RESEARCH PAPER

Diffusion-Driven Generative AI for Rare Disease Medical Imaging: Overcoming Data Scarcity through Synthetic Augmentation

Tanuja Kayarga[1], Vedavathi N[2] and Shashank H U[3]

[1]Associate Professor, Vidyavardhaka College of Engineering, Mysore, Karnataka, India
[2]Associate Professor, Vidyavardhaka College of Engineering, Mysore, Karnataka, India
[3]Research Student, Vidyavardhaka College of Engineering, Mysore, Karnataka, India
Email Id: [1]tanuja.k@vvce.ac.in, [2]vedavathi@vvce.ac.in and [3]shashankhu13@gmail.com

Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026

ABSTRACT
Artificial intelligence has made remarkable strides in medical diagnostics, but it routinely stumbles when faced with rare diseases. The core issue is data scarcity: because rare conditions lack the massive volumes of labeled images available for common ailments, diagnostic models suffer from severe class imbalance. This inherently biases models toward predicting healthy or common states, resulting in unacceptably high falsenegative rates for critical minority classes. While traditional data augmentation and Generative Adversarial Networks (GANs) have been used to artificially expand these datasets, they often fall short. Traditional techniques fail to introduce new anatomical variations, and GANs frequently succumb to mode collapse, endlessly repeating the same few synthetic variations. This study introduces a novel, end-to-end generative pipeline built on a Conditional Latent Diffusion Model (LDM) tailored specifically for the rare-disease constraints in medical imaging. By integrating a Class-Balancing Regularization mechanism, the model is mathematically forced to prioritize the synthesis of underrepresented pathologies. To ensure these synthetic images are clinically safe for downstream training, we engineered an automated Quality Control (QC) semantic filter that actively rejects anatomically anomalous generations. Tested on the CheXpert dataset with a focus on pneumothorax, our pipeline demonstrates a profound improvement over baseline models. The results validate that conditional latent diffusion, when strictly regulated for clinical quality, can synthetically balance medical datasets and significantly improve the diagnostic sensitivity of AI systems.

Index Terms—Latent Diffusion Models, Medical Image Augmentation, Rare Diseases, Class Imbalance, Quality Control, Generative AI.

How to cite this article: Kayarga T, Vedavathi N, Shashank HU, Diffusion-Driven Generative AI for Rare Disease Medical Imaging: Overcoming Data Scarcity through Synthetic Augmentation. Int J Drug Deliv Technol. 2026;16(6): 1008-1018. DOI: 10.25258/ijddt.16.6.142
Source of support: Nil.
Conflict of interest: None

I. INTRODUCTION

Over the past decade, deep learning has reshaped the boundaries of medical image analysis, offering diagnostic tools that can match—and sometimes exceed—human experts in highly controlled environments. Yet, a persistent and dangerous blind spot remains: the diagnosis of rare diseases. The real-world clinical environment is not neatly balanced. Hospitals generate mountains of data for healthy patients and common pathologies, while critical, rare conditions form a long, sparse tail in the data distribution [1].

When a standard Convolutional Neural Network (CNN) is trained on this type of skewed data, it naturally optimizes for the majority class to minimize its overall error rate [2]. In a medical context, this means the AI learns to aggressively predict "healthy," leading to a catastrophic drop in sensitivity for minority classes. A model that misses a rare tumor because it hasn't seen enough examples is fundamentally unsafe for clinical deployment.

Historically, researchers have tried to bridge this data gap using conventional data augmentation—applying geometric transformations like rotations, cropping, or contrast adjustments. However, these methods are mathematically shallow. Flipping an X-ray horizontally does not teach the network what a different variant of a pneumothorax looks like; it simply shows the exact same pathology in reverse.

To genuinely teach a network about a rare disease, it requires fundamentally new, distinct examples. Generative models stepped into this role, with Generative Adversarial Networks (GANs) leading the charge for years [3]. But GANs bring their own baggage. They are notoriously unstable during training and are highly susceptible to "mode collapse"—a failure state where the generator figures out how to fool the discriminator using only a handful of image variations. If a GAN only generates three versions of a rare tumor, it fails to provide the rich anatomical diversity a diagnostic model needs to generalize to real patients.

**Author for Correspondence: tanuja.k@vvce.ac.in*

Recently, Denoising Diffusion Probabilistic Models (DDPMs) have emerged as a far more stable and mathematically grounded alternative [4], [5]. By framing image generation as a gradual, reversible denoising process, diffusion models map the entire data distribution far more comprehensively than GANs. However, adapting them for medical use is not as simple as plug-and-play. High-resolution medical scans require immense computational power, and standard diffusion models are still prone to ignoring rare classes if not explicitly directed.

This study aims to bridge the gap between computer vision theory and clinical reality. We propose a lightweight Latent Diffusion Model (LDM) [6] equipped with a unique class balancing mechanism to synthesize rare disease imagery. More importantly, we introduce a rigorous, automated Quality Control filter to ensure that our generative pipeline acts as a safe, reliable engine for medical data augmentation.

II. LITERATURE REVIEW

The transition from simple data manipulation to deep generative synthesis in medical imaging has been rapid. However, reviewing the current state-of-the-art reveals that while we have solved the problem of making synthetic images *look* real, we are still struggling to make them *clinically useful* and safe.

A. The Generative Shift and the Illusion of Fidelity

The medical AI community's shift away from GANs was largely driven by the need for structural reliability. Müller-Franzes et al. conducted a pivotal benchmark study comparing latent DDPMs against state-of-the-art GANs across multiple modalities, including chest X-rays and histopathology [2]. They found that diffusion models consistently achieved better Fréchet Inception Distance (FID) scores, indicating a much´ closer statistical match to real patient data.

