STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, November 2025, pp 2534–2558.
Published online in International Academic Press (www.IAPress.org)

CAT-VAE: A Cross-Attention Transformer-Enhanced Variational Autoencoder for Improved Image Synthesis

Khadija Rais [1,*], Mohamed Amroune [1], Mohamed Yassine Haouam [1], Abdelmadjid Benmachiche [2]

[1] *Laboratory of mathematics, informatics and systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, 12002, Algeria*
[2] *Department of Computer Science, LIMA Laboratory, Chadli Bendjedid, University, El-Tarf, PB 73, 36000, Algeria*

Abstract Deep generative models are increasingly useful in medical image analysis for solving various issues, including class imbalance in classification tasks, which motivates the development of multiple methods. The Variational Autoencoder (VAE) is recognized as one of the most popular image generators. However, the utilization of convolutional layers in VAEs weakens their ability to model global context and long-range dependencies. This paper presents CAT-VAE, a hybrid approach combining VAE and Cross-Attention Transformers (CAT), in which a cross-attention mechanism is utilized to foster long-range dependencies and enhance the quality of the generated images. The CAT-VAE achieved better image quality and the highest SSIM, PSNR, FID, and MSE compared to the standard VAE. An experiment was conducted where a CNN classifier model was trained without data augmentation, with augmentation based on VAE, and using synthetic data generated by CAT-VAE. The CNN achieved the highest accuracy of 97.50% with the Ultrasound breast cancer dataset and 93.62% with the MRI Brain tumor dataset, based on CAT-VAE synthetic images, which improves generalization and resilience. These results highlight CAT-VAE's ability to produce diverse and realistic synthetic datasets.

Keywords Variational Autoencoder (VAE), Cross-Attention Transformers (CAT), Synthetic images, Imbalanced classification, Data augmentation.

DOI: 10.19139/soic-2310-5070-2546

1. Introduction

Medical image analysis is increasingly contributing to disease diagnosis, where deep learning techniques have remarkably succeeded in automating the classification and segmentation of medical images, typically achieving performance comparable to or better than that of human experts. One of the main challenges to developing accurate and resilient models is the problem of class imbalance, a common issue in medical datasets where some diseases are rare. This imbalance can lead to models prioritizing the majority classes, reducing their diagnostic accuracy and reliability [23]. This hurdle must be overcome to ensure that computer-aided diagnostic tools are accurate and clinically viable.

To address the problems brought by class imbalance, data augmentation strategies have been widely researched [6]. Conventional techniques such as rotating, flipping, or intensity rescaling the image often do not provide sufficient diversity and may not accurately portray the character of minority or rare classes [12]. In recent years, advanced deep generative models like Generative Adversarial Networks (GANs) and VAEs have gained favor as efficient means for producing realistic medical images [24], balancing data while enhancing model performance. Although GANs have shown great potential, their susceptibility to training instability, model collapse, and high

*Correspondence to: Khadija Rais (Email: khadija.rais@univ-tebessa.dz). Laboratory of mathematics, informatics and systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, 12002, Algeria.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

complexity introduces challenges in their applicability in the medical field, where accuracy and high-quality image generation are essential. Instead, VAEs are more stable, offering a probabilistic solution. Despite their usefulness, traditional VAEs are limited in addressing long-distance dependencies and the complex structure of medical images [10].

The traditional VAEs primarily employ CNNs for encoding and decoding, which excel at local feature detection but struggle to handle global spatial relationships in images. This might lead to blurry image generations, thereby reducing their suitability for medical image augmentation. This study presents CAT-VAE, a novel deep generative model that integrates transformer-based cross-attention mechanisms and VAEs to facilitate image generation. Unlike typical VAEs, which employ fully convolutional networks, CAT-VAE utilizes cross-attention mechanisms to optimize the interaction between the latent space and the encoder's feature maps. CAT-VAE ensures the generated images retain significant spatial patterns, improve contextual coherence, and are more robust than normal VAEs. The key contributions are:

- We introduce cross-attention mechanisms in the VAE, where the latent vector (query) interacts with the encoder's feature maps (keys and values). This allows for a more structured transformation of latent representations into high-quality synthesized images.
- The classical VAEs cannot learn the long-range dependencies since they use convolutional layers. With the introduction of self-attention and multi-head attention (MHA), CAT-VAE can learn complex spatial relations, producing more realistic and finer generations of images.
- We deploy CAT-VAE for the classification of breast cancer (Ultrasound images) and brain tumor (MRI images), a problem where class imbalance is an essential concern. Synthetic images are created by our model to balance the datasets, enhancing the performance for differentiation between classes.

The rest of the paper presents section 2, which introduces related work about VAEs and other generative models for medical image synthesis, followed by section 3, which presents the methodology behind CAT-VAE. Section 4 highlights our results and comparison, section 5 discusses the results, and section 6 concludes the paper.

2. Related works

VAEs have been extensively applied to image generation since they can capture complex data distributions into a continuous latent space. They can produce high-quality, diverse, and realistic images, which are useful in artistic and scientific contexts [14]. Scientists in this work [11] present the Hamiltonian VAE (HVAE) for medical image synthesis and mask synthesis, improving upon traditional VAEs by providing a better posterior distribution estimation. The method is designed to generate high-quality, diverse images and precise tumor masks and performs better than GAN-based approaches when data are limited. Its capability is demonstrated through experiments with the BRATS and HECKTOR datasets for different medical imaging modalities. The authors in another study [19] compare Discriminator-VAE (Disc-VAE) with GANs in terms of SSIM, PSNR, and accuracy. The findings highlight the significance of synthetic image quality in enhancing the generalization and robustness of AI models for medical applications. The PAVAE model generates synthetic brain lesion images to augment small datasets for Laser Interstitial Thermal Therapy (LITT) procedures. It uses two networks: a mask generation network and a mask-guided lesion synthesis network. With the utilization of condition and mask embedding blocks, PAVAE produces realistic lesions and improves segmentation performance over conventional data augmentation methods [8]. The EndoVAE algorithm utilizes VAEs for endoscopy-image synthesis to address data constraints and confidentiality issues of medical images. Different from GANs, EndoVAE avoids mode collapse as well as stability issues, representing a stable and efficient data replacement or augmentation scheme for the overall dataset. The method exhibits prospective promise in endoscopy-image synthesis with varied realistic images, especially in medical image analysis using deep-learning model training [5]. Researchers use a VAE for reconstructing PET brain images in the diagnosis of Alzheimer's disease. It studies the effect of including or not including certain disease stages on image reconstruction accuracy. Results suggest that the addition of cognitively

