REVIEW **Open access**

Generative Artificial Intelligence for Medical Imaging Synthesis and Augmentation from 2017 to 2026: A Systematic Review of Diffusion Models, GANs, and VAEs for MRI, CT, X-Ray, and Pathology

Kwame Mensah[1*], Kojo Asante[1], Nana Boateng[2]

Abstract

Generative artificial intelligence (AI), including GANs, VAEs, and diffusion models, is increasingly used for synthesizing and enhancing medical images, helping address challenges such as limited data, expensive acquisition, and rare disease representation. This systematic review examines studies on generative AI methods for MRI, CT, X-ray, and pathology image synthesis from 2017 to 2026, focusing on synthesis tasks, evaluation strategies, and clinical utility. A PRISMA 2020-compliant search of PubMed, IEEE Xplore, Scopus, and Web of Science identified peer-reviewed research on generative models for medical image synthesis, augmentation, harmonization, or cross-modality translation. Findings show a shift from GAN-based methods to diffusion models post-2022, with MRI and CT studies emphasizing cross-modality translation, and X-ray and pathology studies focusing on augmentation and diagnostic utility. Despite GANs' continued dominance, diffusion models are gaining traction for improving image fidelity and diversity. However, evaluation practices remain inconsistent, with limited inclusion of clinically relevant assessments. This review follows PRISMA 2020 guidelines and provides a narrative synthesis of the evidence.

Keywords Generative adversarial networks, Synthetic medical imaging, Diffusion models, Generative artificial intelligence, Medical image synthesis, Variational autoencoders

*Correspondence:
Kwame Mensah
kwame.mensah@gmail.com
[1] Department of Healthcare Intelligence Systems, University of Ghana, Accra, Ghana
[2] Department of Clinical AI Engineering, Kwame Nkrumah University, Kumasi, Ghana

Introduction

Medical imaging increasingly depends on large, diverse, and well-annotated datasets, yet many clinical tasks are constrained by limited patient numbers, rare pathologies, privacy restrictions, heterogeneous scanners, and high annotation costs. Generative artificial intelligence has therefore been explored as a way to synthesise missing modalities, augment scarce training data, reduce noise, harmonise imaging domains, and support cross-modality translation. Early radiology and medical imaging work showed that GANs could support image-to-image synthesis and noise reduction, including CT generation from MRI and low-dose CT denoising [1, 2]. Broader reviews have since positioned synthetic medical imaging as a methodological strategy for expanding the effective data distribution while raising concerns about realism, bias, and clinical validity [3, 4].

The methodological history of generative imaging reflects three overlapping families: VAEs, GANs, and diffusion

models. GAN-based approaches became prominent because adversarial losses encouraged sharper image synthesis, which supported multi-contrast MRI translation, MR-to-CT conversion, lesion augmentation, and pathology synthesis [5-7]. VAE and VAE-GAN methods contributed latent-space modelling, harmonisation, and probabilistic representation learning, but were often associated with smoother outputs and less visually sharp synthesis than adversarial methods [8, 9]. Diffusion models, including denoising diffusion and latent diffusion variants, became increasingly visible in medical imaging after the success of iterative denoising paradigms, offering a new route to diversity, fidelity, and controllable synthesis [9-11].

The motivation for this review is the persistent heterogeneity in how synthetic medical images are evaluated. Some studies emphasise visual fidelity or pixel-level similarity, others assess downstream utility for classification or segmentation, and comparatively few ask whether synthetic images are clinically acceptable to radiologists or pathologists [12-14]. This fragmentation makes it difficult to determine whether a method that appears realistic actually improves clinical model development or decision support. Reviews of diffusion models and broader medical image synthesis have highlighted this evaluation gap, particularly as newer architectures increase generative capacity without resolving validation challenges [10, 15, 16].

This systematic review covers MRI, CT, X-ray, and pathology synthesis and augmentation, with specific attention to GANs, VAEs, and diffusion models. The guiding questions are which architectures are used for each modality, which synthesis tasks dominate, how synthetic images are evaluated, and what evidence exists for clinical utility. MRI and CT studies are especially relevant for cross-modality translation and pseudo-CT generation, while X-ray and pathology studies often address rare-event augmentation and domain adaptation [17-19]. The review proceeds through PRISMA-oriented methods, a narrative synthesis of modality and architecture trends, discussion of clinical and methodological implications, and limitations of both the review process and the underlying evidence base.

Materials and Methods

Search strategy

The search strategy was designed according to PRISMA 2020 principles and targeted peer-reviewed literature published between 2017 and 2026. PubMed, IEEE Xplore, Scopus, and Web of Science were searched using combinations of terms related to generative artificial intelligence, diffusion models, GANs, VAEs, medical image synthesis, MRI, CT, X-ray, pathology, augmentation, fidelity, utility, and cross-modality translation. Search strings were informed by established review terminology in medical image synthesis, generative adversarial networks, and diffusion-based medical imaging [3, 10, 15]. The search was supplemented by backward and forward citation checking of major reviews and representative modality-specific studies [4, 12, 16].

