Synthetic Brain MRI Generation Using Class-Conditional Latent Diffusion Models

Pronoy Ghosh
Department of Computing Technologies
SRM Institute of Science and Technology
Chennai, India
pg7994@srmist.edu.in

Ananya Kumar
Department of Computing Technologies
SRM Institute of Science and Technology
Chennai, India
ak5454@srmist.edu.in

Dr. Nancy P
Department of Computing Technologies
SRM Institute of Science and Technology
Chennai, India
nancyp@srmist.edu.in

***Abstract*—The scarcity of annotated training samples and severe distributional skew across diagnostic categories remain critical impediments in automated medical image analysis, most notably when developing reliable deep learning frameworks for the identification and classification of brain MRI tumors. Generative modeling offers a technically grounded avenue for enriching constrained clinical collections with realistic synthetic imagery. However, Pixel-level adversarial architectures exhibit well-known susceptibilities to optimization instability and frequently sacrifice fine structural detail. This paper presents a class-guided latent diffusion framework engineered for the targeted synthesis of brain MRI volumes. The methodology couples a Variational Autoencoder (VAE), responsible for deriving a dense, geometrically consistent latent embedding of MRI scans, with a conditioning-aware diffusion network that learns the category-stratified latent distribution for normal and tumour tissue independently. Grayscale brain MRIs at full resolution are mapped to a compact low-dimensional representation without sacrificing essential morphological cues. A self-attention-enhanced conditional latent diffusion model then generates category-aligned latent codes, subsequently reconstructed into the image domain through the frozen VAE decoder. Experimental results show that the framework produces synthetic volumes that are structurally coherent and have anatomical signatures that can be recognized. The method provides a computationally feasible, adversary-free synthesis pathway directly applicable to the enhancement of training data and the fortification of downstream classifiers in clinical MRI pipelines.**

***Index Terms*—Latent Diffusion Model, Variational Autoencoder, Brain MRI Synthesis, Class-Conditional Generation, Medical Image Augmentation, Deep Learning, Generative Models**

I. Introduction

Neural networks utilizing deep hierarchical representations have emerged as the benchmark paradigm for a growing array of medical image analysis challenges, consistently demonstrating superior performance in lesion classification, anatomical delineation, and computer-assisted diagnosis across MRI modalities. However, practical implementation is often hindered by ongoing data scarcity, uneven label distributions, and variability caused by diverse imaging hardware—challenges that are particularly pronounced in limited clinical archives. Sparse ground-truth annotations and disparities in class proportions represent the most significant challenges in the compilation of training corpora for clinical decision-support systems. In the neuroimaging field, getting labeled pathological scans is very difficult because it requires a lot of planning and following rules. This makes it hard to create large, balanced datasets. Generative synthesis has consequently garnered persistent interest as a method for focused dataset augmentation. Alrashedy et al. [1] presented BrainGAN, which combines adversarial image generation with convolutional classification to show that using synthetic samples along with real training data can lead to measurable improvements in diagnosis. Adversarial methods, however, are still vulnerable to recognized failure modes, including unstable gradient dynamics, modal coverage collapse, and the attenuation of subtle anatomical structures. Denoising Diffusion Probabilistic Models (DDPMs) [2] have become popular as a principled alternative because they provide more stable optimization landscapes and a wider range of outputs. The Latent Diffusion Model (LDM) extension [3] further addresses computational cost by confining the denoising iterations to a compressed representation rather than full pixel space. This work leverages these advances to develop a class-conditional LDM specific to neuroimaging synthesis. Three primary contributions are made: (i) a VAE optimised on grayscale brain MRI to yield a structured, compact latent space; (ii) a self-attention-augmented class-conditional diffusion network operating solely within that latent domain; and (iii) an end-to-end generation pipeline that produces label-guided normal and tumour MRI images without recourse to adversarial objectives.

II. Related Work

Vijayalakshmi [4] demonstrated that classical machine learning pipelines applied to dermatoscopic imagery can achieve clinician-grade accuracy in melanoma identification, providing an early and widely referenced validation of computational tools in dermatological screening.