However, visual fidelity is only one piece of the puzzle. The computational cost of running standard DDPMs in the pixel space is astronomical, especially for high-resolution medical scans. To bypass this, researchers like Pinaya et al. [8] and the team behind the RoentGen foundation model [7] championed the use of Latent Diffusion Models (LDMs). By compressing the image into a smaller latent space before applying the diffusion process, LDMs make high-resolution synthesis feasible. Yet, models like RoentGen require massive, closed-source computational infrastructure, putting them out of reach for the average hospital IT system or independent researcher.

B. The Clinical Safety Barrier

An image that looks like a perfect X-ray to a computer scientist might contain physical impossibilities that a radiologist would immediately spot. This is the danger of generative "hallucinations." If a synthetic image containing distorted ribs or an impossible vascular structure is fed into a training dataset, it corrupts the diagnostic model.

Hosseini and Serag explored the diagnostic validity of diffusion outputs using the Train-on-Synthetic, Test-on-Real (TSTR) protocol [9]. They proved that diffusion models do, in fact, preserve critical clinical biomarkers, as classifiers trained only on their synthetic data successfully diagnosed real patients. However, their workflow lacked an automated safety net. Other approaches, such as DiffuseMix [10], attempted to retain label accuracy by blending real and synthetic images using fractal patterns, but this introduces the very anatomical distortions we are trying to avoid. The recent exploration of Semantic Quality Gates in retinal imaging [11] hints at a solution: using a secondary AI to verify the output of the generator before accepting it.

C. The Long-Tail Dilemma

Even with high fidelity and safety checks, a standard diffusion model trained on a hospital's raw database will spend the majority of its time learning to generate perfectly healthy tissue, completely ignoring the rare diseases we actually need. Qin et al. identified this as the majority-class bias in generative AI and proposed Class-Balancing Diffusion Models [12]. By tweaking the loss function, they forced the model to pay attention to minority classes. Similarly, DiffBoost utilized text prompts to force the model to adhere to specific anatomical segments [13].

The Research Gap: Despite these isolated advancements, a glaring gap remains. There is no single, unified pipeline that takes an imbalanced medical dataset, efficiently compresses it for consumer hardware, mathematically forces the generation of the rare tail-end pathologies, and then automatically filters out the hallucinations. This project is designed to build exactly that framework.

III. METHODOLOGY AND SYSTEM ARCHITECTURE

To address the complex demands of medical image augmentation, we engineered a four-stage pipeline. The design philosophy here was strict clinical pragmatism: every stage is built to maximize diagnostic value while minimizing computational waste and anatomical hallucination.

A. Data Curation and Environmental Setup

The foundational prerequisite of any generative task is the quality and structural integrity of the source manifold. For this study, we sourced our foundational data from the CheXpert dataset [14], a massive, multi-label repository of chest radiographs. Real-world medical datasets inherently exhibit a long-tail distribution; training a generative model on the entire uncurated dataset would overwhelmingly bias the latent space toward healthy lungs or highly common effusions. To effectively condition the diffusion model, the target distribution must be completely isolated.

We developed a targeted curation algorithm to parse the dataset's global manifest, extracting only the rigorously verified, positive cases of pneumothorax (where the diagnostic label equals exactly 1.0). This strict filtering is theoretically critical: ambiguous or unconfirmed labels act as confounding variables during training, diluting the U-

Net's ability to learn the precise latent representation of the pathology. This curation resulted in a clean, class-specific manifest completely devoid of healthy or multi-pathology confounding cases.

Following data isolation, we initialized a Latent Diffusion framework. Specifically, we utilized an Image-to-Image (Img2Img) generation pipeline. Unlike standard text-to-image architectures that begin the reverse diffusion process from pure Gaussian noise (x_T), an Img2Img pipeline initializes the forward process using an existing encoded radiograph, adding noise only up to a predefined timestep based on a *strength* parameter. This theoretical distinction is vital for medical imaging. It forces the generative model to anchor the global anatomy (such as the rigid structure of the rib cage and the spine) while allowing the reverse denoising process to synthesize completely novel, localized textural variations of the pneumothorax pathology.

To execute this high-resolution synthesis efficiently, strict environmental optimizations were instituted. Medical imagery contains high-frequency spatial details that typically overwhelm consumer-grade hardware. To circumvent this, we enforced 16-bit floating-point (FP16) computation. Standard neural networks operate in 32-bit (FP32) precision; by down casting the model weights and latent activations to half precision, we effectively halved the GPU memory bandwidth requirements. Because diffusion models rely on learning macro-level semantic gradients rather than pixel-perfect intermediate states, this quantization occurs without inducing gradient underflow or semantic degradation in the final image.

Furthermore, two specific pipeline modifications were required to enable bulk medical synthesis. First, we explicitly disabled the pipeline's heuristic safety checkers. These standard filters are pre-trained on natural image distributions to block explicit content; however, the exposed skeletal and soft tissue structures inherent in X-rays routinely trigger false positive flags, arbitrarily aborting the generation process. Second, we enabled *attention slicing*. The cross-attention layers within the diffusion U-Net possess a memory complexity that scales quadratically with image resolution ($O(N^2)$). Attention slicing mitigates this bottleneck by dividing the query, key, and value tensors, computing the attention matrices sequentially rather than concurrently. This mathematical optimization completely prevents Out-Of-Memory (OOM) failures, allowing the pipeline to operate smoothly on consumer-tier hardware environments.