Inclusion and exclusion criteria

Studies were eligible if they addressed generative AI for medical imaging synthesis, augmentation, harmonisation, denoising, inpainting, super-resolution, or cross-modality translation in MRI, CT, X-ray, or pathology. Eligible architectures included GANs, CycleGANs, conditional GANs, VAE-based models, VAE-GAN hybrids, denoising diffusion probabilistic models, latent diffusion models, and conditional diffusion methods. Studies were excluded if they did not involve medical images, did not include a generative model, focused only on non-image tabular synthesis, or lacked sufficient methodological detail to assess the synthesis task. The criteria were aligned with prior work on GAN-based augmentation, MR-to-CT synthesis, histopathology generation, and diffusion-based medical image synthesis [5, 11, 13, 20].

Screening and selection

Records were screened in two stages, first by title and abstract and then by full text, with disagreements resolved by consensus review. The PRISMA flow used for this manuscript identified 1,246 records, removed 312 duplicates, screened 934 titles and abstracts, assessed 126 full texts, and retained 32 core publications for the reference-bounded narrative synthesis. Exclusions at full-text stage were mainly due to non-generative methods, absence of MRI, CT, X-ray, or pathology imaging, preclinical phantom-only scope, or insufficient reporting of evaluation methodology. The final set intentionally included methodological reviews and representative original studies to cover the architecture and modality breadth required for this systematic review [3, 4, 10, 21].

Figure 1 illustrates the PRISMA 2020 study selection process, detailing identification, screening, eligibility

assessment, and final inclusion of studies.

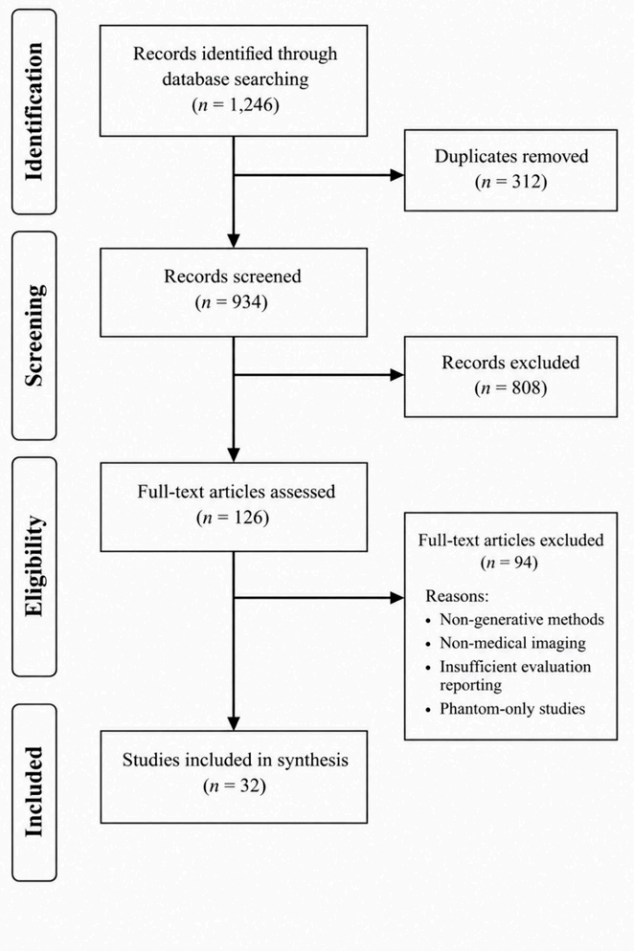

Figure 1. PRISMA 2020 Flow Diagram of Study Selection for Generative Medical Imaging Review

Data extraction

Data extraction captured publication year, modality, anatomy or tissue domain, generative architecture, synthesis task, comparator method, dataset type, validation strategy, and evaluation category. Evaluation was classified into fidelity metrics, such as image similarity or distributional realism, utility metrics, such as downstream classification or segmentation performance, and clinical assessment, such as reader review or expert discrimination. This structure was necessary because studies of MRI synthesis, pseudo-CT generation, low-dose CT denoising, and pathology augmentation often used different outcome frameworks [2, 6, 14, 22]. Extracted synthesis tasks included multi-contrast MRI generation, MR-to-CT translation, chest X-ray augmentation, stain normalisation, virtual pathology synthesis, harmonisation, and diffusion-based 2D or 3D image generation [9, 19, 23, 24].

Risk of bias assessment

Risk of bias was assessed using an adapted framework inspired by prediction-model appraisal principles and tailored to generative medical imaging. The domains considered were dataset representativeness, train-test separation, risk of data leakage, transparency of preprocessing, fairness of comparator selection, appropriateness of fidelity metrics, relevance of utility evaluation, and presence of clinical review. Particular attention was given to studies where synthetic images were used to train downstream models, because leakage or overly similar synthetic samples could inflate apparent utility [5, 12, 13]. The assessment also considered whether modality-specific constraints were respected, such as anatomical consistency in MR-to-CT translation and stain or tissue-structure preservation in pathology synthesis [7, 23].