normal information improves reconstruction quality, whereas the late mild cognitive impairment group spoils it. Performance is gauged through PSNR and SSIM metrics, which reflect the importance of class selection in training for improved disease representation [9]. VAE-GAN was trained to synthesize cine MRI images at high resolution from low-resolution tagged MRI scans. This approach reduces the need for additional MRI acquisitions, yielding time and cost savings. The model was trained and fine-tuned on a self-generated dataset of 3,774 tagged and cine MRI image pairs from 20 normal subjects. The method performed superiorly in generating natural cine MRI images, preserving anatomical information, and allowing subsequent motion analysis and segmentation processes [15]. HCAL is a novel deep learning framework to generate realistic and diverse brain structural connectivity networks. It employs graph VAE (GVAE) and a hemisphere-separated generator with a cross-connectome aggregation mechanism to capture both local and global topological properties of brain networks. In experiments on the ADNI dataset, HCAL improves the diversity and quality of generated brain networks and the accuracy of disease diagnosis [27]. DACMVA is a deep architecture specifically designed for cross-modal data augmentation in datasets with missing values. DACMVA utilizes VAE to learn cross-modality mappings and impute missing data, aiding downstream prediction tasks. In the case of predicting cancer survival using tabular gene expression data, DACMVA significantly outperforms state-of-the-art alternatives like TDImpute and naive oversampling, particularly in the presence of high missing rates. The algorithm greatly enhances prediction accuracy in scenarios with limited data, achieving substantial performance improvements [21].

Attention mechanisms enhance medical image quality by emphasizing vital features and eliminating redundant noise. They contribute to the generation of clearer, higher-resolution, and more detailed images, allowing finer structures and textures to be observable. It leads to better image reconstruction and preservation of anatomical detail, making images closer to anatomical structures. Thus, image clarity and medical image task diagnosis are significantly boosted. Moreover, transformers significantly enhance the quality of medical images by retaining both local and global information, which means better image super-resolution, denoising, reconstruction, synthesis, and registration. Because they can learn to capture long-range dependencies, the images are sharper and more detailed with fewer artifacts and noise. Transformers also enhance texture and structural detail preservation, producing high-fidelity, high-resolution images that improve the reliability and accuracy of medical image interpretation [13]. Advanced variants of VAEs are powerful tools, including transformer-based approaches such as DALL·E that combine VAEs with transformers for zero-shot text-to-image generation, producing realistic images from text descriptions [14]. ResViT is a novel medical image synthesis generative adversarial model that combines the strengths of vision transformers and CNNs. Compared to traditional CNN-based methods, which have no contextual awareness, ResViT uses aggregated residual transformer (ART) blocks to enhance global context and local precision. The model employs residual connections for learning varied features, a channel compression module for efficient information extraction, and a weight-sharing strategy to prevent redundant computational costs [4]. Pan et al. propose a Swin-Transformer-based diffusion model for medical image synthesis because of the scarce training data in AI-based medical imaging. It employs a forward Gaussian noise process and a transformer-based denoising reverse process. It has been trained on chest X-rays, heart MRIs, pelvic CTs, and abdomen CT datasets [18]. Deep Convolutional GAN (DCGAN) combined with a Vision Transformer (ViT) in this work [7] to augment the dataset for brain tumor detection. DCGAN generates synthesized images to augment the small dataset, which further enhances the performance. The hybrid methodology achieved high accuracy with low training loss, which significantly outperformed the model without data augmentation. 4D-VQ-GAN is a generative model employed to predict Idiopathic Pulmonary Fibrosis (IPF) progression using longitudinal 3D CT scans. It combines a 3D vector-quantized GAN with a Neural ODE-based temporal model to generate realistic CT volumes at a given time point. The approach helps in modeling disease progression continuously and predicting survival outcomes [25]. MC-DDPM is a transformer-based denoising diffusion probabilistic model that aims to generate high-quality synthetic CT (sCT) images from MRI examinations. The procedure simplifies radiation treatment planning by eliminating the need for CT scans, reducing registration errors, and patient radiation exposure. Using Swin-VNet in a diffusion process, MC-DDPM effectively captures MRI-to-CT correspondence to produce accurate sCTs for brain and prostate datasets [17]. CALF-GAN is a cross-modal medical image generative adversarial model that boosts modality-specific feature synthesis, as well as long-range dependency modeling, over CNNs without the high cost of computing needed for transformers. It uses a latent attribute

separation module and a multi-scale convolutional attention mechanism [26]. MSG-SAGAN is an attention-drawn multi-scale GAN model that is capable of synthesizing diverse, high-quality, synthetic biomedical X-ray images. MSG-SAGAN addresses common training issues in GANs, such as mode collapse and instability, through its application of attention mechanisms and learning multi-scales of gradients. The model has improved diversity and stability in image generation over previous models, such as MSG-GAN [22]. The authors of another paper suggest a deep learning model for liver tumor classification by combining data augmentation with GANs and Convolutional Block Attention Module (CBAM) and Enhanced Channel Attention (ECA) mechanisms to enhance feature extraction. Using the Duke Liver dataset, the model (CBAM with VGG19) achieves high classification accuracy, demonstrating improved diagnostic ability for liver diseases and the potential for improved patient outcomes [3]. AttnGAN incorporates an attention mechanism within its generator to focus on essential regions within medical images while the generation proceeds. AttnGAN produces real, high-resolution synthetic images using selective focus on major features that closely simulate the original BraTS20 set of brain tumor images[20].

