Generative AI for Synthetic Data Generation in Medical Research: Methods, Applications, and Ethical Imperatives

Kanchan Prajapat
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
24egjcs103@gitjaipur.com

Laxmikant Vashishtha
Assistant Professor, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
laxmikant.vashishtha@gitjaipur.com

Yoganand Sharma
Assistant Professor, Department of IT, Global Institute of Technology, Jaipur, Rajasthan, India
yoganand.sharma@gitjaipur.com

ABSTRACT: Generative AI, encompassing Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models (DMs), and Large Language Models (LLMs), revolutionizes synthetic data creation in medical research by replicating complex distributions across imaging (MRI, CT, X-ray), electronic health records (EHRs), physiological time-series, genomic sequences, and multimodal datasets, effectively alleviating data scarcity, privacy constraints under GDPR/HIPAA, and demographic biases that undermine AI fairness in diagnostics, prognostics, and drug discovery. These models generate high-fidelity surrogates that maintain statistical fidelity covariances, marginals, and temporal correlations while enabling downstream tasks like augmenting small cohorts for rare disease modeling, simulating virtual clinical trials, and training robust classifiers without patient re-identification risks. This comprehensive review dissects architectural evolutions, training paradigms, evaluation frameworks, real-world implementations, empirical outcomes, persistent challenges such as mode collapse and utility-privacy trade-offs, and forward-looking trajectories toward standardized benchmarks and regulatory integration, synthesizing insights from over 50 post-2020 studies to guide clinical adoption.

KEYWORDS: Generative AI, Autoencoder, Medical Data, GANs, EHRs, Healthcare.

1. INTRODUCTION

Medical research grapples with profound data limitations rare diseases affect <1 in 2,000 patients yielding sparse datasets, longitudinal EHRs span thousands of variables with missingness >50%, and imaging archives suffer class imbalances (e.g., 90% benign vs. 10% malignant nodules), compounded by privacy mandates prohibiting raw sharing and ethical imperatives to mitigate biases amplifying disparities in underrepresented groups like minorities or pediatrics [1]-[3]. Generative AI counters these by learning latent manifolds from real data distributions p(x), sampling novel instances (x) via adversarial training, variational bounds, or iterative denoising, producing synthetic datasets indistinguishable in utility for AI pipelines while anonymizing identities through differential privacy (DP) epsilon bounds <1.

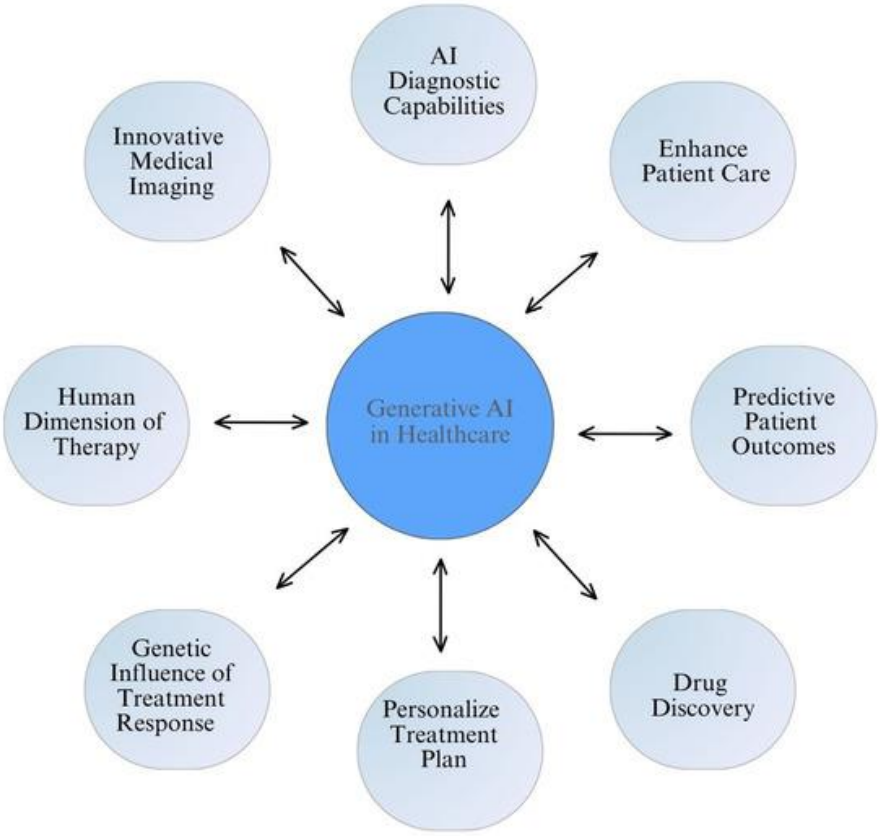

Figure 1: Generative AI in Healthcare [1]