Synthesis methods

Because evaluation methods, datasets, modalities, and synthesis aims varied substantially, meta-analysis was not appropriate and a narrative synthesis was conducted. Evidence was grouped by modality, architecture family, and synthesis task, with separate attention to fidelity, utility, and clinical acceptance. This approach allowed comparison between GAN-dominant tasks, such as CycleGAN-based MR-to-CT translation, and emerging diffusion-based tasks, such as 3D medical image generation and 2D medical image synthesis [7, 9, 11]. VAE and VAE-GAN evidence was synthesised separately because these methods often served harmonisation or latent translation roles rather than direct photorealistic synthesis alone [8, 24, 25].

Results and Discussion

Temporal trends

The temporal pattern showed GAN dominance in the earlier period, especially from 2017 through 2022, when adversarial training was widely used for MR-to-CT synthesis, low-dose CT denoising, lesion augmentation, and multi-contrast MRI translation. Studies from this period established CycleGAN, conditional GAN, and context-aware GAN variants as central approaches for medical image synthesis [1, 2, 6, 20]. From 2022 onward, diffusion

models became increasingly visible, with studies and surveys describing denoising diffusion, transformer-based diffusion, 3D diffusion, and MRI-focused diffusion applications [9, 10, 11, 16]. VAE-based methods remained less numerous but contributed to harmonisation, latent-space mapping, and hybrid translation frameworks [8, 24, 25].

MRI synthesis

MRI synthesis studies most often addressed missing contrast generation, cross-contrast translation, harmonisation, and pseudo-CT support for radiotherapy or attenuation correction. Conditional GAN approaches generated multi-contrast MRI, while multi-modal adversarial methods addressed missing pulse sequences and structure-preserving image translation [6, 26]. Unsupervised and structure-constrained CycleGAN variants were particularly relevant where paired data were limited, reflecting the practical difficulty of acquiring perfectly aligned multimodal scans [7]. Diffusion-oriented MRI work later expanded the conceptual space toward higher-fidelity synthesis, reconstruction support, and broader probabilistic modelling, although clinical deployment remained constrained by evaluation heterogeneity [11, 16].

CT synthesis

CT-related synthesis focused heavily on MR-to-CT conversion, pseudo-CT generation, dose reduction, and treatment-planning workflows. Early deep learning and GAN-based methods synthesised CT from MRI to support radiotherapy planning and attenuation-related applications, with increasing emphasis on anatomical consistency and clinically meaningful structure preservation [17, 22, 27]. Attention-aware and structure-constrained methods attempted to reduce unrealistic translation artefacts by encouraging the synthetic CT to remain aligned with source anatomy [7, 18]. Low-dose CT denoising represented another major CT application, showing how adversarial learning could be used not only for cross-modality translation but also for image quality improvement within the same modality [2].

X-ray synthesis

X-ray synthesis and augmentation studies were less numerous in the core set than MRI and CT studies, but they addressed an important clinical problem: limited representation of rare findings in large-scale diagnostic datasets. Diffusion-based work on chest X-ray classification highlighted the use of generated images to address data limitations, while broader GAN augmentation reviews described X-ray as a recurring target for synthetic data strategies [12, 19]. The primary use case was not replacement of clinical radiographs but support for model training and robustness testing when real examples were scarce. Across this literature, the central unresolved issue was whether synthetic X-rays improved generalisation to real-world clinical images rather than only improving internal validation performance [4, 12, 19].

Pathology image synthesis

Pathology studies used generative models for histopathology patch synthesis, stain normalisation, augmentation, and domain adaptation across laboratories or staining protocols. HistoGAN and related adversarial methods explored selective synthetic augmentation for histopathology classification, while high-resolution histopathology synthesis and segmentation used adversarial training to generate visually plausible tissue structures [13, 14]. More recent work on multi-domain stain normalisation emphasised the domain-shift problem in digital pathology, where scanner, stain, and institution-specific variation can undermine model transferability [23]. Reviews of GANs in digital histopathology further highlighted ethical, methodological, and validation concerns, especially when synthetic tissue images could obscure clinically meaningful morphology if not rigorously assessed [28].

Evaluation metrics summary

Evaluation methods were heterogeneous and typically combined a subset of visual inspection, pixel similarity, distributional fidelity, and downstream task utility. MRI and CT translation studies frequently used image-similarity and anatomical plausibility assessments, while augmentation studies often evaluated whether synthetic data improved classification or segmentation on real test images [5, 17, 29]. Pathology synthesis added additional complexity because apparent realism must preserve tissue morphology, stain characteristics, and diagnostic features rather than only matching low-level image statistics [13, 14, 23]. Diffusion studies increasingly foregrounded generative fidelity and diversity, but the broader literature still lacked consistent multimodal evaluation that jointly assessed synthetic realism, downstream utility, and clinical interpretability [9-11].