Li et al. [5] conducted a systematic review of deep learning strategies across heterogeneous medical imaging domains, documenting the diversity of clinical use cases while establishing data volume and annotation quality as the overriding determinants of generalisation capability.

Caseneuve et al. [6] tackled image quality control in thoracic radiograph pipelines by creating automated rejection systems to get rid of blurry, underexposed, or otherwise poor-quality images. This is a preprocessing method that can also be used in MRI curation workflows.

Alrashedy et al. [1] integrated adversarial synthesis with convolutional classification in the BrainGAN system for brain MRI analysis; the reported accuracy improvements resulting from synthetic augmentation highlighted the necessity for generative architectures that provide enhanced structural fidelity and training reliability.

Ho et al. [2] showed that DDPMs can produce adversarial generation quality while allowing for much more stable training. Rombach et al. [3] expanded this concept to high-resolution synthesis through LDMs functioning in compressed latent spaces with significantly diminished computational requirements—the theoretical basis for the proposed methodology.

III. Dataset

A. Dataset Composition

The experimental corpus consists of 400 grayscale brain MRI images partitioned into two diagnostic categories:

- **Normal:** 170 MRI images
- **Tumour:** 230 MRI images

B. Preprocessing Pipeline

Before training the model, each image was normalized to 256×256 pixels and rescaled to the intensity range [0,1]. Figure 1 shows that the preprocessing workflow has three steps that happen in order: spatial resampling through bilinear interpolation, intensity rescaling through min-max normalization, and online data augmentation that adds random horizontal reflections, small-angle rotations, and contrast perturbations to improve generalization capacity.

The collection's distributional skew (230 tumor images compared to 170 normal images) could cause a bias in the learned conditional distribution. Balanced class-conditional mini-batch sampling during diffusion model training helps fix this problem by making sure that no one category has too much of an effect on the learned generative distribution.

```
class MRIDataset(Dataset):
    def __init__(self, root, split="Train"):
        self.samples = []
        self.transform = transforms.Compose([
            transforms.Resize((256,256)),
            transforms.RandomHorizontalFlip(),
            transforms.ToTensor(),
            transforms.Normalize([0.5],[0.5])
        ])
```

Fig. 1. MRI dataset class structure and preprocessing pipeline. The pipeline applies three sequential stages — resize, normalize, and augment — to all 400 grayscale brain MRI images before training.

IV. Variational Autoencoder

A. Overview

A Variational Autoencoder (VAE) is a probabilistic generative architecture that estimates the underlying data distribution, thereby enabling both high-fidelity reconstruction of observed inputs and the synthesis of novel samples drawn from the learned prior. In contrast to deterministic autoencoders that compress each observation to a single fixed-point representation, a VAE parameterises the encoder output as a probability distribution characterised by a mean vector $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\sigma}^2$, from which latent codes are drawn via the reparameterisation identity, preserving end-to-end differentiability across stochastic sampling operations.

Given an input image $\mathbf{x}$, the encoder network infers an approximate posterior distribution over latent codes:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\boldsymbol{\mu},\, \boldsymbol{\sigma}^2\mathbf{I}\big) \tag{1}$$

A latent code $\mathbf{z}$ is subsequently obtained through the reparameterisation identity:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

which decouples the stochastic component from the learnable parameters, enabling gradient-based optimisation to proceed through the sampling node without approximation.

VAEs are particularly well-suited to medical image analysis contexts because their regularised training objective yields smooth, topologically coherent latent manifolds, within which semantically related inputs occupy proximate regions. This structural property makes it easy to generate reliable downstream data and makes it possible to interpolate between latent codes in a meaningful way. These are very useful features when training data is limited.

B. Architecture

The encoder ingests 256×256 grayscale MRI inputs and employs a succession of strided convolutional layers to systematically halve the spatial resolution: $256 \to 128 \to 64 \to 32$. The terminal convolutional feature map is linearly projected into two distinct parameter tensors representing the posterior mean $\boldsymbol{\mu}$ and log-variance $\log \boldsymbol{\sigma}^2$, jointly defining the approximate posterior. The resulting latent code occupies a space of dimensionality 8 channels $\times\, 32 \times 32$.