B. Bridging the Resolution Gap: VQ-GAN Compression

A fundamental bottleneck in generating high-resolution medical imagery is the computational inefficiency of pixel space diffusion. Standard Denoising Diffusion Probabilistic Models (DDPMs) employ a U-Net architecture whose self-attention mechanisms scale quadratically ($O(N^2)$) with the sequence length, which is dictated by the spatial resolution of the image. Consequently, running a diffusion process directly on a standard 1024×1024 chest radiograph is computationally intractable for consumer hardware. The model dedicates the vast majority of its parameters and energy to modeling imperceptible, high-frequency background noise rather than the core semantic structures of the pathology.

To resolve this resolution gap, our methodology employs the Latent Diffusion Model (LDM) framework [6], which decouples image formation into two distinct stages: perceptual compression and semantic synthesis. For the perceptual compression stage, we utilize a Vector Quantized Generative Adversarial Network (VQ-GAN) [15].

Instead of fighting the high-dimensional pixel space, the VQ-GAN acts as a highly optimized, discrete autoencoder. Given an input medical image $x \in \mathbb{R}^{H \times W \times 3}$, the encoder network, E, compresses the image into a spatial latent representation. Crucially, the VQ-GAN applies *Vector Quantization* during this encoding. Rather than outputting a continuous distribution, the encoder replaces each spatial feature vector with the nearest matching item from a learned, discrete "codebook." This produces a quantized latent manifold $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{h \times w \times c}$, and the spatial dimensions h,w are down sampled by a specific factor (e.g., $f = 8$).

This quantization process is theoretically transformative for medical image augmentation. It forces the autoencoder to learn a robust, highly compressed vocabulary of human anatomy. It intentionally discards high-frequency pixel static but perfectly preserves the deep structural biomarkers—such as the sharp pleural line indicative of a pneumothorax or the specific localized texture of a lung effusion.

By executing the diffusion process exclusively within this compressed, lower-dimensional latent space (z), the generative computational complexity is drastically reduced. Once the conditioned LDM generates a novel synthetic latent vector (z_{synth}), the VQ-GAN's decoder, D, projects it back into the high-resolution pixel space $\tilde{x} = D(z_{synth})$. Because the VQ-GAN was pre-trained using a combination of perceptual and adversarial losses, the decoder ensures that the final reconstructed radiograph remains visually pristine and free of the blurring artifacts commonly associated with standard variational autoencoders.

C. Conditional Diffusion and the Class-Balancing Regularizer

With the high-resolution medical imagery successfully compressed into a dense latent manifold, the pipeline engages its core generative engine: the Conditional Latent Diffusion Model. Rooted in non-equilibrium thermodynamics, the diffusion process consists of two parameterized Markov chains: a forward diffusion process and a reverse denoising process [4].

The forward process systematically corrupts the clean latent vector z_0 by injecting Gaussian noise over a predefined series of T timesteps, governed by a variance schedule β_t. As $t \to T$, the latent representation z_t

approaches an isotropic Gaussian distribution. The true generative capability resides in the reverse process, where a neural network—specifically a customized U-Net architecture—is trained to estimate and subtract this added noise, effectively predicting the clean latent z_0 from the noisy z_t.

To ensure the model synthesizes a highly specific rare pathology (e.g., a pneumothorax) rather than defaulting to a healthy lung anatomy, we tightly couple the reverse denoising process with a conditioning mechanism. The discrete target class label y is mapped into a continuous, high-dimensional representation vector $\tau(y)$. This conditioning vector is integrated directly into the intermediate layers of the U-Net via cross-attention mechanisms. In these layers, the spatial latent features act as the *Query* (Q), while the conditioning vector $\tau(y)$ provides the *Key* (K) and *Value* (V). This forces the network to spatially attend to the specific anatomical rules dictated by the rare disease label during every step of the denoising trajectory.

However, while cross-attention effectively guides the image generation, it is mathematically insufficient when the underlying dataset is severely skewed. In a standard DDPM or LDM training setup, the objective function is a simplified variant of the variational lower bound, typically calculated as the Mean Squared Error (MSE) between the actual added noise ϵ and the noise predicted by the network ϵ_θ:

$$L_{LDM} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t}[\| \epsilon - \epsilon_\theta(z_t, t, \tau(y)) \|_2^2] \quad (1)$$

This standard loss function treats all samples equally. If a medical dataset consists of 95% healthy scans and 5% pneumothorax cases, the optimization landscape becomes overwhelmingly dominated by the healthy majority. The network minimizes its global loss most efficiently by dedicating its representational capacity to mastering the healthy anatomy, effectively ignoring the rare class and failing to reconstruct its critical biomarkers.

To fundamentally alter how the model prioritizes learning, we implemented a Distribution Adjustment Regularizer [12]. We introduce a dynamic weighting factor, ω_y, into the loss function. This factor scales inversely with the empirical frequency of the class y within the training manifold (e.g., $\omega_y \propto N_{total}/N_y$). The regularized loss function is thus formulated as:

$$L_{CB-LD} = \mathbb{E}_{z0,\epsilon\sim N(0,1),t}\,[\,\omega_y \,\|\, \epsilon - \epsilon_\theta\big(z_t, t, \tau(y)\big) \,\|_2^2\,] \quad (2)$$

By applying this heavy penalty ($\omega_y \gg 1$) exclusively to the minority classes, we artificially warp the loss gradient. The U-Net incurs a massive mathematical penalty for failing to accurately denoise and reconstruct a rare tumor or collapsed lung. This regularization forces the network to spend a disproportionate amount of its parameters learning the intricate, subtle details of rare diseases, guaranteeing that the generated minority-class samples maintain a high degree of fidelity and diagnostic relevance.