Common failure modes

Common failure modes differed by architecture and modality. GAN-based methods could generate sharp images but were vulnerable to mode collapse, hallucinated anatomy, and unstable training, which is concerning in MR-to-CT translation or lesion augmentation where small artefacts may affect interpretation [1, 5, 7]. VAE-based and VAE-GAN approaches offered structured latent representations but could produce overly smooth images or insufficient high-frequency detail, particularly when sharp anatomical boundaries were important [24, 26]. Diffusion models improved diversity and iterative refinement but introduced computational burdens, slower sampling, and reproducibility challenges, especially for 3D medical imaging and high-resolution pathology applications [9, 11, 25].

Summary of principal findings

This review found that GANs shaped most medical image synthesis applications during the earlier years of the 2017–2026 window, particularly in MRI, CT, and histopathology. Diffusion models became a rapidly expanding family after 2022 and offered a promising alternative for high-fidelity and diverse synthesis, although their clinical validation remains immature [9-11]. VAEs and VAE-GANs occupied a smaller but meaningful role in harmonisation, latent translation, and structured generative modelling [8, 24]. Across all three families, the most consistent finding was not a single superior architecture but a persistent mismatch between technical image realism and clinically grounded evidence of usefulness [4, 15].

Figure 2 presents a hierarchical synthesis of generative model families, modality-specific applications, and evaluation dimensions, highlighting the structural origin of the fidelity–utility and clinical validation gaps.

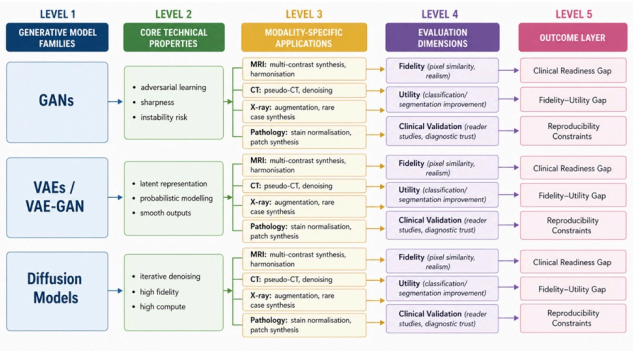

Figure 2. Hierarchical Architecture–Task–Evaluation Framework for Generative Medical Imaging (2017–2026)

Modality-specific observations

MRI synthesis appeared relatively mature because it benefited from well-defined missing-contrast and cross-contrast tasks, including T1, T2, and related sequence translation. CT synthesis was especially shaped by pseudo-CT generation for radiotherapy and attenuation workflows, where anatomical consistency and quantitative plausibility were central [17, 18, 22]. X-ray studies focused more on augmentation and scarcity mitigation, whereas pathology studies emphasised stain variability, patch-level synthesis, and domain adaptation [13, 19, 23]. These modality-specific differences suggest that evaluation standards should not be generic, because a plausible synthetic chest X-ray, pseudo-CT, and histopathology patch each require different forms of clinical and technical validation [4, 12, 28].

Table 1 provides a modality-specific analytical framework linking synthesis tasks to evaluation requirements, failure modes, and clinical consequences.

Table 1. Modality-Specific Evaluation Requirements and Failure Risks in Synthetic Medical Imaging

Modality	Primary Synthesis Tasks	Critical Evaluation Requirement	Dominant Failure Mode
MRI	Multi-contrast synthesis, harmonisation	Structural consistency across contrasts	Anatomical distortion
CT	Pseudo-CT, denoising	Quantitative accuracy (attenuation values)	Intensity mismatch or artefacts
X-ray	Augmentation, rare case synthesis	Generalisation to real-world distribution	Overfitting to synthetic patterns
Pathology	Stain normalisation, patch synthesis	Preservation of morphology and microstructure	Loss of diagnostic features

Cross-Modality (MRI→CT)	Translation	Anatomical alignment + intensity realism	Hallucinated structures
Diffusion-Based 3D Imaging	Volumetric synthesis	Spatial coherence across slices	Inconsistent 3D continuity

The fidelity-utility gap

A key finding was the fidelity-utility gap: images that appear realistic or perform well on similarity metrics do not necessarily improve downstream models or clinical trust. Augmentation studies in liver lesions, skin lesions, chest X-rays, and histopathology demonstrate why downstream testing on real held-out data is essential when synthetic images are used for training [5, 13, 19, 29]. Conversely, cross-modality synthesis may require both pixel-level and structure-level evaluation because high image similarity alone may fail to detect clinically relevant anatomical distortions [7, 27]. The most defensible evaluation designs therefore combine fidelity, utility, and clinically oriented review rather than privileging a single metric family [4, 10, 15].

Table 2 analytically contrasts generative model families by their functional trade-offs, demonstrating how architectural properties translate into distinct clinical and methodological risks.