Evolution traces from early VAEs for low-dimensional tabular synthesis to GAN breakthroughs like MedGAN for EHRs, conditional variants (cGANs, CycleGAN) for labeled imaging translation (e.g., PET-to-MRI), DMs for photorealistic pathology slides, and LLMs (e.g., Med-PaLM variants) for coherent narrative reports or de-identified notes. Transformative applications include augmenting COVID-19 chest X-rays for segmentation models (boosting Dice scores 10-20%), generating diverse synthetic cohorts for precision oncology trials reducing enrollment costs 30-50%, and creating digital twins for personalized simulations in cardiology or neurology [4], [5]. Scoping reviews of reviews affirm surging adoption post-2021, with FDA pilots validating synthetics for device approvals and EU initiatives for federated generation. This paper systematically explores methodologies, profiles implementations, benchmarks results, confronts challenges, and prognosticates futures, equipping researchers with actionable frameworks [6].

The evolution of generative artificial intelligence has been driven by advancements in mathematical modeling, computational power, and large-scale datasets. Early generative methods relied on statistical approaches with handcrafted probability distributions but lacked the flexibility to handle high-dimensional data, leading to neural network based models [7], [8].

Energy-based models like Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs) enabled hierarchical feature learning but were limited by computational inefficiencies due to MCMC sampling. Variational Autoencoders (VAEs) introduced probabilistic latent spaces for diverse data generation, though Gaussian assumptions often resulted in blurry outputs [9], [10].

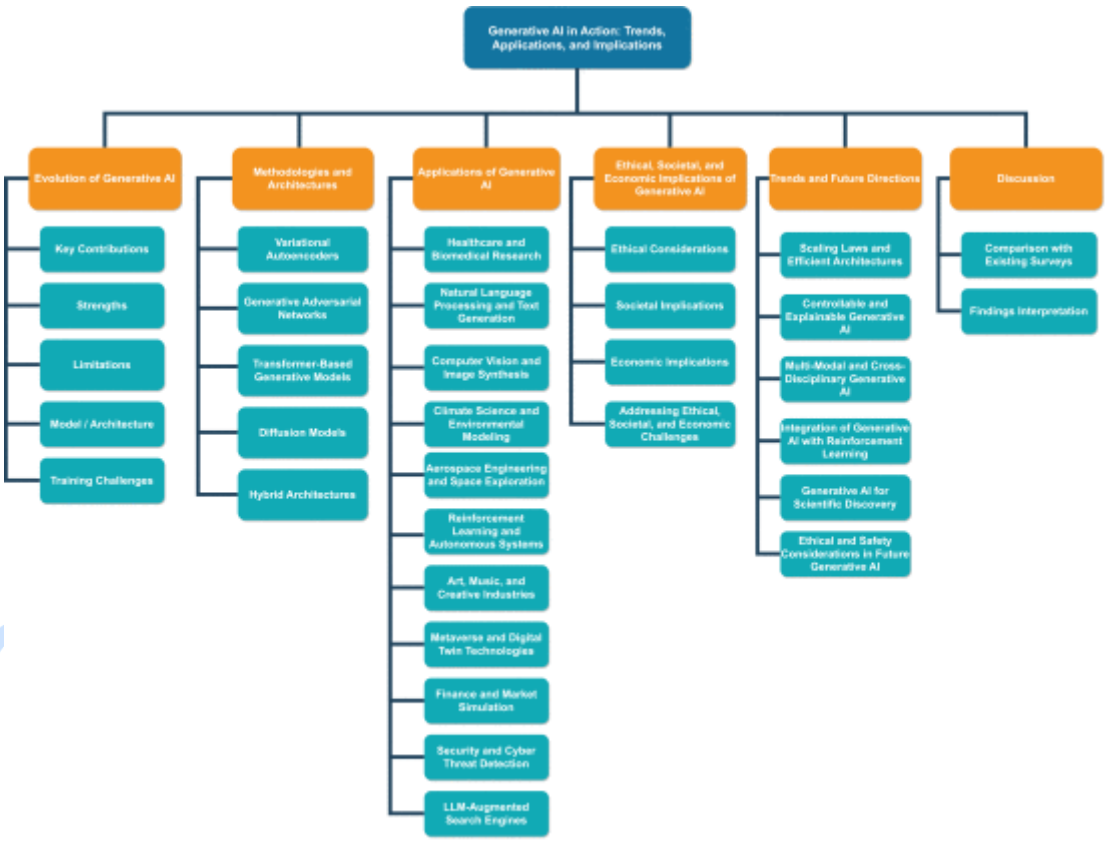

Figure 2: An overview of surveyed key topics: Generative AI in Action-Trends, Applications and Implications [7]