The decoder reconstructs full-resolution outputs through a mirrored series of transposed convolutions that progressively restore spatial dimensions: $32 \to 64 \to 128 \to 256$. LeakyReLU activations are used throughout the encoder to mitigate vanishing gradients, while standard ReLU activations operate within the decoder; layer normalisation is applied at intermediate stages to maintain training stability. The complete encoder-decoder architecture is presented in Fig. 2.

C. Loss Function

VAE training minimises a compound objective that jointly enforces reconstruction accuracy and latent space regularity:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \beta\, \mathcal{L}_{\text{KL}} \tag{3}$$

The reconstruction fidelity term is computed as the ℓ_1 norm between the ground-truth image $\mathbf{x}$ and its decoded estimate $\hat{\mathbf{x}}$:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_1 \tag{4}$$

```python
class Encoder(nn.Module):
    def __init__(self):
        super().__init__()
        self.conv = nn.Sequential(
            nn.Conv2d(1, 32, 4, 2, 1), nn.ReLU(),    # 128
            nn.Conv2d(32, 64, 4, 2, 1), nn.ReLU(),   # 64
            nn.Conv2d(64,128,4,2,1), nn.ReLU(),      # 32
        )
        self.mu = nn.Conv2d(128, 8, 1)
        self.logvar = nn.Conv2d(128, 8, 1)

    def forward(self, x):
        h = self.conv(x)
        return self.mu(h), self.logvar(h)

class Decoder(nn.Module):
    def __init__(self):
        super().__init__()
        self.deconv = nn.Sequential(
            nn.ConvTranspose2d(8,128,4,2,1), nn.ReLU(),
            nn.ConvTranspose2d(128,64,4,2,1), nn.ReLU(),
            nn.ConvTranspose2d(64,32,4,2,1), nn.ReLU(),
            nn.Conv2d(32,1,3,1,1),
            nn.Tanh()
        )
```

Fig. 2. Encoder and Decoder architecture of the proposed VAE. The encoder compresses 256×256 MRI inputs through successive strided convolutions to produce $\boldsymbol{\mu}$ and $\log\boldsymbol{\sigma}^2$ in an $8\times32\times32$ latent space; the decoder reconstructs the original resolution via transposed convolutions.

The KL divergence penalty enforces proximity of the aggregate posterior to a standard Gaussian prior:

$$\mathcal{L}_{\mathrm{KL}} = -\frac{1}{2}\sum_{j=1}^{d}\Big(1+\log\sigma_j^2-\mu_j^2-\sigma_j^2\Big) \tag{5}$$

The scalar weighting coefficient $\beta = 0.001$ deliberately down-weights the regularisation term relative to reconstruction, yielding a more expressive latent space at the cost of mild posterior departure from the prior—a deliberate trade-off that benefits the downstream diffusion process.

D. VAE Training and Loss Convergence

The VAE was optimized for 50 epochs with the Adam optimizer and a starting learning rate of 1×10^{-3}. The left panel of Fig. 4 shows that the training loss drops quickly and smoothly. It starts at about 0.21 and drops sharply during the first seven epochs as the encoder-decoder pair learns rough structural representations of the brain.

After 10 epochs, the total loss reduction is more than 80%, and the value stays close to 0.050. The next phase, which lasts from epochs 10 to 25, results in small further reductions from about 0.050 to about 0.030 as the network gets better at picking out more and more precise anatomical details. After epoch 25, the loss stabilizes and stays between 0.025 and 0.030, which is a strong sign that the model is converging without overfitting.

V. Latent Diffusion Model

A. Overview

The iterative procedure in the proposed framework operates within the compressed latent space $\mathbf{z}$ encoded by the frozen VAE, rather than in the original high-dimensional pixel domain. This architectural choice reduces both memory consumption and computational complexity, following the to the LDM paradigm of Rombach et al. [3]. The VAE distills the salient anatomical content of the MRI scans into a compact representation; the diffusion network therefore needs to only model the distribution over these efficient latent codes.