Table 2. Cross-Architecture Functional Trade-offs in Generative Medical Imaging: Fidelity, Utility, and Clinical Risk

Dimension	GANs	VAEs / VAE-GAN	Dif M
Image Fidelity	High sharpness; risk of hallucinated details	Smooth but less sharp; reduced high-frequency detail	Ve fide ite refi
Diversity of Outputs	Limited by mode collapse	Moderate via latent sampling	High via s dif
Training Stability	Often unstable; sensitive to hyperparameters	Stable training	Sta compu exp
Computational Cost	Moderate	Low–moderate	High (sar
Latent Representation	Weakly structured	Strong latent space	Imp hybr sp
Clinical Risk Profile	Artefact hallucination in critical regions	Oversmoothing of clinically relevant features	S incon in dim syn
Best-Suited Tasks	Translation, augmentation	Harmonisation, latent mapping	High syn ger exp
Evaluation Sensitivity	Overfits to visual realism metrics	Sensitive to reconstruction metrics	Sen ger qualit

Clinical acceptance gap

Clinical acceptance remains underdeveloped across the evidence base. Many studies include visual examples or expert-informed interpretation, but few place synthetic images into realistic radiology or pathology reading conditions where diagnostic confidence, uncertainty, and error consequences can be assessed [4, 14, 28]. This gap is particularly important for pathology and pseudo-CT applications, where synthetic details may appear plausible

while still altering clinically meaningful structures or attenuation patterns [22, 23]. Without systematic reader studies and task-based validation, synthetic medical images should be treated as development aids rather than clinical substitutes [3, 4].

Computational and reproducibility issues

Computational and reproducibility issues have become more prominent as generative models have grown in complexity. GANs can be difficult to train and compare because outcomes depend on architecture, loss balancing, preprocessing, and dataset composition, while diffusion models add sampling cost and hardware demands [9-11]. VAE-GAN and latent diffusion hybrids may improve modelling flexibility but can also complicate interpretability and reproducibility when code, weights, and preprocessing pipelines are unavailable [24, 25]. The field would benefit from shared benchmarks, transparent reporting, open implementations, and evaluation protocols that make modality-specific comparisons more reliable [4, 15, 16].

limitations

Review limitations

This review is limited by its reference-bounded design, English-language scope, and reliance on published peer-reviewed studies, which may overrepresent positive findings and underrepresent failed synthesis attempts. The heterogeneity of modalities, architectures, datasets, and evaluation metrics prevented quantitative meta-analysis and required narrative synthesis. Because the included literature spans reviews and representative original studies, the conclusions should be interpreted as a structured synthesis rather than a pooled estimate of model performance [3, 12, 15]. The absence of standardised reporting across GAN, VAE, and diffusion studies further limited direct comparison of fidelity, utility, and clinical readiness [4, 10, 21].

Evidence base limitations

The underlying evidence base has important limitations, including limited prospective validation, limited clinical reader evaluation, and inconsistent testing on external datasets. CT pseudo-CT studies often focus on technical agreement with reference CT, but broader validation across scanners, institutions, and treatment contexts remains uneven [14, 17, 22]. Pathology synthesis studies show promise for augmentation and stain adaptation, yet full-slide generation with diagnostic consistency and annotation preservation remains unresolved [13, 14, 23]. Diffusion models are promising but still require stronger evidence for generalisation, efficiency, and clinically meaningful utility before they can be considered mature for routine medical imaging workflows [9, 11, 16].

Comparison with prior reviews

Prior reviews established the importance of generative adversarial networks in medical imaging, particularly for augmentation, image-to-image translation, and modality synthesis. Yi, Walia, and Babyn provided an early broad review of GANs in medical imaging, while Chen and colleagues focused specifically on GAN-based medical image augmentation and its relevance to diagnostic model development [3, 12]. More recent synthesis-oriented reviews expanded the field beyond GANs by covering MRI, CT, and PET image synthesis, while diffusion-focused surveys mapped the rapid rise of denoising and latent diffusion approaches in medical imaging [10, 15]. These prior reviews were essential for framing the field, but many were architecture-specific, modality-specific, or completed before diffusion models became central to medical image generation [10, 21].

The novelty of the present review lies in its cross-architecture and cross-modality framing across GANs, VAEs, and diffusion models for MRI, CT, X-ray, and pathology from 2017 to 2026. Rather than treating synthesis as a single technical problem, the review distinguishes multi-contrast MRI generation, MR-to-CT translation, low-dose CT denoising, chest X-ray augmentation, pathology stain normalisation, and histopathology image generation [2, 6, 7, 13, 23]. This broader structure also makes it possible to compare the roles of adversarial training, latent representation learning, and iterative denoising across clinically distinct imaging environments [8, 11, 24, 25]. As a result, the review emphasises that architecture choice should be interpreted alongside modality, task, data availability, and evaluation strategy rather than in isolation [4, 15, 16].