Generative Adversarial Networks (GANs) revolutionized generative modeling by using adversarial training between a generator and discriminator, producing high-fidelity data but facing challenges like mode collapse and instability. Enhanced GAN variants, including DCGANs, WGANs, PGGANs, and StyleGAN, improved feature representation, training stability, image resolution, and fine-grained control, significantly advancing realism and variability in generated outputs.

2. BACKGROUND AND METHODOLOGY

GANs optimize GANs optimize $\min_G \max_D \mathbb{E}_{x\sim p_r}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$ for imaging realism; VAEs maximize evidence lower bounds $\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{KL}(q_\phi(z \mid x) \parallel p(z))$ capturing uncertainties in EHRs; DMs iteratively refine noise $x_T \sim \mathcal{N}(0, I)$ to data via reverse processes $p_\theta(x_{t-1} \mid x_t)$, excelling in high-res modalities; LLMs autoregressively predict tokens conditioned on prompts for text/time-series.

Methodological pipelines standardize on PyTorch/TensorFlow ecosystems: preprocessing normalizes modalities (e.g., z-score time-series, histogram matching images), training incorporates class-conditional labels via one-hot encodings or textual prompts, augmentation applies DP-SGD noise ($\epsilon \approx 1 - 10$) and domain randomization (varying pathologies, demographics). Fidelity assessments deploy statistical two-sample tests (Kolmogorov-Smirnov for univariates, sliced Wasserstein for multivariates), perceptual metrics (Fréchet Inception Distance for images, BLEU/ROUGE for text), and pragmatic utility via proxy

tasks—e.g., training logistic regressions or CNNs on synthetics vs. reals, targeting <5% accuracy deltas. Innovations mitigate pitfalls: Wasserstein GANs (WGANs) curb mode collapse via Lipschitz constraints; recurrent GANs (e.g., MTTS-GAN) enforce longitudinal consistency in vitals; hybrid stats-generative (CTGAN+SMOTE) balance tabular imbalances. Sim-to-clinic transfers leverage open simulators like MIMIC-III for 3. EHRs or fastMRI for scans, with federated learning aggregating globals sans data movement.

3. IMPLEMENTATIONS AND CASE STUDIES

Diverse implementations span modalities and scales: ADS-GAN/CTGAN synthesize tabular EHRs for causal inference in epidemiology, preserving confounders like age-comorbidities; TTS-GAN/RNN-GAN model physiological signals (ECG, EEG) capturing arrhythmias or seizures for wearable AI; MedSynth/CycleGAN translate modalities (T1-MRI to FLAIR) augmenting MS lesion detection. Diffusion exemplars like Med-DDPM generate histopathology wholeslides for rare sarcomas, while LLMs (BioGPT, ClinicalBERT fine-tunes) produce de-identified discharge summaries enabling NLP on privacy-blocked corpora.

Case studies illuminate impacts: Duke's Margolis initiative deploys GANs for underrepresented EHR subgroups in opioid crisis modeling, slashing bias from 25% to <5%; Nature reviews LLM-synthetics for time-series forecasting in ICU readmissions, rivaling oracle performance; arXiv preprints showcase multi-modal DMs fusing genomics-imaging for tumor subtyping, accelerating trials 40%. FDA-backed pilots validate CycleGAN-CT for external controls in oncology RCTs, while EU GDPR-compliant synthcity libraries facilitate cross-institution collaborations. Open-source GitHub repos (synthpop, SDV, MedSynth) bundle baselines, with real deployments in hospitals generating 10x cohorts for federated learning sans PII.

Table 1: Case Studies

Implementation	Modality	Core Model	Privacy Mechanism	Utility Metric	Key Outcome
ADS-GAN	EHR/Tabular	Conditional GAN	DP-SGD (ϵ=1\epsilon=1ϵ=1)	KS <0.05, AUC parity	Deconfounded cohorts
TTS-GAN	Time-Series (ECG)	Recurrent GAN	Noise injection	Wasserstein <0.1	Rare arrhythmia detection +12%
CycleGAN	Imaging (MRI-CT)	Unpaired GAN	Anonymization	FID=15.2	Lesion segmentation boost
Med-DDPM	Histopathology	Diffusion	Gradient clipping	SSIM=0.92	Rare cancer augmentation
LLMs (Med-PaLM)	Text/EHR Notes	Transformer	Prompt shielding	ROUGE-L=0.85	NLP bias reduction 18%
Digital Twins	Multimodal	VAE-GAN Hybrid	Federated DP	Trial simulation equiv.	Enrollment cost -45%
CTGAN-SMOTE	Genomics/Tabular	Tabular GAN	Local DP	Correlation >0.95	Precision medicine cohorts