B. Forward Diffusion Process

Beginning with a pristine latent code $\mathbf{z}_0$, Gaussian noise is progressively introduced over $T = 1000$ discrete timesteps in accordance with a cosine-annealed variance.

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\Big(\mathbf{z}_t; \sqrt{1-\beta_t}\,\mathbf{z}_{t-1}, \beta_t\mathbf{I}\Big) \tag{6}$$

Marginal sampling at any target timestep t is accomplished analytically as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{7}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ represents the cumulative schedule product. The cosine schedule is better than the linear one because it makes noise transitions smoother near both ends of the diffusion trajectory, which helps keep the gradient magnitudes uniform and the training stable throughout the optimization.

C. Reverse Diffusion and UNet Architecture

A conditional UNet models the reverse denoising trajectory by predicting the noise component $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, y)$ that is added at each forward step, based on the discrete class label $y \in \{\text{normal}, \text{tumour}\}$. The UNet has four main architectural parts that work together:

1) **Residual blocks with timestep embeddings:** allowing the network to dynamically modulate its behaviour as a function of the current noise level.
2) **Class-label embeddings:** explicitly steering the generative process toward anatomical features characteristic of normal or tumour tissue.
3) **Self-attention layers:** positioned at the bottleneck between the contracting and expanding paths to capture long-range spatial dependencies.
4) **Skip connections:** transmitting fine spatial information across the encoder-decoder boundary to improve reconstruction fidelity.

Classifier-free guidance is incorporated by stochastically discarding class conditioning during training with probability 0.1, enabling the model to learn both unconditional and conditional generation. At inference, a linear combination of conditional and unconditional score estimates amplifies adherence to the specified class.

D. Self-Attention Mechanism

Self-attention is applied to spatially-flattened intermediate feature maps, enabling every position to compute weighted interactions with every other position in the sequence:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are learned projections of the query, key, and value, respectively, and d_k is the dimensionality of the key projection. This operation is especially important when it comes to pathological brain MRI, where tumor masses may occupy spatially distributed, non-contiguous regions that are outside the receptive field of regular convolutional kernels.

E. Noise Scheduling

The cosine variance schedule establishes a monotonically increasing sequence $\beta_1, \beta_2, \ldots, \beta_T$, which dictates the amount of noise introduced at each forward step. The complementary quantities $\alpha_t = 1\beta_t$ and their cumulative product $\bar{\alpha}_t$ control how much of the original signal is kept at each time step. In the beginning of the forward process, noise increments are small, which keeps most of the latent structure. As time goes on, the original representation is slowly overwhelmed until the distribution is close to isotropic Gaussian noise. The cosine schedule's smooth transition profile keeps the noise level from changing too quickly, giving consistent gradient information across all time steps.

F. Training Pipeline

The overall training scheme comprises two strictly sequential, non-overlapping phases:

Stage 1 — VAE Pre-training: Pre-training the VAE The VAE is improved over 50 epochs by using the composite reconstruction-plus-KL-divergence objective with $\beta = 0.001$. After training is done, all of the encoder's parameters are set in stone and will stay that way for the next stage.

Stage 2 — Diffusion Training: The frozen VAE encoder is used on all of the training images to make a set of hidden representations. Then, the conditional UNet is trained for 100 epochs. At each iteration, a random timestep is chosen, the corresponding noisy latent is made, and the network is updated by minimizing the mean squared error between its noise prediction and the actual injected noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, y)\right\|^2\right] \tag{9}$$

An Exponential Moving Average (EMA) of model parameters is kept throughout all diffusion training steps. The EMA-smoothed weights are used as the inference model, which makes the generated outputs more consistent and visually stable than the raw training checkpoint.

VI. Results

A. VAE Reconstruction Quality

Following completion of the 50-epoch training schedule, VAE exhibits robust reconstruction capability. Fig. 3 presents paired comparisons of original MRI images with their corresponding VAE reconstructions. The outputs faithfully reproduce the overall anatomical topology, skull boundary, and gross hemispheric symmetry of the originals.