Table 1 shows the generative methods in medical imaging that have advanced but still have gaps in many areas. Approaches based on VAE like HVAE, Disc-VAE, and PAVAE ensure more stable training and better posterior approximation. They're great when there's not much data available, but they struggle to generate realistic images in complex scenarios. For example, EndoVAE tackles data augmentation of endoscopic images but generates lesser-quality images than GANs. The VAE-GAN hybrid model uses the advantages of both architectures to produce high-resolution images while ensuring anatomical correctness. However, this model is derived from healthy subjects, meaning the results may not be entirely applicable to subjects with conditions. Transformer models like ResViT and Swin-Transformer Diffusion are very context-aware and can generate high-fidelity images. With their ability to model long-range dependency through attention mechanisms, such methods enable improved overall performance in various challenging synthesis tasks. Often, they require substantial computational resources and meticulous tuning to produce optimal results. Like ML-DDPM, approaches like CALF-GAN and MC-DDPM add multi-scale convolutional attention and diffusion processes for better realism and modeling of disease progression. These methods focus on improving image diversity and feature extraction through attention models such as MSG-SAGAN, ViT combined with DCGAN, and GAN CBAM ECA. However, they are quite stable and accurate. Most of these methods are complex to train and require large datasets, which are often not accessible in healthcare settings. Despite significant progress made in generative models for the synthesis of medical images, various issues remain open. Existing approaches have difficulty generating high-quality, diverse, and semantically accurate medical images. Most models are prone to mode collapse, training instability, high computational demands, and poor generalizability across multiple datasets or different medical imaging modalities. These difficulties limit their use in real clinical settings where good-quality synthetic data should be used for training robust diagnostic models, addressing the issue of data scarcity, and maintaining patient confidentiality. Therefore, there is a need for novel approaches that can generate realistic and diverse medical images efficiently and remain computationally tractable. This research focuses on the following main questions:

- How can the quality and diversity of synthetic medical images be improved to overcome the challenges of existing generative models?

- Can a new generative framework improve the diagnostic performance of downstream AI when trained with synthetic medical images, particularly in medical imaging where data is sparse and class imbalances exist?

To overcome these issues, a new approach, CAT-VAE, is proposed in this research. The CAT-VAE model includes CAT within the VAE structure to enhance the ability of the VAE to learn both global dependencies and local details in medical images. CAT-VAE learns to generate more diverse and clinically useful synthetic data tailored to specific diagnostic categories by conditioning the generation process on class labels. The proposed model aims to improve existing models' flaws by reducing computational costs, improving training stability, and generating high-fidelity images suitable for increasing training datasets in AI diagnostic systems.

Table 1. Summary of Transformer-Based, Attention Mechanism, and VAE approaches for medical image synthesis and augmentation

Approach	Authors, Year	Description	Dataset	Evaluation Metrics	Advantages	Limits
Based on VAE Approaches						
HVAE	Kebaili et al., 2023 [10]	HVAE for medical image synthesis improving posterior approximation.	BRATS, HECKTOR	DSC, PSNR, SSIM	High-quality images, useful in data-scarce conditions	May produce poor quality in complex cases
Disc-VAE	Rais et al., 2024 [19]	VAE with discriminator for medical image augmentation.	BraTS2020, Breast cancer dataset	SSIM, PSNR, Accuracy	Improves diversity and AI model performance	Dataset acquisition remains challenging
VAE	John et al., 2021 [9]	VAE trained on PET scans for AD stage reconstruction.	PET brain scans dataset	PSNR, SSIM	Highlights importance of CN data for reconstruction	LMCI group reduces accuracy
PAVAE	Huo et al., 2022 [8]	Brain lesion synthesis framework using mask-guided generation.	T1-weighted MRI dataset	PSNR, SSIM, NMSE, Dice, Jaccard, AS, HD	Enhances data diversity, improves segmentation	Requires expert validation
VAE-GAN	Liu et al., 2021 [15]	VAE-GAN with dual-cycle constraints for cine MRI synthesis.	Tagged/cine MRI dataset	PSNR, SSIM	Reduces acquisition cost, maintains anatomical structure	Focused on healthy subjects, segmentation not integrated
HCAL	Zuo et al., 2023 [27]	GVAE-based synthesis of diverse brain structural connectivity.	ADNI dataset	MMD, Accuracy, Sensitivity, Specificity	Improves classification accuracy, high-quality networks	Focused on structural connectivity synthesis
EndoVAE	Diamantis et al., 2022 [5]	VAE-based augmentation for endoscopic image synthesis.	KID dataset	AUC	Avoids GAN issues, data-efficient	May produce unrealistic images, needs validation
DACMVA	Rajaram et al., 2023 [8]	Cross-modal data augmentation using VAEs for missing data imputation.	Cancer survival gene dataset	Prediction accuracy, Wilcoxon test	Effective in low-data scenarios, improves prediction	Limited to tabular data, scalability not discussed
Based on Transformers and Attention Mechanism						
ResViT	Dalmaz et al., 2022 [4]	Transformer-based GAN model integrating ART blocks to combine CNN precision with transformer context-awareness.	IXI Dataset, BRATS dataset	SSIM, PSNR, FID	Improved contextual sensitivity, high-quality synthesis	High computational cost, complex training
Swin-Transformer Diffusion Model	Pan et al., 2023 [18]	Diffusion-based medical image synthesis framework using a Swin-transformer network for denoising.	Chest X-rays, heart MRI, pelvic CT, abdomen CT	Visual Turing test, IS, FID, DS	Generates high-quality images, improves AI training	Requires careful tuning
4D-VQ-GAN	Zhao et al., 2025 [25]	Generates longitudinal 3D CT images of IPF patients to model disease progression.	IPF CT scan dataset	MSE, SSIM, PSNR, C-Index	Models disease progression, aids personalized treatment	Requires large datasets, complex training
ViT with DCGAN	Haque et al., 2023 [7]	Uses DCGAN for augmentation and ViT for brain tumor detection.	Brain tumor MRI dataset	Accuracy, Loss	High accuracy, reduces overfitting	Computationally intensive, requires tuning
MC-DDPM	Pan et al., 2024 [17]	Diffusion probabilistic model combining Swin-Vnet Transformer for MRI-to-CT synthesis.	Brain, Prostate dataset	MAE, PSNR, MS-SSIM, NCC	High-quality sCT, improves radiation therapy planning	Requires accurate MRI, computationally intensive
GAN + CBAM + ECA	Bandaru et al., 2024 [3]	Combines GANs with CBAM and ECA attention for liver tumor classification.	Duke Liver Dataset	Accuracy, Precision, Recall	High accuracy, enhanced feature focus	Needs validation on other datasets, high computational cost
MSG-SAGAN	Saad et al., 2022 [22]	Attention-guided multi-scale GAN for biomedical X-ray synthesis.	COVID-19 dataset	MS-SSIM, FID	Enhances image diversity, stabilizes GAN training	High computational cost, limited dataset variety
CALF-GAN	Zhu et al., 2025 [26]	Uses multi-scale convolutional attention for cross-modal medical image synthesis.	BraTS2020 dataset	PSNR, MSE, SSIM	Captures long-range dependencies, generalizes well	Struggles with complex modality transformations
AttnGAN	Rais et al., 2023 [20]	Attention-based GAN generating high-quality images for augmentation.	BraTS2020	Accuracy	Improves CNN classification accuracy	Requires tuning, complex architecture, training stability issues