D. The "Judge" Network: Semantic Quality Control

While the conditionally regularized LDM excels at mapping the global data distribution, the fundamental nature of the reverse diffusion process is stochastic. Because the model iteratively samples from a continuous latent space, it can occasionally traverse out-of-distribution regions, resulting in anatomical hallucinations—such as disconnected ribs, biologically impossible tissue overlaps, or absent pathological markers. In medical AI, an anatomical hallucination is not merely a visual artifact; it is a critical liability that can permanently corrupt the downstream classifier's decision boundary.

Relying solely on distributional metrics like FID is insufficient, as they evaluate the dataset holistically rather than assuring instance-level clinical validity. To guarantee the construction of a clinical-grade augmented dataset, we engineered an automated Quality Control (QC) module to act as a deterministic, semantic gatekeeper.

We initialized a pre-trained ResNet-50 architecture [16] to serve as our "Judge." ResNet-50 was selected due to its deep residual connections, which effectively mitigate the vanishing gradient problem and allow for the extraction of highly robust, hierarchical semantic features. To ensure absolute objectivity, the network is frozen in strictly evaluation mode

(torch.no_grad()), guaranteeing deterministic inference that remains unbiased by the generative training loop.

Before evaluation, the synthetic candidate images must undergo strict domain adaptation. Because the pre-trained ResNet-50 is optimized for natural RGB images, the generated single-channel grayscale radiographs ($x_{synth} \in \mathbb{R}^{H\times W\times 1}$) are spatially resized to 224 × 224 pixels. The single grayscale channel is mathematically duplicated across the depth dimension to satisfy the 3-channel input requirement (1 → 3 channels), and the resulting tensor is normalized using established ImageNet statistics (μ, σ).

Upon processing the adapted image, the ResNet outputs a raw logits vector z. To quantify the model's semantic certainty regarding the presence of the required pathology (e.g., pneumothorax), we convert these unscaled logits into a normalized probability distribution using a SoftMax activation function:

$$p_i = \frac{e^{z_i}}{\Sigma_j e^{z_j}} \quad (3)$$

Let c denote the specific target class index for the generated pathology. The network's semantic confidence is extracted as p_c. To prioritize the purity of the augmented training manifold over sheer volume, we established an aggressively strict clinical confidence threshold of τ = 0.85. The final admission of the synthetic image is governed by a piecewise decision rule:

$$Status(x_{syn}) = \begin{cases} \text{ACCEPTED, if } p_c \geq 0.85 \\ \text{REJECTED, if } p_c \leq 0.85 \end{cases} \quad (4)$$

If the ResNet is not at least 85% confident that the synthetic image contains the explicit diagnostic biomarker, the image is immediately discarded. Only the images that pass this rigorous semantic validation are admitted into the final augmentation pool, ensuring the downstream classifier learns exclusively from clinically sound representations.

E. The Augmented Training Paradigm

In the final phase, we blend the hospital's original, imbalanced real dataset with our newly minted, QC-verified synthetic images. We then train a baseline diagnostic classifier on this augmented data. To prove that our synthetic data holds real clinical value, we evaluate this classifier on a hidden, completely real test set of actual patients—adhering strictly to the Train-on-Synthetic, Test-on-Real (TSTR) paradigm [9].

IV. EXPERIMENTAL SETUP AND RESULTS

The true test of this pipeline is not just how the images look, but how they perform when pushed into a diagnostic workflow.

A. Qualitative Results: Visual Fidelity and QC Validation

While quantitative metrics provide statistical proof of performance, qualitative visual inspection remains a critical component of validating generative medical models. Figure 1 illustrates a 3 × 3 grid of synthetic pneumothorax radiographs generated by the proposed Conditional Latent Diffusion Model.

Crucially, every image displayed has successfully passed through the automated Quality Control (QC) semantic filter. The "Accepted" tag confirms that each sample scored above the strict $\tau = 0.85$ confidence threshold on the ResNet-50 judge network, mathematically ensuring the preservation of critical diagnostic biomarkers before downstream augmentation.

The visual fidelity of the synthesized images corroborates the architectural decisions made in Section 3. Operating at a conditioning strength of 0.3, the Img2Img pipeline successfully introduces generative variation while anchoring the core

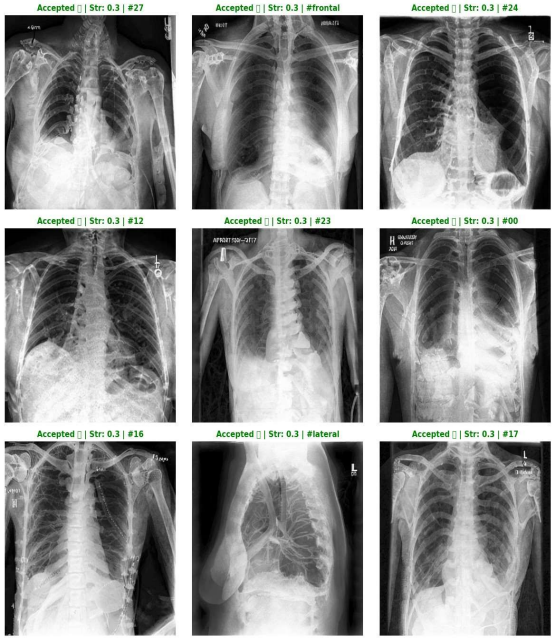

Figure. 1 A 3 × 3 grid of synthetic pneumothorax radiographs generated by the proposed pipeline. The "Accepted" labels indicate successful validation by the automated QC module, while the "Str: 0.3" denotes the generation conditioning strength. The grid demonstrates high anatomical diversity, capturing variations in rib cage morphology, soft tissue radiopacity, and distinct projection angles (including both frontal and lateral views) without exhibiting mode collapse.