The quality of these reconstructions show that the latent manifold encodes semantically meaningful and structurally informative representations. Cortical features such as sulcal patterns and folding morphology are unfortunately, only partially retained . A degree of spatial smoothing is attributable to the ℓ_1 reconstruction objective and the inherent information bottleneck of the $8 \times 32 \times 32$ latent dimensionality.

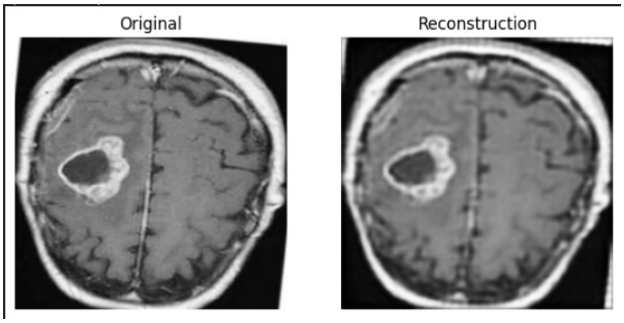

Fig. 3. VAE reconstruction results after 50 training epochs. Original brain MRI images (left column) are compared with VAE-reconstructed counterparts (right column). The reconstructions faithfully preserve gross anatomical structure, brain boundaries, and hemispheric symmetry, confirming the semantic fidelity of the learned $8\times32\times32$ latent space.

B. Training Loss Analysis

Fig. 4 presents the optimisation trajectories for both model components: the VAE (left panel) and the latent diffusion model (right panel).

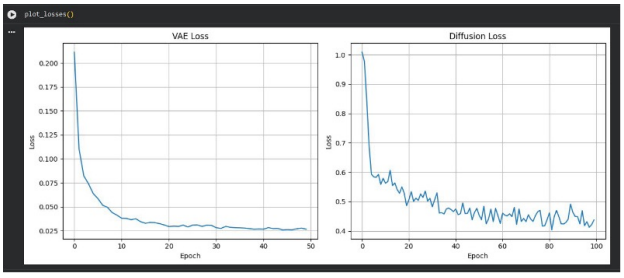

Fig. 4. Training loss curves for both model stages. *Left:* VAE loss over 50 epochs, showing smooth monotonic convergence from ≈ 0.21 to ≈ 0.027. *Right:* Diffusion model loss over 100 epochs, exhibiting a steep initial decline followed by stable stochastic oscillations in the range 0.43–0.50, characteristic of random timestep-sampling training dynamics.

VAE Loss Convergence. The VAE loss follows a smooth, monotonically decreasing trajectory throughout training. An initial value of approximately 0.21 at epoch one gives way to a sharp descent across epochs 1–7, during which the model rapidly acquires representations of dominant brain structures. By epoch 10, the loss has contracted to roughly 0.050, a reduction exceeding 80% of the initial value. The descent continues in smaller increments through epochs 10–25, converging from ≈ 0.050 toward ≈ 0.030, before entering a stable plateau in the range 0.025–0.030 for the remainder of training, confirming well-regularised optimisation without overfitting.

Diffusion Model Loss Convergence. The diffusion training loss initialises near 1.0 before undergoing a pronounced decline across the first seven epochs. This is driven by the UNet's rapid acquisition of noise-prediction capability at highly corrupted

timesteps. From epoch 7 through epoch 20, incremental reductions bring the loss from approximately 0.60 to 0.50; the visible stochastic fluctuations during this phase are an expected consequence of random timestep sampling, which introduce gradient variance across consecutive batches, yet the macro-level trend remains consistently downward-pointing.

Between epochs 20 and 60, the loss continues to decay in small steps, settling into the range 0.43–0.50. Beyond epoch 60, negligible reduction is observed, and the loss oscillates within this narrow band for the duration of training. These persistent fluctuations show the stochasticity of the diffusion objective rather than any form of training divergence. The loss plateau suggests that the model has approached its representational ceiling given the 400-image corpus, and that larger datasets would likely enable continued incremental improvement beyond this threshold.