3. Methodology

CAT-VAE, a deep generative model, combines the probabilistic latent features of a VAE with the global feature-learning capabilities of a CAT. CAT-VAE aims to improve the quality of synthesized images by improving the generation process using self-attention mechanisms instead of relying solely on convolutional layers.

3.1. Variational autoencoder (VAE)

The VAE is an autoencoder variant that learns a probabilistic latent space. Instead of mapping an input x to a fixed latent representation, a VAE encodes it as a distribution.

Encoder: The encoder maps an input image x to a probabilistic latent space. It outputs parameters of a distribution $q(z \mid x)$ (typically Gaussian), including the mean $\mu(x)$ and variance $\sigma^2(x)$ of the latent variable z:

$$q(z \mid x) = \mathcal{N}(z; \mu(x), \sigma^2(x)I) \tag{1}$$

Latent Space: The latent space is sampled using the reparameterization trick to ensure the gradient can propagate back through the sampling step during training:

$$z = \mu(x) + \sigma(x) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

Decoder: The decoder generates the image from the sampled latent vector z. The decoder is typically a fully connected or CNN that generates the image $\hat{x}$.

Loss: The VAE is trained using the ELBO (Evidence Lower Bound) objective, which consists of:

- **Reconstruction loss:** Measures the difference between the original image x and the generated image $\hat{x}$.
- **KL divergence:** Ensures that the learned latent space distribution $q(z \mid x)$ stays close to the prior distribution $p(z)$, typically standard Gaussian.

$$L_{\text{VAE}} = \mathbb{E}_{q(z|x)} \left[\log p(x \mid z)\right] - D_{\text{KL}}(q(z \mid x) \parallel p(z)) \tag{3}$$

3.2. Cross-Attention Transformer (CAT)

CAT is a mechanism used to enhance feature learning by allowing different feature maps to attend to each other. It is based on the mechanism used in Transformers.

Attention Mechanism: Given an input feature map $X \in \mathbb{R}^{N \times d}$ (where N is the number of spatial locations, and d is the feature dimension).

Let Q, K, and V be the query, key, and value matrices of an input feature map. These are derived as follows:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{4}$$

Where:

- $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices,
- d_k is the attention dimension.

The attention scores are computed as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{5}$$

The output for self-attention is:

$$SA(X) = AV \tag{6}$$

Where A is the attention matrix that allows each spatial location in the decoder's feature map to attend to every other location, enabling global feature refinement.

Cross-Attention: Q comes from one input, while$K$, and V come from another.

Multi-Head Attention (MHA): Instead of using a single attention mechanism, MHA applies multiple attention layers in parallel:

$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}\left(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h\right) W_O \tag{7}$$

Where each attention head computes:

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{8}$$

This parallel processing allows the model to capture diverse and complementary relationships between spatial regions.

3.3. Integrating Cross-Attention into VAE

In a standard VAE, the decoder takes a fixed latent vector z and generates a reconstruction of the input image. However, in CAT-VAE, cross-attention is introduced into the decoder. This allows the latent vector z to dynamically interact with spatial feature representations from the encoder, leading to more refined and informative image generation

Latent Vector as Query: The latent vector z, which is the compressed representation of the image, is passed through a linear transformation to become the query (Q):

$$Q = W_Q z \tag{9}$$

where:

- $z \in \mathbb{R}^{d_z}$ is the latent vector.
- $W_Q \in \mathbb{R}^{d_z \times d_k}$ is a learnable weight matrix that projects z into the query space.

Encoder Feature Map as Key (K) and Value (V): The encoder processes the input image and outputs a feature map F, which contains spatial information about the input. This feature map is used to compute both the key and value matrices:

$$K = W_K F, \quad V = W_V F \tag{10}$$

where:

- $F \in \mathbb{R}^{N \times d_f}$ (where N is the number of spatial locations and d_f is the feature dimension).
- $W_K, W_V \in \mathbb{R}^{d_f \times d_k}$ are learnable projections for keys and values.

Cross-Attention Mechanism: The attention mechanism computes the similarity between the latent query Q and the encoder keys K. It uses this to compute a weighted sum over the values V, producing the attended feature:

$$\text{Attended Feature} = \text{Cross-Attn}(Q,K,V) \tag{11}$$

This allows the decoder to selectively focus on important spatial regions of the encoder's feature map, guided by the latent vector z.

MHA: The MHA in CAT-VAE acts as a refinement step that enables the latent space to interact with itself dynamically.