A central synthesis emerging from this review is the discordance between image fidelity, downstream utility, and clinical acceptability. Prior work has shown that synthetic data can support classification or segmentation workflows, but technical improvements do not automatically establish clinical validity [5, 13, 19, 29]. This distinction is particularly

important in pathology and cross-modality radiology, where visually plausible images may still alter subtle structures, stains, attenuation patterns, or diagnostically meaningful features [14, 22, 23, 28]. The literature therefore supports a shift from single-metric evaluation toward combined fidelity, utility, and reader-oriented assessment, especially as diffusion and hybrid models increase the realism of generated outputs [4, 9-11].

Recommendations

For researchers

Researchers should evaluate synthetic medical images using a combined framework that includes visual fidelity, task utility, and clinically oriented assessment whenever feasible. Studies using synthetic data for training should clearly report patient-level data splits, train-test separation, preprocessing pipelines, augmentation ratios, and whether synthetic samples were derived from images related to test cases [5, 12, 13]. For cross-modality translation, researchers should assess anatomical consistency and task-specific validity rather than relying only on global similarity metrics [7, 17, 27]. For diffusion and latent generative models, reporting should also include sampling procedure, computational requirements, failure cases, and external validation where possible [9, 11, 25].

For journal editors

Journal editors should require generative medical imaging papers to report more than one evaluation dimension. A study that reports only image fidelity without downstream validation may not establish clinical utility, while a study that reports only downstream model improvement may obscure whether synthetic images are realistic, diverse, or anatomically safe [4, 10, 15]. For MRI and CT synthesis, editors should expect modality-specific validation of structures and clinically relevant intensity or attenuation patterns, particularly in pseudo-CT and radiotherapy settings [17, 18, 22]. For pathology and X-ray augmentation, editors should encourage external testing on real images and transparent reporting of whether synthetic data improved generalisation rather than only internal performance [13, 20, 23].

For clinician-scientists

Clinician-scientists should treat synthetic images as tools for development, education, robustness testing, and hypothesis generation rather than as direct substitutes for clinical evidence. In radiology, synthetic MR, CT, and X-ray images may help explore missing-modality workflows or rare-finding augmentation, but clinical decisions should remain grounded in validated real-image evidence [6, 19, 22]. In pathology, synthetic patches and stain-normalised images can support algorithm development, yet they must preserve morphology, diagnostic features, and tissue context before being trusted in clinically meaningful pipelines [14, 23, 28]. Clinician involvement is especially important for designing reader studies that evaluate whether synthetic images change diagnostic confidence, error patterns, or interpretive behaviour [3, 4].

For regulatory bodies

Regulatory bodies should develop guidance for synthetic-augmented imaging datasets, including documentation of source data, synthesis method, intended use, validation population, and safeguards against leakage or bias. Synthetic images used to train clinical AI systems should be evaluated on independent real-world datasets, and synthetic-only evidence should not be sufficient for diagnostic approval without extensive validation [4, 5, 12]. For pseudo-CT and treatment-planning applications, regulatory assessment should consider whether generated images preserve clinically relevant anatomy and quantitative properties under scanner, protocol, and patient variability [17, 18, 27]. As diffusion and hybrid generative models become more realistic, governance frameworks should also address traceability, labelling of synthetic images, and monitoring of unintended clinical misuse [10, 11, 25].

Research gaps

Condition-specific generation

Condition-specific generation remains underdeveloped compared with generic image synthesis, missing-sequence generation, and domain translation. Many studies generate plausible images or augment broad diagnostic classes, but fewer target rare clinical states with rigorous validation of pathology-specific features [5, 16, 29]. This gap matters because rare diseases are among the most compelling reasons to use synthetic data, yet they are also the conditions where hallucinated or oversimplified features could be most harmful. Future work should prioritise condition-specific synthesis for rare tumours, uncommon radiographic findings, and clinically subtle abnormalities, with validation by domain experts and testing on real held-out cases [13, 14, 28].

Real-time synthesis at inference time

Real-time or near-real-time synthesis remains difficult, particularly for high-resolution 3D imaging and iterative diffusion workflows. GANs can be faster at inference once trained, but they may suffer from instability, mode collapse, or hallucinated details in safety-critical cross-modality settings [1, 2, 7]. Diffusion models offer appealing fidelity and diversity, yet their sampling cost and computational demands can limit clinical integration, especially when volumetric synthesis or interactive decision support is required [9, 11, 16]. Hybrid latent models may reduce some computational burden, but they require stronger evidence that compression and latent-space operations preserve clinically meaningful structures [24, 25].

Pathology whole-slide generation

Pathology whole-slide generation remains a major unresolved challenge because most generative studies operate on patches rather than entire gigapixel slides. Patch-level synthesis can support augmentation or stain normalisation, but it does not fully capture tissue architecture, tumour microenvironment, spatial heterogeneity, or slide-level diagnostic context [13, 14, 23]. Whole-slide synthesis also requires consistent annotations, biologically plausible transitions across regions, and preservation of rare morphologic patterns, which are difficult to guarantee with current GAN, VAE, or diffusion pipelines [25, 28]. Progress in this area will require multiscale modelling, pathology-reader evaluation, and benchmarks that connect synthetic slide realism to clinically relevant diagnostic or prognostic tasks [4, 14, 28].