C. Synthetic MRI Generation — 50 Epochs

Figs. 5 and 6 display representative synthetic brain MRI images produced by the diffusion model at 50 training epochs, across both the Normal and Tumour classes.

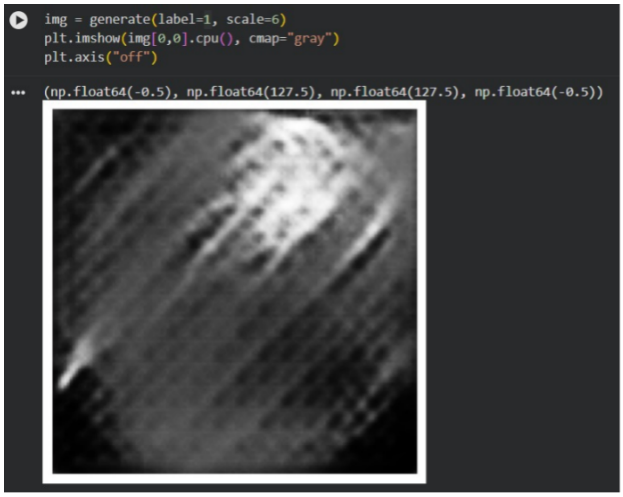

Fig. 5. Synthetic brain MRI generated at 50 diffusion training epochs (Sample 1). The image exhibits early structural emergence: a recognizable oval brain boundary and rough internal contrast gradients are visible, but anatomical detail remains coarse and textures are largely uniform.

At the 50-epoch checkpoint, generated images start showing early-stage convergence of the generative model: an almost oval cranial boundary and coarse internal contrast gradients are discernible, intra-tissue textures still remain largely homogeneous. This behaviour is consistent with the characteristic coarse-to-fine optimisation trajectory of diffusion-based generative models.

D. Synthetic MRI Generation — 100 Epochs

Figs. 7 and 8 illustrate representative outputs obtained at the 100-epoch milestone, revealing pronounced qualitative advances over the earlier checkpoint.

The 100-epoch outputs demonstrate a clear advancement in perceptual quality, with more pronounced differentiation

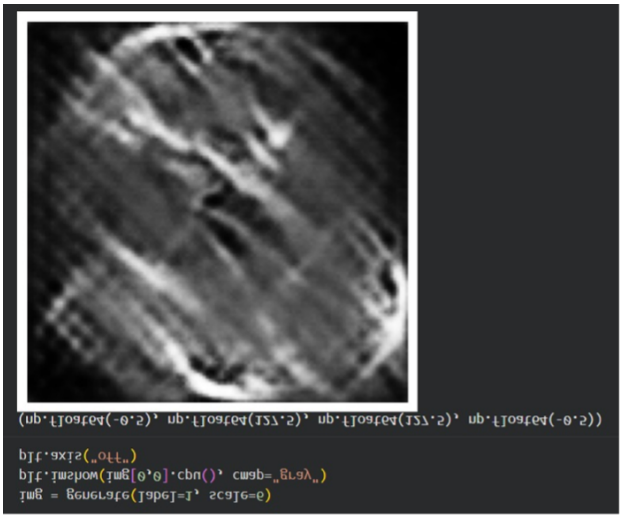

Fig. 6. Synthetic brain MRI generated at 50 diffusion training epochs (Sample 2). Similar to Sample 1, the output demonstrates initial global structural formation while lacking high-frequency anatomical texture, consistent with the coarse-to-fine learning progression of the diffusion process.