Image Generation in the Decoder: The attended features produced by the cross-attention mechanism are passed through the decoder's layers, progressively reconstructing the output image:

$$\hat{x} = \text{Decoder}(\text{Cross-Attn}(Q,K,V)) \tag{12}$$

Algorithm 1 CAT-VAE Model (Cross-Attention Transformer Integration)

function BUILD_ENCODER(img_shape, latent_dim)
    Standard VAE Encoder:
    inputs ← Input(shape=img_shape)
    x ← Conv2D(filters)(inputs) → Flatten → Dense(units)
    z_mean ← Dense(latent_dim)(x)
    z_log_var ← Dense(latent_dim)(x)
    z ← Lambda(sampling)([z_mean, z_log_var])
    CAT-Specific:
    encoder_features ← Conv2D(filters)(x) ▷ Intermediate features for cross-attention
    return Model(inputs, [z_mean, z_log_var, z, encoder_features])
end function
function CROSS_ATTENTION_BLOCK(query, key, value)
    CAT Core Operation:
    attn ← MultiHeadAttention(query=query, key=key, value=value) ▷ Cross-attention
    out1 ← AddNorm(attn + query) ▷ Residual + LayerNorm
    ffn ← FeedForward(out1) ▷ Position-wise FFN
    out2 ← AddNorm(ffn + out1) ▷ Another residual + LayerNorm
    return out2
end function
function BUILD_DECODER(latent_dim)
    CAT-Enhanced Decoder:
    latent_inputs ← Input(shape=(latent_dim,))
    encoder_features ← Input(shape=(...)) ▷ From encoder (e.g., (32, 32, 64))
    x ← Dense(...) → Reshape ▷ Project latent vector
    CAT Integration Point:
    x ← Reshape((...,))(x) ▷ Flatten for attention (query)
    encoder_flat ← Reshape((...,))(encoder_features) ▷ Flatten for attention (key/value)
    x ← cross_attention_block(x, encoder_flat, encoder_flat) ▷ Apply CAT
    x ← Reshape((...))(x) ▷ Restore spatial dims
    Standard Upsampling:
    x ← Conv2DTranspose(...)(x)
    return Model([latent_inputs, encoder_features], x)
end function
function BUILD_VAE(img_shape, latent_dim)
    CAT-VAE Assembly:
    inputs ← Input(shape=img_shape)
    encoder ← build_encoder(img_shape, latent_dim)
    $z\_mean, z\_log\_var, z, encoder\_features$ ← encoder(inputs)
    CAT Critical Link:
    decoder ← build_decoder(latent_dim) ▷ Decoder uses encoder_features
    outputs ← decoder([z, encoder_features]) ▷ CAT connects here
    outputs ← VAELossLayer()([inputs, outputs, z_mean, z_log_var])
    return Model(inputs, outputs)
end function
function VAE_LOSS(inputs, outputs, z_mean, z_log_var)
    Standard VAE Loss: ▷ CAT does not modify loss
    recon_loss ← binary_crossentropy(inputs, outputs)
    kl_loss ← $-0.5 * sum(1 + z\_log\_var - z\_mean^2 - \exp(\text{z\_log\_var}))$; total_loss ← recon_loss + kl_loss
    return total_loss
end function

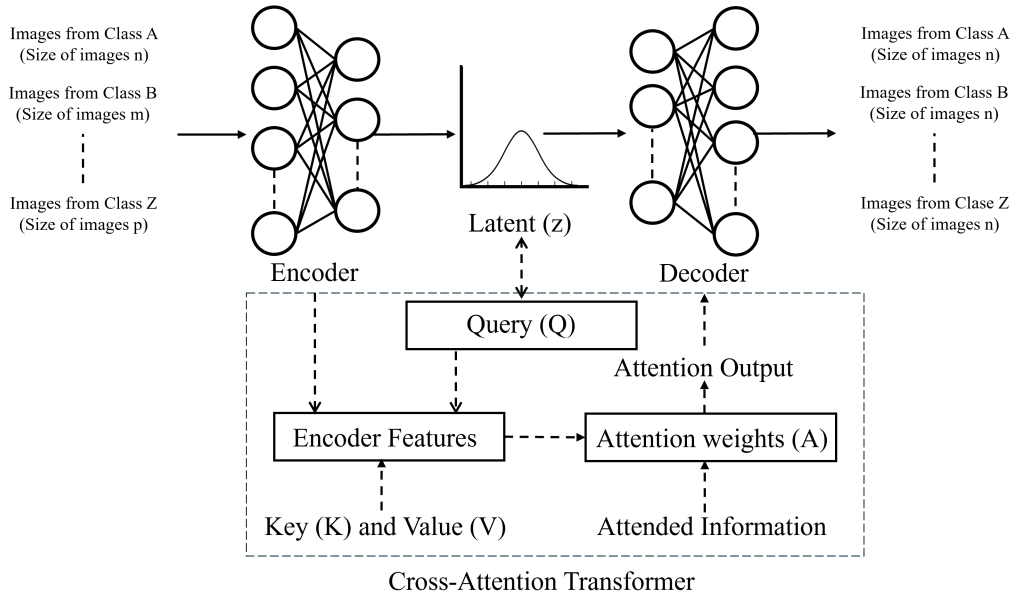


Figure 1. Architecture of the CAT-VAE framework

Figure 1 presents the architecture of the CAT-VAE framework, where the query (Q) is the latent space, and the key K and the value V are the features extracted by the encoder. The attention mechanism calculates the attention weights (A), which determines the importance of each part of the encoder features. Its output is the result of applying attention weights A to the value V, which represents the encoder features. This process highlights the most relevant parts of the features for decoding. The cross-latent information is the input to the decoder. Our framework is designed for imbalanced classes, where images and their labels are transferred from the encoder to the decoder, generating fake images.

3.4. Dataset

The CAT-VAE algorithm was evaluated on two datasets: the Breast Cancer Ultrasound and the Breast Cancer MRI datasets.

The Breast Ultrasound Dataset consists of ultrasound images collected in 2018, classified into three categories: benign, malignant, or normal. Each image is a PNG file of dimensions 500×500 pixels, accompanied by ground truth labels. The dataset can be used for three machine learning tasks: classification, detection, and segmentation of breast cancer.