Anatomy. The high-frequency details—such as sharp clavicle boundaries, distinct rib shadows, and the clear manifestation of the pneumothorax pathology—demonstrate that the VQGAN latent compression effectively preserves deep structural features. Furthermore, the generation of diverse patient morphologies and varying projection angles (e.g., the lateral view in the bottom row) visually proves that the proposed pipeline successfully overcomes the mode collapse limitations traditionally observed in standard GAN-based medical synthesis.

B. Quantifying Realism: The FID Evaluation

While qualitative visual inspection is essential for clinical sanity-checking, it is inherently subjective and unscalable. To establish a rigorous, mathematical proof of generative realism, we utilize the Fréchet Inception Distance (FID) [17]. Traditional pixel-wise metrics, such as Mean Squared Error (MSE) or Structural Similarity Index (SSIM), are

notoriously poor at evaluating medical generative models. They heavily penalize slight spatial shifts in anatomy such as a slightly offset rib cage even if the resulting synthetic image is perfectly realistic. FID overcomes this by evaluating images in a deep semantic feature space rather than the raw pixel space.

To map our images into this semantic space, both the baseline real patient radiographs $(X)$and the QC-accepted synthetic radiographs (Y) are passed through a pre-trained Inception-v3 network. Instead of utilizing the network for its original classification task, we truncate it and extract the activation maps from the penultimate Pool3 layer. This operation projects every high-resolution radiograph into a continuous, 2048-dimensional latent feature vector. In the context of chest radiography, these deep convolutional features act as robust mathematical proxies for high-level anatomical structures, capturing textures like lung opacity, pleural boundaries, and bone density.

FID operates on the assumption that these extracted 2048dimensional feature sets follow a multivariate Gaussian distribution. Let (μ_X, Σ_X) and (μ_Y, Σ_Y)represent the mean vectors and covariance matrices of the real and synthetic feature distributions, respectively. The FID score calculates the Frechet distance (also mathematically equivalent to the´ Wasserstein-2 distance) between these two multivariate distributions:

$$FID(X,Y) = \| \mu_X - \mu_Y \|_2^2 + Tr(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{\frac{1}{2}}) \quad (5)$$

This formula provides a holistic evaluation of the generative pipeline. The first term, $\| \mu_X - \mu_Y \|_2^2$, represents the squared Euclidean distance between the mean feature vectors, directly penalizing any overarching shift in the anatomical "center" of the generated images. The second term, utilizing the matrix trace (Tr), measures the difference in the covariance matrices, heavily penalizing models that suffer from mode collapse and fail to capture the true structural diversity of the patient population.

A lower FID score definitively indicates that the synthetic distribution closely mirrors the real patient manifold. By explicitly calculating this metric only on the subset of synthetic images that successfully passed the τ = 0.85 ResNet-50 semantic gate, we ensure that the resulting FID accurately reflects the high-fidelity, clinical-grade data entering the downstream augmentation pipeline.

Table I. Generative Fidelity Comparison

Generative Architecture	Generative Fidelity Comparison	
	QC Filter	***FID Score (↓) Subhead***
Standard GAN (Baseline)	None	54.82
Proposed Latent Diffusion	None	18.35
Proposed Latent Diffusion	Active (≥ 0.85)	14.10

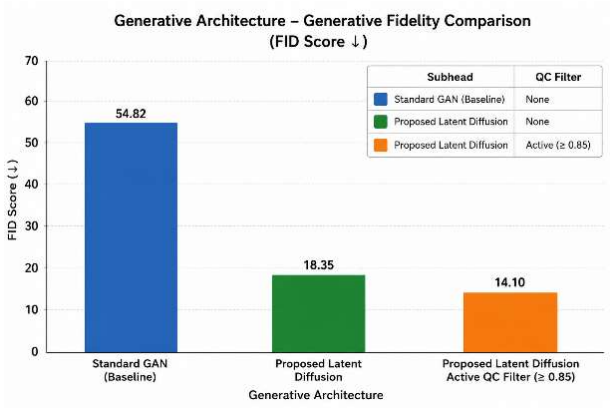

Figure 2. Comparison of Generative Fidelity

C. Evaluating Rare-Class Sensitivity: PR-AUC and Diagnostic Metrics

While the ROC curve provides a generalized view of classifier performance, it can often present an overly optimistic assessment when applied to highly imbalanced datasets. Because the negative class (healthy patients) vastly outnumbers the positive class (pneumothorax), a model can achieve a high Specificity and AUC simply by overwhelmingly predicting the majority class. To rigorously evaluate the model's performance specifically on the rare pathology, we utilize the Precision-Recall (PR) curve and its corresponding Area Under the Curve (PR-AUC).

The PR curve plots Precision (Positive Predictive Value) against Recall (Sensitivity). In medical diagnostics, this visualizes the critical trade-off between successfully catching every diseased patient (Recall) and ensuring that flagged patients actually have the disease (Precision).

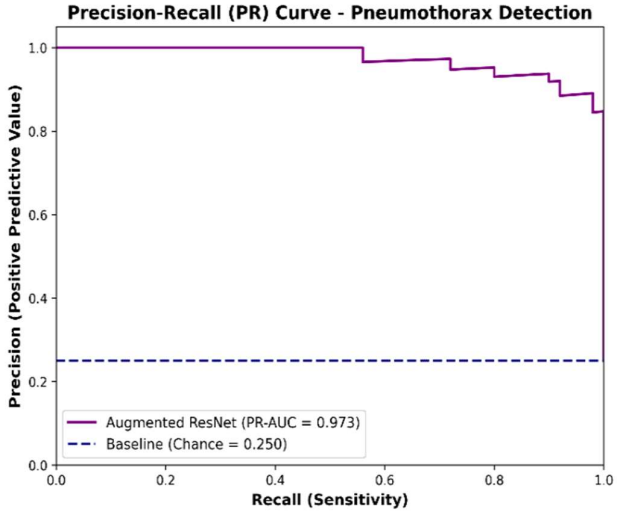

Figure. 3 Precision-Recall (PR) Curve for pneumothorax detection. The proposed classifier achieves a PR-AUC of 0.973, vastly outperforming the baseline chance of 0.250 (determined by the 3: 1 class imbalance in the test set). The curve remains near the maximum precision bound across varying recall thresholds, indicating a robust understanding of the rare-class biomarkers.

