmentation, cGAN exhibited improved performance for detecting pathological lungs, particularly in the detection of peripheral lung nodules.

Similarly, the detection of brain metastases can be improved through lesion augmentation using generative AI. The detection algorithm with YOLOv3 exhibited a significant improvement in performance compared with that of training using real images only when the bounding box for detection and randomly shaped tumors were generated simultaneously using GAN [91]. The sensitivity improved from 67% to 77% (training with 2813 real images vs. +4000 GAN-based augmented images); however, the number of false positives per slice also increased from 4.11 to 7.65, which was considered clinically acceptable.

Several studies have demonstrated the utility of generative AI in generating synthetic pseudo-healthy images from diseased images, eventually generating anomaly maps based on the differences between the synthetic and generated images. Pseudo-healthy images were applied to brain MR sets and CXR images using denoising diffusion implicit models [96] in a previous study to improve tumor and pleural effusion detection, respectively. Detail-consistent image-to-image translation was performed without altering the architecture or training procedure to create a qualified anomaly map. Bercea et al. [90] created pseudo-healthy images from diseased input using an auto-de-noising diffusion probabilistic model trained on two public datasets of healthy brain MR images. Compared with classical diffusion models without explicit noise level tuning, the model applied to a public dataset for detecting ischemic infarct lesions demonstrated significantly higher robustness. The area under the precision-recall curve increased from 4.25 to 14.48, and the maximum Dice coefficient increased from 8.39 to 22.75.

Table 4. Generated images for augmentation of training data for imaging-based supervised deep learning tasks

Clinical utility	Authors	Network	Image data	Study purpose	Number of patients (training/validation/test set)	Evaluation methods
Enhancing detection by generating pathologic conditions	Jin et al. [92]	3D conditional GAN	Chest CT	1. Generate lung nodules on chest CT 2. Measure detection and segmentation performance of lung nodules	LIDC dataset 998/0/22	DSC and average surface distance
	Han et al. [91]	GAN	Brain MRI contrast-enhanced T1	1. Generate tumor and bounding box 2. Segment tumor	126/18/36	Mean average precision, sensitivity, and false positives per slice
Enhancing detection by generating pseudo-healthy images	Wolleb et al. [96]	Denoising diffusion implicit models	Chest X-ray/brain MRI T1, T2, FLAIR and contrast-enhanced T1	Generate pseudo-healthy chest X-ray and/or brain MRI and enable anomaly detection	CheXpert dataset (30955 images, split is unclear) BRATS dataset (16205 slices, split is unclear)	DSC and area under the receiver operating characteristics curve
	Bercea et al. [90]	DDPMs	Brain MRI T1	Generate pseudo-healthy brain MRI and enable anomaly detection	IXI 581 (split is unclear) Fast MRI 176 (131/15/30)	Segmentation area under the precision recall curve, and DSC
	Lee et al. [43]	Style GAN	Brain CT	1. Generate pseudo-healthy brain CT and enable anomaly detection 2. Implement anomaly detection system and evaluate clinical workflow efficiency	34085/271/273 (internal) and 1795 (external)	Area under the receiver operating characteristics curve, wait time, radiology turnaround time, and reading time
Enhancing segmentation	Rosnati et al. [95]	DDPMs	Chest X-ray/brain MRI T1, T2, FLAIR and contrast-enhanced T1	Improve organ segmentation of lung by adding generated chest X-ray and brain MRI, respectively	UKbiobank (34230/4280/4280) and BRATS dataset (269/36/33)	DSC
	Sandfort et al. [83]	Cycle GAN	Abdomen contrast-enhanced CT	Improve organ segmentation of the kidney, liver, and spleen by adding generated non-contrast CT images	Kidney NIH (50/3/13) liver dataDecathlon (179/9/43), Spleen dataDecathlon (30/2/8)	DSC and error volume of segmentation
	Preetha et al. [70]	Cycle GAN	Brain MRI T1, T2 and FLAIR	1. Generate contrast-enhanced T1 from pre-T1, T2, and FLAIR 2. Calculate volumetry and estimate association with time to progression	1540/NA/521	SSIM, MAE, DSC for volumetry, and hazard ratio