The BRATS20 dataset ([1],[2],[16]) consists of 3D multimodal MRI scans (T1, T1ce, T2, and FLAIR) and corresponding tumor segmentation masks. For this study, we only used the FLAIR modality. We extracted 2D slices in the axial, sagittal, and coronal planes, classifying each as 'tumor' (non-zero mask) or 'NoTumor' (zero mask). Non-informative slices (e.g., those with black backgrounds) were discarded to maintain data quality. Following preprocessing, we preserved 1,107 high-quality FLAIR images, where 830 were 'tumor' and 277 were 'NoTumor'.

3.5. Model Architecture

The proposed framework is implemented using the TensorFlow/Keras library and incorporates several custom layers and modules to facilitate efficient training and evaluation of the VAE model. The key components are as follows:

- **CrossAttentionTransformerBlock**: A custom module that implements multi-head cross-attention, layer normalization, and a feed-forward neural network to enhance feature representation through attention mechanisms.
- **VAELossLayer**: A specialized layer that encapsulates the Evidence Lower Bound (ELBO) loss computation, balancing the reconstruction loss and the Kullback-Leibler divergence.
- **Encoder**: A modular subnetwork that outputs the latent mean, log-variance, and a reparameterized sample. It also returns intermediate encoder features used for cross-attention in the decoder.
- **Decoder**: Designed to accept both the latent vector and the encoder-derived feature maps as dual inputs, enabling feature-aligned reconstruction through cross-attention.
- **Full VAE Model**: Integrates the encoder and decoder into a unified VAE architecture with a custom loss layer, supporting end-to-end training.

Figure 2 presents the comprehensive system architecture diagram, which illustrates:
Encoder: The encoder maps an input image of size $128 \times 128 \times 3$ to a probabilistic latent space. It consists of:

- Two convolutional layers:
  - `Conv2D(32, kernel_size=3, strides=2)`
  - `Conv2D(64, kernel_size=3, strides=2)`
- Output feature map: shape $32 \times 32 \times 64$ (retained for cross-attention in decoder)
- After flattening, the feature map is reshaped to a vector of size $(\text{None}, 32768)$
- Latent variables:
  - Mean: $\mathbf{z}_{\text{mean}} = \texttt{Dense}(100)$
  - Log-variance: $\log \mathbf{z}_{\text{var}} = \texttt{Dense}(100)$
- Sampling using the reparameterization trick:

$$\mathbf{z} = \mathbf{z}_{\text{mean}} + \exp\left(\frac{1}{2}\log \mathbf{z}_{\text{var}}\right) \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{13}$$

Decoder with CAT Block:

- **Encoder feature map:**
  - Input shape: $(32, 32, 64)$
  - Reshaped to $(1024, 64)$ (used as key/value)
- **Latent vector processing:**
  - `Dense`$(16 \times 16 \times 64)$, reshaped to $(16, 16, 64)$
  - Reshaped to sequence of shape $(256, 64)$ (used as query)
- **CAT Block:**
  - MHA: `num_heads=8`, `key_dim=64`
  - Layer normalization and dropout (`rate=0.1`)
  - Feed-forward network with ReLU activation

- **Decoder Upsampling:**
  - `Conv2DTranspose(64, 3, strides=2)`
  - `Conv2DTranspose(32, 3, strides=2)`
  - `Conv2DTranspose(3, 3, strides=2)` with sigmoid activation (final output shape: $128 \times 128 \times 3$)

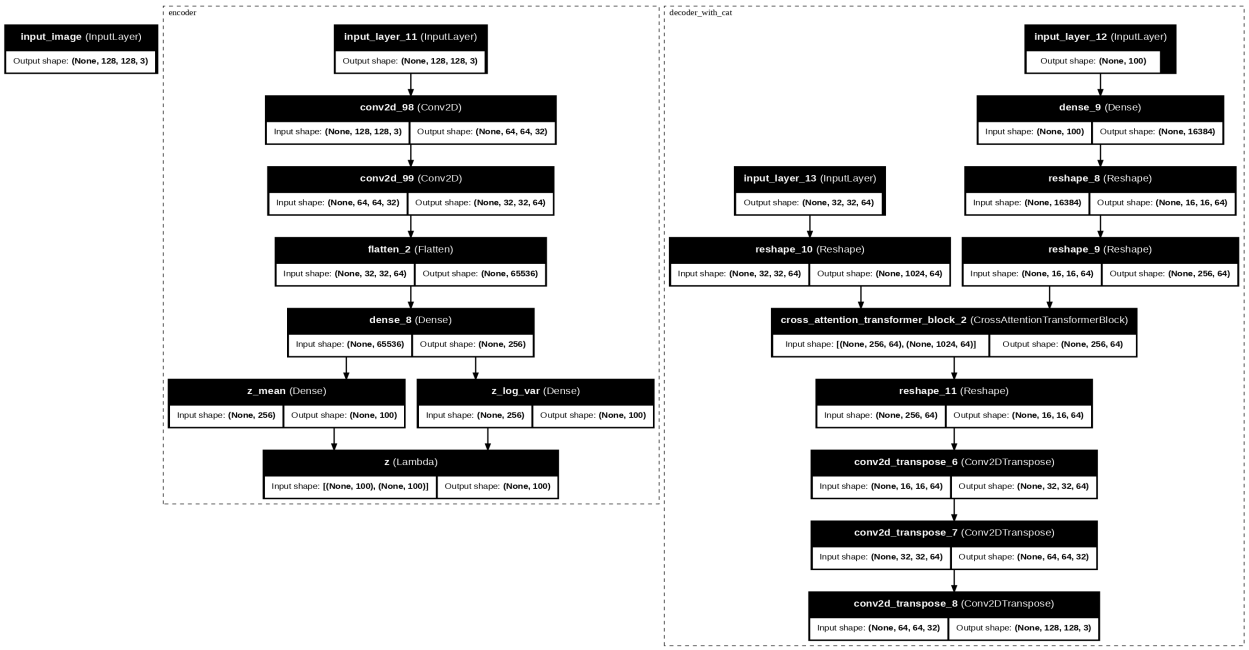


Figure 2. CAT-VAE Full System Architecture

The CAT block consists of the following components, as show in Figure 3, the first part of the CAT architecture is MHA (8 heads), which allows it to attend to relevant features. Each attention head independently looks at the input sequence with scaled dot-product attention, and concatenates all heads back into a single tensor as outputs. The block employed standard transformer techniques, including residual connections around both the attention and feed-forward layers, layer normalization for stable training, and dropout for regularization. The block included a two-layer feed-forward network (64→256→64 dimensions, with ReLU) to model nonlinear relationships. The output of the block has the same 256×64 dimensions as the input, allowing direct addition to subsequent transposed convolutional layers in the decoder.