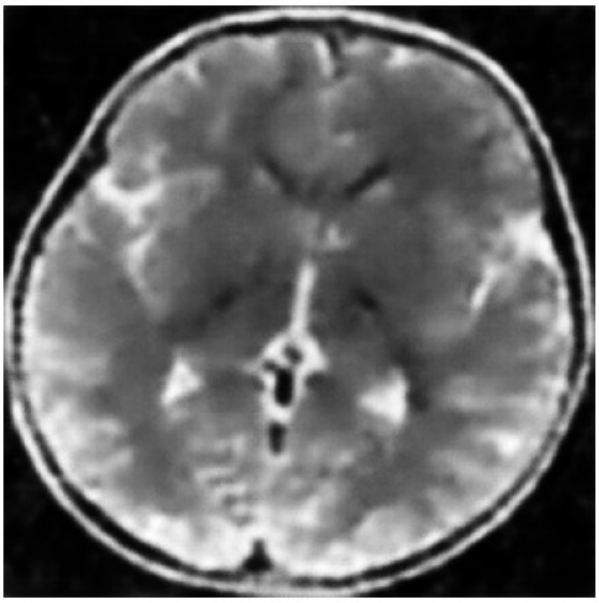

Fig. 7. Synthetic brain MRI generated at 100 diffusion training epochs (Sample 1). Compared with the 50-epoch output, the image shows improved structural differentiation, more defined internal brain architecture, and enhanced anatomical plausibility.

between normal and tumour morphologies, sharper delineation of internal anatomical boundaries, and improved tissue contrast throughout the synthesised volume. The progressive quality improvement from 50 to 100 epochs substantiates the claim that the model is acquiring progressively richer representations of cerebral anatomy over the course of training. Fine-grained texture and high-frequency detail remain constrained, a limitation attributable to the small dataset of the 400-image training corpus.

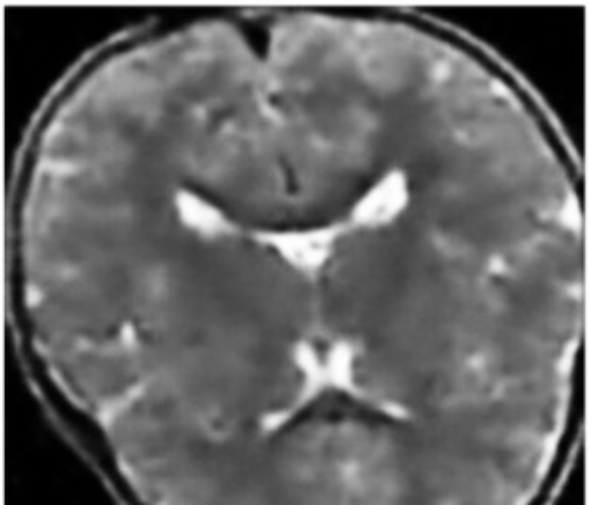

Fig. 8. Synthetic brain MRI produced at 100 diffusion training epochs (Sample 2). The output shows that the morphology of each class is getting better over time, with clearer anatomical boundaries and better tissue contrast than the 50-epoch sample.

E. Image Difference Analysis

To measure the difference in perceptual fidelity between real and synthetic MRI images, we made pixel-wise absolute difference maps for pairs of real and synthetic images that were similar. For a real image $\mathbf{x}$ and its corresponding synthetic image $\hat{\mathbf{x}}$, both normalized to the range $[0, 1]$, the difference map $\mathbf{D}$ is formally defined as:

$$\mathbf{D}_{ij} = \left|\mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij}\right| \tag{10}$$

where (i, j) indexes pixel spatial locations. Elevated values in $\mathbf{D}$ (brighter pixels) correspond to regions of high local discrepancy, while suppressed values indicate areas where generation and reality closely coincide.

The difference maps show that reconstruction errors are mostly found at high-frequency texture boundaries and sulcal margins. In contrast, large-scale structural regions, like the cranial outline and major anatomical compartments, have much lower error magnitudes. This error distribution is typical of latent-space generative models, which naturally capture global structure more accurately than fine-grained local texture, especially when training data is limited.

VII. Discussion

The experimental findings confirm that the proposed latent diffusion framework generates meaningful coarse-level structural representations of brain MRI anatomy. The VAE is able to achieve strong reconstruction accuracy, and reproduces the anatomical organisation of input scans with high fidelity. This indicates that the encoded latent manifold is suitable for generative modelling.