Table 4. Generated images for augmentation of training data for imaging-based supervised deep learning tasks (continued)

| Clinical utility	Authors	Network	Image data	Study purpose	Number of patients (training/validation/test set)	Evaluation methods
Improving diagnostic performance	Park et al. [94]	Style GAN	Brain MRI FLAIR and contrast-enhanced T1	Improve diagnostic performance of IDH-mutation high-grade glioma by generating IDH-mutant high-grade glioma images	110/0/44	Turing's test, area under the receiver operating characteristics curve, and 2 by 2 diagnostic accuracy test
	Moon et al. [93]	Score-based diffusion models	Brain MRI FLAIR and contrast-enhanced T1	Improve diagnostic performance of IDH-mutation by generating IDH-mutant and IDH-wild type phenotype images	565/86/119 (internal) and 108 (external)	Turing's test, area under the receiver operating characteristics curve, and 2 by 2 diagnostic accuracy test

GAN = generative adversarial network, CT = computed tomography, DSC = Dice similarity coefficient, MRI = magnetic resonance imaging, FLAIR = fluid-attenuated inversion recovery, DDPM = denoising diffusion probabilistic model, NA = not available, SSIM = structural similarity index measurement, MAE = mean absolute error, IDH = isocitrate dehydrogenase

The clinical implications were demonstrated through the automatic creation of an anomaly map. Pseudo-healthy images were generated using actual brain CT images of patients who visited the emergency room [43]. A style-based GAN was developed to create synthetic normal-looking images by training over 34085 brain CT images of healthy individuals. An anomaly detection algorithm (ADA) that determines critical findings based on the discrepancy between real and synthetic images was developed (Fig. 9). ADA achieved an area under the curve (AUC) of 0.85 and 0.87 in the internal and external validation datasets, respectively. In addition, ADA triage was performed to change the radiologists' cues for reading and wait times in the clinical simulation test. Notably, the turnaround and reading times were significantly reduced following the implementation of ADA.

Improvement of Segmentation

Several studies have demonstrated the potential utility of augmented data in enhancing segmentation tasks. An ensemble diffusion model was developed for semi-supervised image segmentation and domain generalization in one study [95]. The DSCs of the model trained on a public CXR image dataset for lung segmentation exhibited equivalent or significantly increased values: the mean DSCs for the fully supervised model, original label efficient diffusion model (LEDM), LEDM trained with diffusion steps (LEDMe), and ensemble diffusion model were 0.973, 0.970, 0.976, and 0.973, respectively, for the in-domain test set and 0.941, 0.944, 0.953, and 0.951, respectively, for the out-of-domain test set.

Sandfort et al. [83] studied the segmentation characteristics of the generated images in an abdominal CT study. Synthetic non-contrast CT images were generated from enhanced CT image inputs, and the segmentation performance of the model trained on real datasets was compared with that of the model trained on combined datasets. The combined dataset included the original contrast CT dataset (in-distribution) and a second dataset from a different hospital comprising only non-contrast CT images (out-of-distribution). The model for kidney segmentation trained using a synthetic augmented dataset exhibited a significant improvement in out-of-distribution performance compared with that of the model trained on contrast-enhanced images (with a Dice index of 0.09 to 0.66, $P < 0.001$). The improvements observed in liver and spleen segmentation for the model trained with generated images were smaller than those for the models trained with real
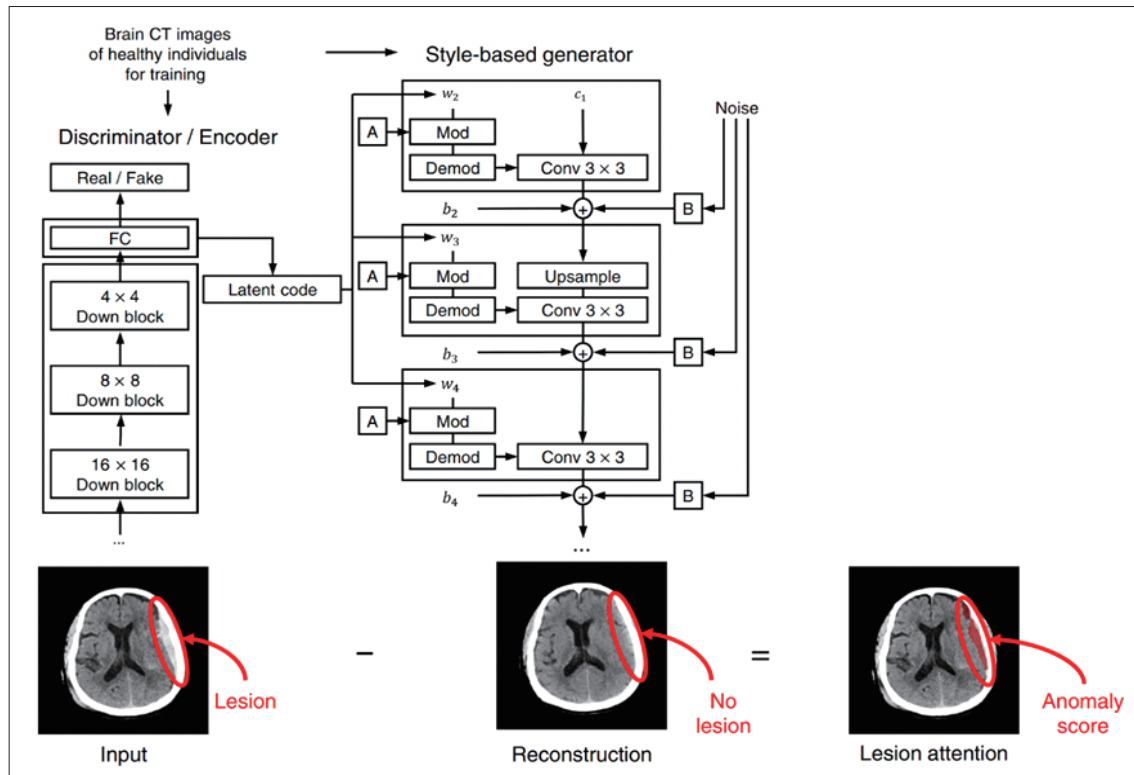
Fig. 9. Anomaly detection map and anomaly score of brain CT can be calculated using style-based generative adversarial network by generating pseudo-healthy CT images. Subtracting the real CT images from the generated pseudo-healthy CT images can create an anomaly detection map. Adapted from Lee et al. *Nat Commun* 2022;13:4251, with permission of Springer Nature BV [43]. CT = computed tomography, FC = fully connected layer