MHA:

- Projects Query/Key/Value into 8 subspaces:
  - Query: Processed latent representation (256x64)
  - Key/Value: Encoder features (1024x64)
  - Output: Attention-weighted features (256x64)
- Computes scaled dot-product attention for each head (Equation 5).
- Concatenates outputs from all heads 256×64

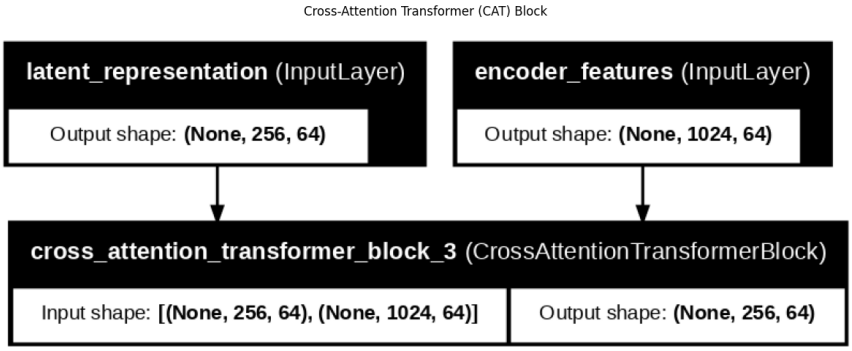


Figure 3. CAT Block architecture

To generate new images the Latent vectors are sampled from a standard normal distribution, the encoder features are averaged over real data, repeated across batches, and generated images are saved as .png. Models are saved in HDF5 format for reuse. Table 2 presents the hyperparameters used in our model.

Table 2. Hyperparameters details for the model.

Parameter	Value / Description
Image Shape	$128 \times 128 \times 3$
Input Image Dimensions	128×128 pixels with 3 color channels
Latent Dim	100
Size of Latent Space	100-dimensional latent vector
Batch Size	64
Epochs	500
Num Heads	8
Key Dim	64
Dropout Rate	0.1
Optimizer	Adam
Activation	ReLU
Final Activation	Sigmoid
FID Model	InceptionV3

3.6. Evaluation Metrics and Performance Tracking

To comprehensively evaluate the performance of the proposed CAT-VAE model, we measure both quantitative generation quality and computational efficiency (Table 3).

1. Reconstruction Quality Metrics

These metrics assess how well the model reconstructs input images during validation/testing:

a) **Mean Squared Error (MSE)**
Measures pixel-wise differences between the original and reconstructed images:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{14}$$

where x_i is the original image and $\hat{x}_i$ is the reconstructed version.

b) **Peak Signal-to-Noise Ratio (PSNR)**
Quantifies the ratio between the maximum possible pixel value and the MSE, expressed in decibels (dB):

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\max(x)^2}{\text{MSE}} \right) \tag{15}$$

Higher values indicate better fidelity.

c) **Structural Similarity Index Measure (SSIM)**
Evaluates structural similarity between original and reconstructed images, considering luminance, contrast, and structure:

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{16}$$

where μ and σ are the mean and standard deviation, and C_1, C_2 are small constants to avoid division by zero. SSIM ranges from -1 to 1, with values closer to 1 indicating higher similarity.

2. Generation Quality Metric

This metric evaluates the perceptual realism of generated synthetic images compared to real ones:

d **Fréchet Inception Distance (FID)**
Measures the distance between feature distributions of real and generated images using an InceptionV3 model:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{17}$$

where:

- μ_r, μ_g: Mean feature vectors of real and generated images.
- Σ_r, Σ_g: Covariance matrices of the features.
- Tr: Matrix trace operator.

Lower FID scores indicate higher similarity between real and generated images.

3. Computational Performance Metrics

These metrics assess the model's computational efficiency and resource usage:

e) **Training Time**
Total time taken to complete training (in seconds), measured per epoch and averaged across all epochs.

f) **Inference Time**
Time required to generate one output image (in milliseconds), computed as the average over batches during test-time prediction.

g) **GPU Memory Usage**
Peak GPU memory consumption during training (in megabytes). Measured using the `pynvml` library if available; otherwise omitted.

Table 3. Overview of evaluation metrics used for assessing model performance.

Metric	Description	Range / Unit	Desired Direction
MSE	Pixel-wise reconstruction error	$[0, \infty)$	Lower is better
PSNR	Reconstruction signal quality	$[0, \infty)$ dB	Higher is better
SSIM	Structural similarity index	$[-1, 1]$	Closer to 1 is better
FID	Realism of generated images	$[0, \infty)$	Lower is better
Training Time	Duration of training	Seconds	Faster is better
Inference Time	Speed of image generation	ms/image	Faster is better
GPU Memory Usage	Resource consumption during training	MB	Lower is better

4. Results and comparaisin

To evaluate the effectiveness of CAT-VAE in generating high-quality medical images, we compare its performance with a traditional VAE.

The key analysis of the baseline VAE and the proposed CAT-VAE model is summarized in Table 4 and 5. The models were evaluated using quantitative metrics (SSIM, PSNR, FID, and MSE) as well as dependent computational efficiency elements (training time, inference time, and memory on GPU), with two medical imaging datasets, include Breast Cancer Ultrasound and Brain Tumor MRI.