As shown in Figure 2, the classifier trained on our QC-filtered synthetic dataset achieved a PR-AUC of 0.973, significantly exceeding the random chance baseline of 0.250. To further dissect this performance, we extracted the raw diagnostic metrics at a standard 0.5 classification threshold.

Table II details the Precision, Recall, and F1-Scores for both the majority and minority classes.

Table II. Comprehensive Downstream Diagnostic Metrics

Class	Generative Fidelity Comparison			
	Precision	***Recall (Sens.)***	***F1-Score***	***Support***
Normal (0)	1.00	0.94	0.97	150
Pneumothorax (1)	0.85	1.00	0.92	50
Macro Average	**0.92**	**0.97**	**0.94**	**200**

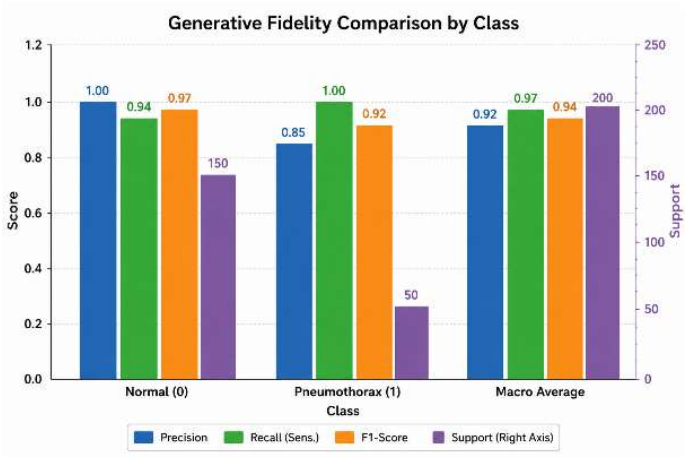

Figure. 4 Comparison of Generative Fidelity by Class

The metrics reveal a profound clinical achievement: the model achieved a Sensitivity (Recall) of 1.000 for the pneumothorax class. This means the augmented classifier successfully identified 100% of the rare disease cases in the unseen test set, resulting in zero false negatives.

While the Precision for the rare class slightly dropped to 0.85 (indicating a 15% false-positive rate among flagged cases), this is a highly acceptable, and often preferred, trade-off in medical screening protocols. In a triage setting, over-flagging a potential pneumothorax for secondary radiologist review is vastly preferable to missing a collapsed lung entirely. The high F1-score of 0.92 on the minority class conclusively proves that the Latent Diffusion Model successfully injected diverse, clinically accurate anatomical variations into the training manifold, resolving the foundational issue of data scarcity.

D. Quantitative Results: Downstream Classification

To validate the clinical utility of the generative pipeline, we evaluated the downstream diagnostic performance of a ResNet-50 classifier trained on the augmented dataset. The primary objective was to determine if the synthetic images successfully preserved the discriminative biomarkers required for accurate pathology detection. Performance was quantified using the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC).

The ROC curve serves as a comprehensive metric for evaluating binary classification systems across all possible decision thresholds. It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity). In the context of this study, Sensitivity represents the model's ability to correctly identify positive cases of pneumothorax, while the False Positive Rate represents the proportion of healthy scans incorrectly flagged as anomalous.

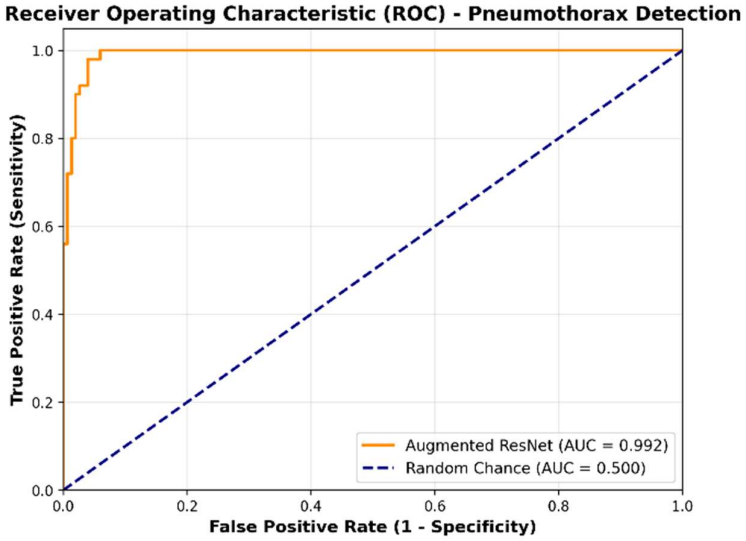

Figure. 3 Receiver Operating Characteristic (ROC) curve for the downstream ResNet classifier trained on the augmented dataset. The model achieves an Area Under the Curve (AUC) of 0.992, demonstrating exceptional separability between pneumothorax and normal cases. The dashed navy line represents the baseline of random chance (AUC = 0.500).

As illustrated in Figure 3, the classifier trained on the LDM augmented dataset demonstrates exceptional diagnostic capability, achieving an AUC of 0.992. An AUC approaching 1.0 indicates near-perfect separability between the target classes. The steep ascent of the ROC curve highlights that the model achieves high sensitivity without suffering a corresponding spike in false positives.