data, with the Dice index from 0.86 to 0.89 and 0.65 to 0.69, respectively.

A DL segmentation model for brain tumors was tested to further investigate the clinical application. Preetha et al. [70] developed a generative AI model to create synthetic contrast-enhanced T1w images from non-contrast T1w image inputs for enhanced segmentation and volume calculations. Contrast-enhanced T1w synthetic images were created using T1w, T2w, FLAIR, and ADC images. Notably, segmentation of the contrast-enhancing tumor from synthetic post-contrast T1w images yielded high CCC (0.782, 0.751–0.807, $P <$ 0.001) with true tumor volume; however, it yielded a slightly underestimated median tumor volume of -0.48 cm^3 (-0.37 to -0.76). Moreover, tumor response was evaluated via segmentation of contrast-enhancing tumors and volumetric assessment. The calculated hazard ratio for survival exhibited no significant difference when compared with real contrast-enhanced images.

Improvement of Diagnosis and Classification

Synthetic images with pathologic conditions/lesions can be utilized as an augmented dataset to develop statistical or DL-based diagnostic models that are particularly effective for rare diseases. In the diagnosis and prognosis prediction of brain gliomas, confirming the isocitrate dehydrogenase (IDH) mutation status is essential. Several studies have focused on using MR phenotyping to predict mutations non-invasively. However, gathering a sufficient number of patients is challenging owing to the rarity of this condition. The following studies generated MR images of IDH mutant or wild-type gliomas and used them for model development to address the insufficient data.