The diffusion model produces synthetic images showing recognisable brain morphology and emergent class-specific differentiation. High frequency texture and fine anatomical detail are still inadequately captured. The constraining factor here is the scale of the training corpus; with only 400 training images, the model's ability is quite bounded. Diffusion models are data-intensive, requiring thousands of training examples and extended training schedules to achieve high-fidelity generation[2].

Analysis of the loss trajectories reinforces this. The chosen encoder-decoder design is appropriate for this task and that KL regularisation operates effectively, seen by the VAE loss converges smoothly and monotonically. The diffusion loss, while exhibiting a clear overall downward trend, retains persistent fluctuations arising from the variance in random timestep sampling. Access to a larger dataset would likely enable continued incremental improvement beyond the observed epoch-60 plateau.

The self-attenuation component improves coherence in generated images, though its efficacy is limited under data scarcity: attention modules learn more meaningful spatial correspondences when exposed to a diverse array of training examples. The class imbalance in the dataset (230 tumour versus 170 normal images) can introduce mild generative bias toward tumour-like structural patterns under conditional generation, which is addressable through cost-sensitive sampling or targeted minority-class augmentation strategies. the framework demonstrates the practical viability of adversary-free latent diffusion as a stable generation strategy, and shows a clear route to substantially improved output quality through systematic dataset expansion.

VIII. Conclusion

This paper proposes a class-conditional latent diffusion network architecture to generate synthetic brain MRI images. Pre-training of a Variational Autoencoder (VAE) resulted in an efficient mapping from grayscale brain images to compact and regularised latent space representation. This demonstrated remarkable reconstruction quality in 50 epochs. The training loss monotonically decreased from 0.21 to 0.027 over the same period of time.

A conditional diffusion model built on top of UNet architecture was then trained for 100 epochs on latent space representation of images. This model learned how to synthesise class-specific latent representations that can be mapped back to image space using VAE pre-trained decoder. Thus, a scalable and train-stable approach has been devised for medical images generation, allowing for producing realistic anatomical structures in synthetic brain MRIs as an alternative to adversarial generator models. Loss analysis confirms efficiency of each training step and shows that training loss for conditional diffusion model decreases from approximately 1.0 to a stable value of 0.43–0.50.

Analysis of pixel-level difference shows that structure is reproduced in generated MRIs very effectively. Texture, however is yet a challenge for future research. In conclusion, it has been confirmed that data availability and costly computations represent two major limitations for medical image synthesis using diffusions. Increased data availability and further training should lead to significantly higher results.

IX. Future Work

Prospective extensions to the current framework encompass:

(i) retraining on large-scale repositories such as BraTS to improve anatomical diversity and model generalisation;

(ii) extending the architecture from two-dimensional slice synthesis to volumetric 3D MRI generation

(iii) incorporating more expressive latent architectures and cross-attention conditioning mechanisms [3];

(iv) rigorous quantitative benchmarking using established perceptual and signal quality metrics including FID (Fréchet Inception Distance), SSIM (Structural Similarity Index), and PSNR (Peak Signal-to-Noise Ratio);

(v) prolonged training beyond 100 epochs augmented with cosine annealing learning rate decay to further improve diffusion model convergence; and

(vi) downstream clinical validation assessing the utility of synthetic images in brain tumour classification and segmentation tasks.

References

[1] H. H. N. Alrashedy, A. F. Almansour, D. M. Ibrahim, and M. A. A. Hammoudeh, "BrainGAN: Brain MRI image generation and classification framework using GAN architectures and CNN models," *Sensors*, vol. 23, no. 5, p. 2528, 2023.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[4] M. Vijayalakshmi, "Melanoma skin cancer detection using image processing and machine learning," *International Journal of Trend in Scientific Research and Development*, vol. 3, pp. 780–784, 2019.

[5] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, p. 1273253, 2023.

[6] G. Caseneuve, I. Valova, N. LeBlanc, and M. Thibodeau, "Chest X-ray image preprocessing for disease classification," *Procedia Computer Science*, vol. 192, pp. 658–665, 2021.