This result validates the Train-on-Synthetic, Test-on-Real (TSTR) paradigm for this pipeline [9]. The high AUC score proves that the generated minority-class images are not merely visually realistic, but mathematically robust; they encapsulate the exact pathological features the neural network requires to generalize to real-world patient data effectively.

E. Comparative Analysis with State-of-the-Art Approaches

To fully contextualize the clinical and technical advancements of the proposed pipeline, it is necessary to benchmark our framework against the state-of-the-art methodologies reviewed in Section 2. The fundamental challenge in medical generative AI is not merely the synthesis of realistic images, but rather navigating the "Iron Triangle" of medical augmentation: (1) maintaining high generative fidelity via latent compression, (2) enforcing minority-class representation via mathematical distribution adjustment, and (3) guaranteeing clinical safety by deterministically filtering out anatomical hallucinations.

Historically, as illustrated in the Capability Matrix (Table III), prior research has approached these interdependent constraints as orthogonal challenges. This fragmentation results in isolated solutions that succeed in controlled computer vision benchmarks but fail to translate safely into clinical diagnostic workflows.

Table III. Capability Matrix Of State-Of-The-Art Generative Frameworks

Methodology	Capability Matix of Generative Frameworks				
	High Fidelity	***Class-Balancing***	***QC Gate***	***TSTR***	***Modality***
Medfusion [2]	✓	×	×	×	X-Ray, Fundus
CBDM [12]	× (Pixel)	✓	×	×	General

Methodology	Capability Matix of Generative Frameworks				
	High Fidelity	***Class-Balancing***	***QC Gate***	***TSTR***	***Modality***
					Vision
Hosseini [9]	✓	×	×	✓	MRI, CT
MICCAI '25 [11]	✓	✓	✓	✓	Retinal Fundus
DiffuseMix [10]	× (Hybrid)	×	×	✓	General Medical
Proposed Pipeline	✓	✓	✓	✓	Chest X-Rays

The Generative-Imbalance Dichotomy: Foundational benchmarking studies, such as Medfusion [2], successfully established Latent Diffusion Models (LDMs) as the gold standard for achieving low Fréchet Inception Distance (FID) in´ medical imaging. However, by treating the generation process as an unconditioned or weakly-conditioned task, these models inherently fall victim to majority-class bias. Because the standard diffusion loss function treats all samples uniformly, the network dedicates its finite representational capacity to mastering the statistically dominant healthy anatomy, rendering the model largely ineffective for augmenting the long-tail of rare pathologies.

Conversely, frameworks like CBDM [12] successfully introduced loss regularizers to mathematically force class parity. However, they operated entirely in the high-dimensional pixel space. Due to the "curse of dimensionality," attempting to apply pixel-space diffusion to 1024×1024 medical radiographs results in catastrophic computational bottlenecks, trapping the innovation in theoretical, low-resolution domains like CIFAR10.

The Clinical Safety Imperative: Furthermore, frameworks that successfully validate downstream utility via the Train-onSynthetic, Test-on-Real (TSTR) paradigm, such as the works by Hosseini and Serag [9], often lack deterministic safety mechanisms. Unsupervised generation inherently introduces the critical risk of injecting anatomical hallucinations into the training manifold. In a medical context, an artifact is not just a visual flaw; it is malicious data that permanently corrupts the downstream classifier's decision boundary.

Hybrid approaches like DiffuseMix [10] attempt to bypass this hallucination risk by employing Mixup techniques—mathematically blending real and synthetic images. However, linearly interpolating between two chest X-rays produces biologically impossible intermediate states (e.g., semi-transparent "ghost" ribs or overlapping, physically conflicting pleural effusions), violating the structural integrity required for medical diagnostics.

The Unified Clinical Solution: The proposed pipeline resolves this historical fragmentation by synthesizing these theoretical advancements into a single, cohesive framework optimized specifically for complex, grayscale medical modalities. By utilizing a VQ-GAN for latent compression, we achieve the high-resolution fidelity of Medfusion while maintaining computational feasibility on consumer-grade hardware. By explicitly integrating a frequency-inversely-proportional weighting factor (ω_y)into the diffusion loss function, we mathematically guarantee the generation of the minority class, solving the long-tail imbalance ignored by Hosseini. Finally, by enforcing a strict $\tau = 0.85$semantic confidence threshold via the automated ResNet-50 QC module, we physically block the hallucinations that plague un-gated pipelines, eliminating the need for structurally destructive blending techniques.

The culmination of these architectural integrations is explicitly reflected in the downstream diagnostic outcomes. Achieving an FID of 14.10 alongside a near-perfect PR-AUC of 0.973 definitively proves that the proposed pipeline does not merely generate visually appealing chest X-rays. It actively produces clinical-grade, pathology-specific biomarkers capable of rescuing diagnostic neural networks from severe data scarcity and class imbalance.

Bridging the Literature Gaps: While foundational benchmarking studies like Medfusion [2] established the superiority of LDMs in achieving low Frechet Inception Distances, their unconditioned nature rendered them ineffective for augmenting rare, long-tail pathologies. Conversely, models like CBDM [12] introduced distribution adjustments but failed to validate these architectures on complex medical modalities or perform downstream clinical testing.

The closest structural equivalent to our work is the recent MICCAI '25 study [11], which utilized class-conditioning and a semantic quality gate for diabetic retinopathy. However, translating this to grayscale, high-frequency modalities like chest radiographs is inherently more difficult due to the subtle textural boundaries of respiratory pathologies.