Park et al. [94] used the World Health Organization 2016 classification to generate synthetic images using GAN for IDH-mutant glioblastoma, a rare subtype. Diagnostic models for predicting IDH mutations have been developed using real IDH-wild-type and IDH-mutant glioblastomas, synthetic IDH-mutant glioblastomas, or both. The diagnostic accuracy of the synthetic image-augmented model was significantly higher than that of the real model. A multivariable diagnostic model using real and synthetic data exhibited slightly higher predictive performance than that of a model based solely on real images when applied to the test set (AUC of reader 1: 0.75 and 0.71, and AUC of reader 2: 0.82 and 0.77).
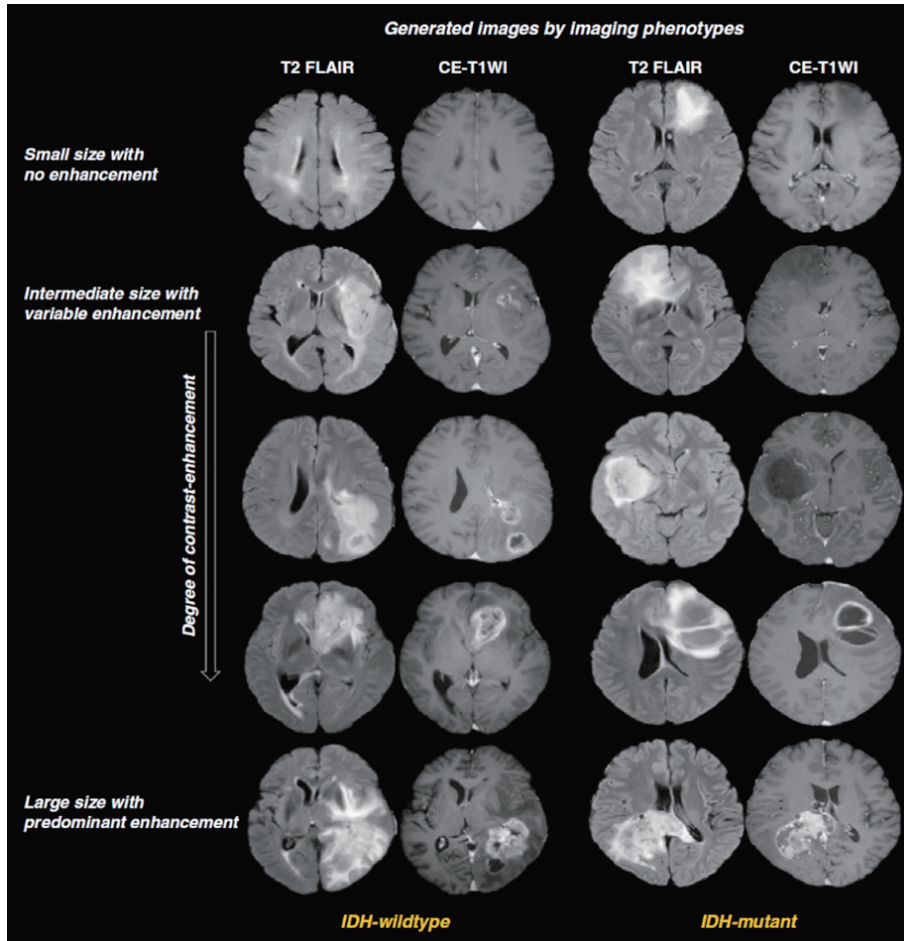
Fig. 10. Score-based diffusion model generates scalable imaging phenotypes of diffuse adult-type glioma in the brain. The size and contrast enhancement vary and are scalable in IDH-wild and IDH-mutant types, providing realistic and natural images. Adapted from Moon et al. *Neuro Oncol* 2024;26:1124-1135, with permission of Oxford University Press [93]. IDH = isocitrate dehydrogenase, FLAIR = fluid-attenuated inversion recovery, CE-T1WI = contrast-enhanced T1-weighted imaging

Moon et al. [93] used score-based diffusion models to generate synthetic images of IDH-mutant and wild-type gliomas. Notably, the diffusion models generated images by imaging phenotypes from small sizes with no enhancement to large sizes with predominant enhancement (Fig. 10). These synthetic images were augmented into a training dataset to develop a DL model for classifying the status of IDH mutation. Optimal augmentation was achieved with 110000 generated slices, with an AUC of 0.938. The augmented diagnostic model exhibited a significant improvement over neuroradiologists in classifying the types of IDH in internal and external test sets (AUC of the model: 0.94 and 0.83, AUC of reader 1: 0.86 and 0.82, and AUC of reader 2: 0.79 and 0.74 in the internal and external test sets, respectively).

Concerns With the Utilization of Generated Images as Augment Data for AI Training

Augmenting training data with generative AI can improve downstream AI models; however, several concerns remain. First, the quality of the models depends on the accuracy and diversity of the training data from the generative AI, which is crucial. The current metrics for evaluating the diversity of generative AI are limited, particularly in terms of diagnostics [46]. A recent study that tested the imaging feature diversity with human readers has proposed a possible method for assessing the diversity of the generated imaging phenotypes. Second, evidence for optimizing the number and ratio of generated images for downstream AI models is insufficient. Various ratios of real to synthetic images must be tested to identify optimal conditions for a broad range of medical applications [97]. Third, generative AI lacks transparency and the process of combining image-generative

AI with its training processes for downstream models further complicates this issue. The types of data used and their applications in the training of generative AI models remain unclear. Thus, it is recommended to test the quality and accuracy of the outputs from generative AI models and continuously monitor and adjust the performance of the downstream model to address this issue.