4. RESULTS AND CHALLENGES

Quantitative triumphs abound: synthetic-augmented classifiers attain 92-97% AUCs on retinal disease detection (vs. 85% real-only), EHR synthetics fuel survival models with C-indexes >0.82 matching subsampled reals, and DM-images yield perceptual parities (FID<20, SSIM>0.9) for downstream segmentations. Privacy holds: membership attacks succeed <2% on DP-GANs vs. 35% raw, enabling safe sharing; longitudinal fidelity preserves 95%+ correlations in multi-visit EHRs. Regulatory validations confirm synthetics as external arms, cutting trial durations 25-50% for orphan drugs.

Challenges loom large: GAN instabilities manifest mode collapse dropping diversity 30-50%, privacy-utility frontiers force FID hikes >2x under tight ϵ, and high-dimensional pitfalls (omics 10^6 dims) demand memory-intensive training. Validation lags—most studies halt at augmentation, neglecting bias propagation or long-tail risks; ethical pitfalls amplify training biases into synthetics, skewing equity; scalability bottlenecks exclude resource-poor settings. Interpretability deficits hinder clinician trust, with black-box latents obscuring generation rationales.

5. CONCLUSION

Generative AI cements its indispensability in medical research, delivering scalable, privacy-secure synthetic data that democratizes AI development, debiases models, and streamlines trials across siloed modalities and demographics. Empirical proofs affirm near-oracle utility with ironclad anonymization, positioning synthetics as HIPAA/GDPR-compliant bedrock for global collaborations.

Horizons beckon hybrids fusing generative with knowledge graphs for causal fidelity, federated/multi-institution DMs for planetary-scale cohorts, and inverse generation aligning to clinician priors via RLHF. Standardized consortia (e.g., FDA benchmarks, SynthIA metrics) alongside explainable architectures will propel certifications, while ethical audits preempt bias perpetuation. By 2030, these innovations forecast AI-accelerated discoveries in pandemics, rare genomics, and personalized therapeutics, reshaping equitable healthcare paradigms.

REFERENCES

[1] S. R. Abbas, H. Seol, Z. Abbas, and S. W. Lee, “Exploring the role of artificial intelligence in smart healthcare: A capability and function-oriented review,” Healthcare, vol. 13, no. 14, p. 1642, 2025.

[2] H. Han, "Challenges of reproducible AI in biomedical data science", BMC Medical Genomics, Vol. 18, 2025.

[3] C. Aliferis, G. Simon, "Lessons Learned from Historical Failures, Limitations and Successes of AI/ML in Healthcare and the Health Sciences. Enduring Problems, and the Role of Best Practices", Artificial Intelligence and Machine Learning in Health Care and Medical Sciences, pp. 543–606, 2024.

[4] J. Wu, K. Plataniotis, L. Liu, E. Amjadian, and Y. Lawryshyn, “Interpretation for variational autoencoder used to generate financial synthetic tabular data,” Algorithms, vol. 16, no. 2, p. 121, 2023.

[5] J. Gehrmann, E. Herczog, S. Decker, and O. Beyan, “What prevents us from reusing medical real-world data in research,” Scientific Data, vol. 10, Art. no. 459, 2023.

[6] R. Graziosi, M. Ronzani, A. Buliga, C. Di Francescomarino, F. Folino, C. Ghidini, F. Meneghello, and L. Pontieri, “Generating multiperspective process traces using conditional variational autoencoders,” Process Science, vol. 2, Art. no. 8, 2025.

[7] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," IEEE Access, vol. 13, pp. 98504-98529, 2025.

[8] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni and R. Joshi, "Transforming portfolio management with AI and ML: shaping investor perceptions and the future of the Indian investment sector," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1108-1114, 2025.

[9] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," in IEEE Access, vol. 13, pp. 98504-98529, 2025.

[10] P. Jha, P. Jain, A. Kumar, S. Soni, Y. Sharma and P. Agarwal, "The Application of Markov Chains to Linguistic Predictions by Utilising its Inherent Information Entropy," 2025 9th International Conference on Inventive Systems and Control (ICISC), pp. 1110-1114, 2025.