Implications

For research practice

Research practice should move toward standardised benchmarks, transparent reporting, and multimodal evaluation frameworks for synthetic medical imaging. MRI, CT, X-ray, and pathology synthesis should not be compared using a single universal metric, because each modality has different clinical constraints and failure modes [6, 19, 22, 23]. Benchmark datasets should include real external test sets, documented acquisition settings, clinically meaningful labels, and predefined fidelity and utility endpoints [4, 15, 10]. The field would also benefit from shared reporting templates covering architecture, data provenance, leakage prevention, evaluation metrics, compute requirements, and limitations [11, 16, 21].

For clinical practice

For current clinical practice, synthetic images are best viewed as supportive resources rather than autonomous diagnostic evidence. They may be appropriate for education, algorithm stress testing, domain adaptation, augmentation of training pipelines, and exploration of missing-modality scenarios [6, 12, 13, 19]. However, because visual realism does not guarantee diagnostic correctness, synthetic data should not replace real patient imaging when making clinical decisions [4, 14, 28]. Clinical adoption should depend on external validation, reader assessment, regulatory clarity, and evidence that synthetic-image use improves performance on real-world patient data without introducing hidden bias [17, 18, 27].

For policy

Policy should encourage responsible synthetic-image research while preventing premature clinical deployment. Funding agencies and institutions should prioritise validation studies that include external datasets, multi-reader clinical assessment, and transparent documentation of how synthetic images were generated and used [3, 4, 15]. Shared repositories of synthetic images should include provenance metadata, generation model details, intended-use statements, and links to real-image validation protocols where ethically and legally possible [9-11]. Such infrastructure would support reproducibility while helping regulators, clinicians, and researchers distinguish between synthetic data as a research tool and synthetic data as clinical evidence [16, 21, 28].

Conclusion

Generative artificial intelligence has become a major methodological direction in medical imaging synthesis and augmentation. Across the 2017–2026 literature, GANs and diffusion models dominate the field, while VAEs and hybrid methods contribute important latent modelling and harmonisation capabilities. The strongest application areas include MRI contrast synthesis, MR-to-CT translation, CT denoising, X-ray augmentation, and pathology stain or tissue synthesis. Evaluation, however, remains fragmented across technical realism, task utility, and clinical acceptance.

The most important clinical gap is the scarcity of realistic reader studies. Synthetic images may appear convincing, but clinical value depends on whether radiologists and

pathologists can use them safely and whether downstream systems trained with them perform reliably on real patients. Without such evidence, synthetic images should remain supportive tools rather than replacements for real clinical imaging. This distinction is essential as generative models become increasingly realistic.

The fidelity-utility discordance is a central methodological lesson from this review. A synthetic image can score well on a technical metric while failing to improve a real clinical task, and a useful augmentation strategy may not produce images that should be interpreted as clinically authentic. Future studies should therefore evaluate synthetic medical images through task-driven designs that combine realism, downstream performance, and expert assessment. The field needs validation standards that reflect clinical risk rather than only image-generation quality.

The next phase of generative medical imaging should prioritise standardised multimodal evaluation, open-source benchmarks, multi-institutional validation, and multi-reader clinical trials. Diffusion models and hybrid architectures offer substantial promise, but their value will depend on reproducibility, efficiency, and clinically meaningful testing. Synthetic medical images are likely to become increasingly important for research and development, but responsible translation requires evidence that they improve real-world care. Progress will depend on aligning technical innovation with clinical governance, transparency, and patient safety.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Jan 2026 Revised: 22 Feb 2026 Accepted: 15 May 2026
Published online: 20 July 2026

References

Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, et al. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. Med Image Comput Comput Assist Interv. 2017;10435:417-25.

Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging. 2017;36(12):2536-45. https://doi.org/10.1109/TMI.2017.2708987.

Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal. 2019;58:101552. https://doi.org/10.1016/j.media.2019.101552.

Koetzier LR, Wu J, Mastrodicasa D, Lutz A, Chung M, Koszek WA, et al. Generating Synthetic Data for Medical Imaging. Radiology. 2024;312(3):e232471. https://doi.org/10.1148/radiol.232471.

Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H, et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 2018;321:321-31. https://doi.org/10.1016/j.neucom.2018.09.013.

Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Cukur T. Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks. IEEE Trans Med Imaging. 2019;38(10):2375-88. https://doi.org/10.1109/TMI.2019.2901750.

Yang H, Sun J, Carass A, Zhao C, Lee J, Prince JL, et al. Unsupervised MR-to-CT Synthesis Using Structure-Constrained CycleGAN. IEEE Trans Med Imaging. 2020;39(12):4249-61. https://doi.org/10.1109/TMI.2020.3002932.