CONCLUSION

This article summarizes image-generative AI techniques employed in radiology. Categorizing clinical and research applications into two approaches has helped provide an overview of the myriad studies on generative AI techniques. The first approach involves the direct utilization of the generated images. The second approach involves the use of the generated images as training data for AI models to perform downstream supervised DL tasks. Image-generative AI produces synthetic data with a broad range of clinical and radiological potential by providing high-quality images, filling in missing sequences, creating images through less harmful processes, reducing inter-vendor variability, and data augmentation for training AI to perform various tasks, such as (anomaly) detection, segmentation, and classification. The evolving evaluation metrics of image-generative AI have become increasingly important in evaluating the fidelity, diversity, and semantic features of the generated images by humans. Image hallucinations and the introduction of new features must be carefully evaluated. Future DL studies should aim to adopt image-generative AI to improve a broad range of image-based tasks. This will help overcome concerns regarding small datasets in the medical field.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2024.0392.

Conflicts of Interest

Ji Eun Park and Namkug Kim, who hold respective positions as Editorial Board Member of the *Korean Journal of Radiology*, were not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Ji Eun Park, Namkug Kim. Funding acquisition: Ji Eun Park, Namkug Kim. Writing—original draft: Ha Kyung Jung, Kiduk Kim, Ji Eun Park. Writing—review & editing: Ji Eun Park, Namkug Kim.

ORCID IDs
Ha Kyung Jung
  https://orcid.org/0000-0003-4571-2688
Kiduk Kim
  https://orcid.org/0000-0002-9659-897X
Ji Eun Park
  https://orcid.org/0000-0002-4419-4682
Namkug Kim
  https://orcid.org/0000-0002-3438-2217

REFERENCES

1. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. *IEEE Signal Process Mag* 2018;35:53-65
2. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean J Radiol* 2024;25:224-242
3. You A, Kim JK, Ryu IH, Yoo TK. Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. *Eye Vis (Lond)* 2022;9:6
4. Skandarani Y, Lalande A, Afilalo J, Jodoin PM. Generative adversarial networks in cardiology. *Can J Cardiol* 2022;38:196-203
5. Qin Z, Liu Z, Zhu P, Xue Y. A GAN-based image synthesis method for skin lesion classification. *Comput Methods Programs Biomed* 2020;195:105568
6. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
7. Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa TL, et al. Foundation model for cancer imaging biomarkers. *Nat Mach Intell* 2024;6:354-367
8. Higgins I, Matthey L, Pal A, Burgess CP, Glorot X, Botvinick MM, et al. beta-VAE: learning basic visual concepts with a constrained variational framework [accessed on March 10, 2024]. Available at: https://api.semanticscholar.org/

CorpusID:46798026

9. Kingma DP. Auto-encoding variational bayes. arXiv [Preprint]. 2013 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.1312.6114

10. Van Den Oord A, Vinyals O. Neural discrete representation learning [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

11. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A. Conditional image generation with PixelCNN decoders [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html

12. Van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks [accessed on March 10, 2024]. Available at: https://proceedings.mlr.press/v48/oord16.html

13. Salimans T, Karpathy A, Chen X, Kingma DP. PixelCNN++: improving the pixelcnn with discretized logistic mixture likelihood and other modifications. arXiv [Preprint]. 2017 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.1701.05517

14. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

15. Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models [accessed on March 10, 2024]. Available at: https://proceedings.mlr.press/v139/nichol21a.html

16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

17. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv [Preprint]. 2017 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.1710.10196

18. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [accessed on March 10, 2024]. Available at: https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf

19. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN [accessed on March 10, 2024]. Available at: https://openaccess.thecvf.com/content_CVPR_2020/papers/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.pdf

20. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models [accessed on March 10, 2024]. Available at: https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf

21. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv [Preprint]. 2020 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2010.02502

22. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. arXiv [Preprint]. 2020 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2011.13456

23. Horvat C, Pfister JP. Denoising normalizing flow [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper/2021/hash/4c07fe24771249c343e70c32289c1192-Abstract.html

24. Papamakarios G, Nalisnick E, Rezende DJ, Mohamed S, Lakshminarayanan B. Normalizing flows for probabilistic modeling and inference. *J Mach Learn Res* 2021;22:1-64