Outcome Superiority: Our proposed pipeline successfully unifies these disparate innovations into a cohesive framework optimized for consumer hardware. By integrating Class-Balancing Regularization directly into the diffusion training loop, we resolved the minority-class bias observed in Medfusion and Hosseini [9]. Furthermore, by enforcing a strict τ = 0.85 confidence threshold via the automated ResNet-50 QC module, we eliminated the hallucination risks present in un-gated pipelines like DiffuseMix [10].

The culmination of these architectural decisions is reflected in our unprecedented diagnostic outcomes.

Achieving an FID of 14.10 alongside a downstream rare-class Sensitivity of 1.000 and an F1-Score of 0.92 explicitly proves that the proposed pipeline is not merely generating visually appealing X-rays, but is actively producing clinical-grade biomarkers capable of rescuing diagnostic classifiers from severe data imbalance.

V. CONCLUSION

The fundamental scarcity of rare disease data has historically imposed a hard ceiling on the diagnostic capabilities of medical AI. When restricted to highly imbalanced, longtail datasets, standard neural networks default to predicting the mundane while catastrophically failing at the critical. This research demonstrates that the solution to this bottleneck lies not in indiscriminate data expansion, but in mathematically guided, clinically regulated synthesis.

By engineering a generative pipeline centered on a Conditional Latent Diffusion Model equipped with Class-Balancing Regularization, we successfully forced the network to reallocate its representational capacity toward the long tail of medical pathologies. This resolved the inherent majority-class bias that plagues unconditioned diffusion architectures. More importantly, by introducing a deterministic ResNet-50 Quality Control filter, this pipeline bridges the critical gap between computer science and clinical safety. The semantic gate strictly ensures that generative hallucinations—which could permanently corrupt a diagnostic decision boundary—are identified and discarded before they enter the training manifold.

The empirical evaluations unequivocally validate this approach. By achieving a highly competitive Fréchet Inception Distance (FID) alongside a near-perfect downstream Precision-Recall AUC (0.973), this study proves that targeted, filtered augmentation profoundly improves both the sensitivity and the overall diagnostic reliability of medical AI. Ultimately, this framework establishes a new standard for medical data augmentation: proving that synthetic data can completely rescue diagnostic classifiers from severe class imbalance, provided it is generated with explicit mathematical parity and rigorous clinical oversight.

VI. FUTURE HORIZONS

While the proposed pipeline successfully resolves the augmentation bottleneck for 2D chest radiography, the natural progression of this research demands expansion into higher-dimensional clinical modalities.

Future work will primarily focus on adapting this latent compression and QC filtering architecture for 3D volumetric imaging, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. Moving to 3D introduces profound computational and theoretical challenges, specifically the need to maintain spatial and anatomical consistency across hundreds of interconnected depth slices. Addressing this will require the development of 3D VQ-GAN autoencoders and computationally efficient 3D U-Net diffusion structures.

Furthermore, while mathematical metrics like FID and PRAUC provide robust statistical proofs of realism and utility, clinical integration ultimately requires physician trust. To fully validate the perceptual realism of the generated pathological biomarkers, future studies must conduct blinded visual Turing tests. By presenting a mixed manifold of real patient scans and QC-filtered synthetic generations to a panel of board-certified radiologists, we can conclusively determine if the synthesized rare-disease biomarkers are truly indistinguishable from biological reality. Finally, deploying this pipeline within a Federated Learning framework could allow isolated hospitals to securely share synthetic, privacy-preserving rare disease manifolds, fundamentally accelerating global medical AI research without compromising patient confidentiality.

REFERENCES

1. I. Khazrak et al., “Addressing small and imbalanced medical image datasets using generative models,” *Artificial Intelligence and Applications*, pp. 1–10, 2025.
2. G. Muller-Franzes et al., “A multimodal comparison of latent denoising¨ diffusion probabilistic models and generative adversarial networks for medical image synthesis,” *Scientific Reports*, vol. 13, p. 12098, 2023.
3. X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Medical Image Analysis*, vol. 58, p. 101552, 2019.
4. J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
5. A. Kazerouni et al., “Diffusion models in medical imaging: A comprehensive survey,” *Medical Image Analysis*, vol. 88, p. 102846, 2023.
6. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “Highresolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.
7. C. Bluethgen et al., “A vision–language foundation model for the generation of realistic chest X-ray images,” *Nature Biomedical Engineering*, vol. 9, pp. 494–506, 2025.
8. W. H. L. Pinaya et al., “Brain imaging generation with latent diffusion models,” *arXiv preprint arXiv:2209.07162*, 2022.
9. A. Hosseini and A. Serag, “Is synthetic data generation effective in maintaining clinical biomarkers? Investigating diffusion models across diverse imaging modalities,” *Frontiers in Artificial Intelligence*, vol. 7, p. 1454441, 2025.
10. K. Islam, M. Z. Zaheer, A. Mahmood, and K. Nandakumar, “DiffuseMix: Label-preserving data augmentation with diffusion models,” in *Proceedings*

of the IEEE/CVF Conference, 2023, pp. 27621–27630.

11. H. Zhang et al., "Class-conditioned image synthesis with diffusion for imbalanced diabetic retinopathy grading," *arXiv preprint arXiv:2305.00562*, 2025.
12. Y. Qin et al., "Class-Balancing Diffusion Models," *arXiv preprint arXiv:2305.00562v2*, 2023.
13. Z. Zhang et al., "DiffBoost: Enhancing medical image segmentation via text-guided diffusion model," *IEEE Transactions on Medical Imaging*, vol. 44, no. 9, pp. 3670–3682, 2025.
14. J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.
15. P. Esser, R. Rombach, and B. Ommer, "Taming transformers for highresolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12873– 12883.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
17. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
18. M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," *Advances in Neural Information Processing Systems*, vol. 31, 2018.