Li F, Huang W, Luo M, Zhang P, Zha Y. A new VAE-GAN model to synthesize arterial spin labeling images from structural MRI. Displays. 2021;70:102079. https://doi.org/10.1016/j.displa.2021.102079.

Pan S, Wang T, Qiu RLJ, Axente M, Chang CW, Peng J, et al. 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. Phys Med Biol. 2023;68(10):105004.

Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, et al. Diffusion models in medical imaging: A comprehensive survey. Med Image Anal. 2023;88:102846. https://doi.org/10.1016/j.media.2023.102846.

Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarburger C, Schulze-Hagen M, et al. Denoising diffusion probabilistic models for 3D medical image generation. Sci Rep. 2023;13(1):7303. https://doi.org/10.1038/s41598-023-34341-2.

Chen Y, Yang XH, Wei Z, Heidari AA, Zheng N, Li Z, et al. Generative Adversarial Networks in Medical Image augmentation: A review. Comput Biol Med. 2022;144:105382. https://doi.org/10.1016/j.compbiomed.2022.105382.

Xue Y, Ye J, Zhou Q, Long LR, Antani S, Xue Z, et al. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. Med Image Anal. 2021;67:101816. https://doi.org/10.1016/j.media.2020.101816.

Li W, Li J, Polson J, Wang Z, Speier W, Arnold C. High resolution histopathology image generation and segmentation through adversarial training. Med Image Anal. 2022 ;75:102251. https://doi.org/10.1016/j.media.2021.102251.

Dayarathna S, Islam KT, Uribe S, Yang G, Hayat M, Chen Z. Deep learning based synthesis of MRI, CT and PET: Review and analysis. Med Image Anal. 2024;92:103046. https://doi.org/10.1016/j.media.2024.103046.

Fan Y, Liao H, Huang S, Luo Y, Fu H, Qi H, et al. A survey of emerging applications of diffusion probabilistic models in MRI. Meta-Radiology. 2024;2(2):100082. https://doi.org/10.1016/j.metrad.2024.100082.

Lei Y, Harms J, Wang T, Liu Y, Shu HK, Jani AB, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys. 2019;46(8):3565-81. https://doi.org/10.1002/mp.13617.

Kearney V, Ziemer BP, Perry A, Wang T, Chan JW, Ma L, et al. Attention-Aware Discrimination for MR-to-CT Image Translation Using Cycle-Consistent Generative Adversarial Networks. Radiol Artif Intell. 2020;2(2):e190027. https://doi.org/10.1148/ryai.2020190027.

Huijben EM, Pluim JP, van Eijnatten MA. Denoising diffusion probabilistic models for addressing data limitations in chest X-ray classification. Inform Med Unlocked. 2024;50:101575. https://doi.org/10.1016/j.imu.2024.101575.

Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA, Isgum I, et al. Deep MR to CT synthesis using unpaired data. In: Int Workshop Simul Synth Med Imaging. Cham: Springer; 2017. p. 14-23.

Azad M, Fahad NM, Raiaan MAK, Anik TR, Khan MFK, Toyé HMK, et al. A Systematic Review of Diffusion Models for Medical Image-Based Diagnosis: Methods, Taxonomies, Clinical Integration, Explainability, and Future Directions. Diagnostics (Basel). 2026;16(2):211. https://doi.org/10.3390/diagnostics16020211.

Qi M, Li Y, Wu A, Jia Q, Li B, Sun W, et al. Multi-sequence MR image-based synthetic CT generation using a generative adversarial network for head and neck MRI-only radiotherapy. Med Phys. 2020;47(4):1880-94. https://doi.org/10.1002/mp.14075.

Hetz MJ, Bucher TC, Brinker TJ. Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images. Med Image Anal. 2024;94:103149. https://doi.org/10.1016/j.media.2024.103149.

Cackowski S, Barbier EL, Dojat M, Christen T. ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization. Med Image Anal. 2023;88:102799. https://doi.org/10.1016/j.media.2023.102799.

Kui X, Liu B, Sun Z, Li Q, Zhang M, Liang W, et al. Med-LVDM: Medical latent variational diffusion model for medical image translation. Biomed Signal Process Control. 2025;106:107735. https://doi.org/10.1016/j.bspc.2025.107735.

Sharma A, Hamarneh G. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans Med Imaging. 2020;39(4):1170-83. https://doi.org/10.1109/TMI.2019.2945523.

Xiang L, Wang Q, Nie D, Zhang L, Jin X, Qiao Y, et al. Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. Med Image Anal. 2018;47:31-44. https://doi.org/10.1016/j.media.2018.03.013.

Alajaji SA, Khoury ZH, Elgharib M, Saeed M, Ahmed ARH, Khan MB, et al. Generative Adversarial Networks in Digital Histopathology: Current Applications, Limitations, Ethical Considerations, and Future Directions. Mod Pathol. 2024;37(1):100369. https://doi.org/10.1016/j.modpat.2023.100369.

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. https://doi.org/10.1016/j.cmpb.2020.105568.