25. Rezende D, Mohamed S. Variational inference with normalizing flows [accessed on March 10, 2024]. Available at: https://proceedings.mlr.press/v37/rezende15.pdf

26. Zhang Q, Chen Y. Diffusion normalizing flow [accessed on March 10, 2024]. Available at: https://proceedings.neurips.cc/paper/2021/file/876f1f9954de0aa402d91bb988d12cd4-Paper.pdf

27. Du Y, Li S, Tenenbaum J, Mordatch I. Learning iterative reasoning through energy minimization [accessed on March 10, 2024]. Available at: https://proceedings.mlr.press/v162/du22d/du22d.pdf

28. Liu N, Li S, Du Y, Tenenbaum JB, Torralba A. Learning to compose visual relations [accessed on March 10, 2024]. Available at: https://dl.acm.org/doi/10.5555/3540261.3542035

29. Xie J, Lu Y, Zhu SC, Wu Y. A theory of generative ConvNet [accessed on March 10, 2024]. Available at: https://proceedings.mlr.press/v48/xiec16.html

30. Xie J, Zhu SC, Wu YN. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE Trans Pattern Anal Mach Intell* 2021;43:516-531

31. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis [accessed on March 10, 2024]. Available at: https://proceedings.nips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

32. Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv [Preprint]. 2016 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.1611.02163

33. Thanh-Tung H, Tran T. Catastrophic forgetting and mode collapse in GANs [accessed on March 10, 2024]. Available at: https://doi.org/10.1109/IJCNN48605.2020.9207181

34. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data [accessed on March 10, 2024]. Available at: https://papers.nips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf

35. Wang Z, Zheng H, He P, Chen W, Zhou M. Diffusion-GAN:

training GANs with diffusion. arXiv [Preprint]. 2022 [accessed on March 10, 2024]. Available at: https://doi.org/10.48550/arXiv.2206.02262

36. Hong GS, Jang M, Kyung S, Cho K, Jeong J, Lee GY, et al. Overcoming the challenges in the development and implementation of artificial intelligence in radiology: a comprehensive review of solutions beyond supervised learning. *Korean J Radiol* 2023;24:1061-1080

37. Moon JH, Lee H, Shin W, Kim YH, Choi E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J Biomed Health Inform* 2022;26:6070-6080

38. Tumanyan N, Geyer M, Bagon S, Dekel T. Plug-and-play diffusion features for text-driven image-to-image translation [accessed on April 2, 2024]. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Tumanyan_Plug-and-Play_Diffusion_Features_for_Text-Driven_Image-to-Image_Translation_CVPR_2023_paper.html

39. Lee H, Kang M, Han B. Conditional score guidance for text-driven image-to-image translation [accessed on March 10, 2024]. Available at: https://dl.acm.org/doi/10.5555/3666122.3667801

40. Yang Q, Li N, Zhao Z, Fan X, Chang EI, Xu Y. MRI cross-modality image-to-image translation. *Sci Rep* 2020;10:3753

41. Wang Z, Yang Y, Chen Y, Yuan T, Sermesant M, Delingette H, et al. Mutual information guided diffusion for zero-shot cross-modality medical image translation. *IEEE Trans Med Imaging* 2024;43:2825-2838

42. Wang K, Chen Z, Zhu M, Li Z, Weng J, Gu T. Score-based counterfactual generation for interpretable medical image classification and lesion localization. *IEEE Trans Med Imaging* 2024 [Epub]. https://doi.org/10.1109/TMI.2024.3375357

43. Lee S, Jeong B, Kim M, Jang R, Paik W, Kang J, et al. Emergency triage of brain computed tomography via anomaly detection with a deep generative model. *Nat Commun* 2022;13:4251

44. Geman D, Geman S, Hallonquist N, Younes L. Visual turing test for computer vision systems [accessed on March 14, 2024]. Available at: https://doi.org/10.1073/pnas.1422953112

45. Borji A. Pros and cons of GAN evaluation measures. *Comput Vis Image Understand* 2019;179:41-65

46. Borji A. Pros and cons of GAN evaluation measures: new developments. *Comput Vis Image Understand* 2022;215:103329

47. Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. *Electron Lett* 2008;44:800-801

48. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 2004;13:600-612

49. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs [accessed on March 14, 2024]. Available at: https://proceedings.neurips.cc/paper_files/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html

50. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S.

GANs trained by a two time-scale update rule converge to a local nash equilibrium [accessed on March 14, 2024]. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

51. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision [accessed on March 14, 2024]. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html

52. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;22:79-86

53. Sajjadi MS, Bachem O, Lucic M, Bousquet O, Gelly S. Assessing generative models via precision and recall [accessed on March 14, 2024]. Available at: https://dl.acm.org/doi/10.5555/3327345.3327429

54. Sculley D. Web-scale k-means clustering [accessed on March 14, 2024]. Available at: https://doi.org/10.1145/1772690.1772862

55. Naeem MF, Oh SJ, Uh Y, Choi Y, Yoo J. Reliable fidelity and diversity metrics for generative models [accessed on March 14, 2024]. Available at: https://proceedings.mlr.press/v119/naeem20a.html

56. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31

57. Faghani S, Khosravi B, Zhang K, Moassefi M, Jagtap JM, Nugen F, et al. Mitigating bias in radiology machine learning: 3. Performance metrics. *Radiol Artif Intell* 2022;4:e220061

58. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. *Radiol Artif Intell* 2021;3:e200126

59. Bae K, Oh DY, Yun ID, Jeon KN. Bone suppression on chest radiographs for pulmonary nodule detection: comparison between a generative adversarial network and dual-energy subtraction. *Korean J Radiol* 2022;23:139-149

60. Chung H, Ye JC. Score-based diffusion models for accelerated MRI. *Med Image Anal* 2022;80:102479

61. Conte GM, Weston AD, Vogelsang DC, Philbrick KA, Cai JC, Barbera M, et al. Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model. *Radiology* 2021;299:313-323

62. Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med Phys* 2018;45:3627-3636

63. Hwang HJ, Kim H, Seo JB, Ye JC, Oh G, Lee SM, et al. Generative adversarial network-based image conversion among different computed tomography protocols and vendors: effects on accuracy and variability in quantifying regional disease patterns of interstitial lung disease. *Korean J Radiol* 2023;24:807-820

64. Küstner T, Munoz C, Psenicny A, Bustin A, Fuin N, Qi H, et al. Deep-learning based super-resolution for 3D isotropic coronary MR angiography in less than a minute. *Magn Reson Med*

2021;86:2837-2852

65. Lee SB, Cho YJ, Hong Y, Jeong D, Lee J, Kim SH, et al. Deep learning-based image conversion improves the reproducibility of computed tomography radiomics features: a phantom study. *Invest Radiol* 2022;57:308-317

66. Lin W, Lin W, Chen G, Zhang H, Gao Q, Huang Y, et al. Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease. *Front Neurosci* 2021;15:646013

67. Lyu J, Fu Y, Yang M, Xiong Y, Duan Q, Duan C, et al. Generative adversarial network-based noncontrast CT angiography for aorta and carotid arteries. *Radiology* 2023;309:e230681

68. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative adversarial networks improve the reproducibility and discriminative power of radiomic features. *Radiol Artif Intell* 2020;2:e190035

69. Ozbey M, Dalmaz O, Dar SUH, Bedel HA, Ozturk S, Gungor A, et al. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans Med Imaging* 2023;42:3524-3539

70. Preetha CJ, Meredig H, Brugnara G, Mahmutoglu MA, Foltyn M, Isensee F, et al. Deep-learning-based synthesis of post-contrast T1-weighted MRI for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study. *Lancet Digit Health* 2021;3:e784-e794

71. Schlaeger S, Li HB, Baum T, Zimmer C, Moosbauer J, Byas S, et al. Longitudinal assessment of multiple sclerosis lesion load with synthetic magnetic resonance imaging-a multicenter validation study. *Invest Radiol* 2023;58:320-326

72. Wicaksono KP, Fujimoto K, Fushimi Y, Sakata A, Okuchi S, Hinoda T, et al. Super-resolution application of generative adversarial network on brain time-of-flight MR angiography: image quality and diagnostic utility evaluation. *Eur Radiol* 2023;33:936-946

73. Xia W, Niu C, Cong W, Wang G. Cube-based 3D denoising diffusion probabilistic model for cone beam computed tomography reconstruction with incomplete data. arXiv [Preprint]. 2023 [accessed on March 20, 2024]. Available at: https://arxiv.org/abs/2303.12861v1

74. Xiao Y, Chen C, Wang L, Yu J, Fu X, Zou Y, et al. A novel hybrid generative adversarial network for CT and MRI super-resolution reconstruction. *Phys Med Biol* 2023;68:135007

75. Xie T, Cao C, Cui Z, Li F, Wei Z, Zhu Y, et al. Brain PET synthesis from MRI using joint probability distribution of diffusion model at ultrahigh fields. arXiv [Preprint]. 2022 [accessed on March 17, 2024]. Available at: https://doi.org/10.48550/arXiv.2211.08901

76. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks [accessed on March 17, 2024]. Available at: https://doi.org/10.48550/arXiv.1611.07004

77. Cui ZX, Cao C, Liu S, Zhu Q, Cheng J, Wang H, et al. Self-score: self-supervised learning on score-based models for MRI reconstruction. arXiv [Preprint]. 2022 [accessed on March 16, 2024]. Available at: https://doi.org/10.48550/

arXiv.2209.00835

78. Choe J, Lee SM, Do KH, Lee G, Lee JG, Lee SM, et al. Deep learning-based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology* 2019;292:365-373

79. Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One* 2016;11:e0164924

80. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol* 2015;50:757-765

81. Meyer M, Ronald J, Vernuccio F, Nelson RC, Ramirez-Giraldo JC, Solomon J, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. *Radiology* 2019;293:583-591

82. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys* 2017;44:1050-1062

83. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci Rep* 2019;9:16884

84. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv [Preprint]. 2023 [accessed on March 20, 2024]. Available at: https://doi.org/10.48550/arXiv.2309.05922

85. Wolterink JM, Mukhopadhyay A, Leiner T, Vogl TJ, Bucher AM, Išgum I. Generative adversarial networks: a primer for radiologists. *Radiographics* 2021;41:840-857

86. Choi J, Kim S, Jeong Y, Gwon Y, Yoon S. ILVR: conditioning method for denoising diffusion probabilistic models. arXiv [Preprint]. 2021 [accessed on April 1, 2024]. Available at: https://doi.org/10.48550/arXiv.2108.02938

87. Zhu J, Shen Y, Zhao D, Zhou B. *In-domain GAN inversion for real image editing*. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. *Computer vision–ECCV 2020*. Cham: Springer, 2020:592-608

88. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks [accessed on March 26, 2024]. Available at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.pdf

89. Shrivastav A. Generative AI hallucinations: revealing best techniques to minimize hallucinations [accessed on April 9, 2024]. Available at: https://www.kellton.com/kellton-tech-blog/generative-ai-hallucinations-revealing-best-techniques

90. Bercea CI, Neumayr M, Rueckert D, Schnabel JA. Mask, stitch, and re-sample: enhancing robustness and generalizability in anomaly detection through automatic diffusion models. arXiv [Preprint]. 2023 [accessed on March 20, 2024]. Available at: https://doi.org/10.48550/arXiv.2305.19643

91. Han C, Murao K, Noguchi T, Kawata Y, Uchiyama F, Rundo L, et al. Learning more with less: conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images [accessed on March 20, 2024]. Available at: https://dl.acm.org/doi/10.1145/3357384.3357890

92. Jin D, Xu Z, Tang Y, Harrison AP, Mollura DJ. *CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation.* In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, eds. *Medical image computing and computer assisted intervention–MICCAI 2018.* Cham: Springer, 2018:732-740

93. Moon HH, Jeong J, Park JE, Kim N, Choi C, Kim YH, et al. Generative AI in glioma: ensuring diversity in training image phenotypes to improve diagnostic performance for IDH mutation prediction. *Neuro Oncol* 2024;26:1124-1135

94. Park JE, Eun D, Kim HS, Lee DH, Jang RW, Kim N. Generative adversarial network for glioblastoma ensures morphologic variations and improves diagnostic model for isocitrate dehydrogenase mutant type. *Sci Rep* 2021;11:9912

95. Rosnati M, Roschewitz M, Glocker B. Robust semi-supervised segmentation with timestep ensembling diffusion models [accessed on March 20, 2024]. Available at: https://proceedings.mlr.press/v225/rosnati23a/rosnati23a.pdf

96. Wolleb J, Bieder F, Sandkühler R, Cattin PC. *Diffusion models for medical anomaly detection.* In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. *Medical image computing and computer assisted intervention–MICCAI 2022.* Cham: Springer, 2022:35-45

97. Chen D, Han Y, Duncan J, Jia L, Shan J. Generative artificial intelligence enhancements for reducing image-based training data requirements. *Ophthalmol Sci* 2024;4:100531