On the breast cancer ultrasound dataset (Table 4), CAT-VAE significantly outperformed VAE in terms of image generation quality. For example, the SSIM values improved from 0.4433 to 0.5924 for benign images, from 0.6291 to 0.6663 for malignant images, and from 0.5794 to 0.6914 for normal images. This indicates that CAT-VAE preserves structural details more effectively than VAE.

Similarly, PSNR values of generated images increased across each class, indicating a noise reduction. The lower FID score indicates that CAT-VAE's synthetic images were of higher quality than VAE's for benign cases. In both benign and malignant classes, MSE values decreased, suggesting a reduction in pixel-level reconstruction errors; however, increased resource utilization was a consequence. The benign class took 546.89 seconds to train with CAT-VAE, much longer than VAE's 143.18 seconds. Inference time for CAT-VAE also doubled, and the GPU memory requirements of CAT-VAE were greater than those of VAE.

Despite the increase in resource utilization, the higher quality and the improved generation from CAT-VAE make it particularly useful for medical image analysis.

Table 4. Breast Dataset Model Performance Metrics

Model	Class	SSIM	PSNR	FID	MSE	Training Time	Inference Time	GPU Memory Usage
VAE	Benign	0.4433	21.34	39.19	0.0075	143.18 sec	132.00 ms per image	2433 MB
	Malignant	0.6291	24.64	23.57	0.0036	94.13 sec	131.84 ms per image	2433 MB
	Normal	0.5794	23.32	19.49	0.0047	79.84 sec	263.00 ms per image	2431 MB
CAT-VAE	Benign	0.5924	23.51	32.66	0.0045	546.89 sec	259.58 ms per image	8581 MB
	Malignant	0.6663	25.23	24.47	0.0031	283.72 sec	259.11 ms per image	8583 MB
	Normal	0.6914	25.13	17.93	0.0031	173.32 sec	260.24 ms per image	8577 MB

CAT-VAE had comparatively higher SSIM scores with the brain tumor MRI dataset (Table 5) that increased from 0.8000 to 0.8264 for tumor images and from 0.7411 to 0.8357 for non-tumor images. This increase indicates that CAT-VAE is better at articulating structures such as areas of complex anatomy.

The PSNR metric followed a similar trend, with CAT-VAE performing a strong 27.67 dB for tumor images and 28.15 dB for non-tumor images, whereas VAE produced only a 26.60 dB and 25.26 dB performance, respectively. The PSNR results handle ambiguity better to say that CAT-VAE provides clearer and less noisy images.

Probably the most pronounced improvement was in the FID scores. CAT-VAE achieved a nearly 26-point decrease in FID for the non-tumor class, from 68.67 to 43.09. The reduction signifies an improvement in the perceived realism of the generated images. Similarly, MSE values were lower overall for CAT-VAE, implying an order of improvement in pixel-wise accuracy.

As expected, CAT-VAE again required more computational resources. Training time increased significantly—for example, from 264.70 seconds to 1030.17 seconds for tumor images. Inference time remained relatively stable across both models, but GPU memory usage was nearly double compared to VAE.

Table 5. Brain Dataset Model Performance Metrics

Model	Class	SSIM	PSNR	FID	MSE	Training Time	Inference Time	GPU Memory Usage
VAE	Tumor	0.8000	26.60	72.68	0.0023	264.70 sec	263.41 ms per image	4481 MB
	NoTumor	0.7411	25.26	68.67	0.0031	119.34 sec	113.27 ms per image	4483 MB
CAT-VAE	Tumor	0.8264	27.67	62.60	0.0017	1030.17 sec	260.10 ms per image	8579 MB
	NoTumor	0.8357	28.15	43.09	0.0016	306.76 sec	259.09 ms per image	8581 MB

The classification performance was evaluated by means of a Breast Ultrasound Dataset, composed of three classes: Normal, Benign, and Malignant. As shown in Table 6, CAT-VAE augmentation enhances classification accuracy compared to no augmentation and standard VAE augmentation. When no data augmentation (DA) was applied, the overall accuracy was 67.31%. A high loss (2.9162) was observed, while precision, recall, and F1-scores were moderate across classes, with a precision of 0.59 for Normal and a recall of 0.85 for Benign. The model successfully identified true positives, further improvements are needed to reduce false positives.

In contrast, when data augmentation was performed using a VAE, the accuracy improved to 84.35%, with reduced loss (0.7018). Precision and recall both increased, especially for the Malignant class, with precision of 0.94 and recall of 0.83, showing strong detection of cancerous cases.

The CAT-VAE augmentation method achieved the best performance. The model achieved 82.9% accuracy with a loss of 0.268 using 437 images. With 2000 images, the model reached 97.5% accuracy with a loss of 0.1676. The results indicate that CAT-VAE produces high-quality, varied synthetic samples that enhance model generalization and classification.

The confusion matrices (Figure 4) provide more evidence of the improvements. Without the use of data augmentation, the Benign and Malignant cases created misclassifications impacted the model and limited the reliability of the diagnosis.

The pattern seen within the CAT-VAE data augmentation showed a near-perfect categorical classification after model training on 2000 images, 16 Normal images incorrectly classified as either Benign or Malignant; 8 Benign images were misclassified as Malignant and 1 as Normal, and 5 Malignant were misclassified as Benign. The performance achieved by training the model on the described data augmentation is critically important related to medical diagnosis, where the prevalence of false negatives (missed cancers) and false positives (misdiagnosed healthy patients) should be minimized.

The training and validation curves (6) provide context for the iterative learning of the model. Specifically, during the best performing configuration (using the CAT-VAE DA) combined with 2000 images, Training Accuracy

starts around ∼0.6, and quickly improves to ∼1.0 after epoch 20. Validation Accuracy eventually achieved ∼0.96, through progressive improvement, and demonstrated a positive level of generalization. Training Loss saw a transition from ∼1.00 down to ¡ 0.1, signalling rapid convergence. Validation loss exhibited a similar trajectory, settling around ∼0.15, and there was minimal evidence of overfitting.

Table 6. Performance on Breast Dataset (Epoch = 20)

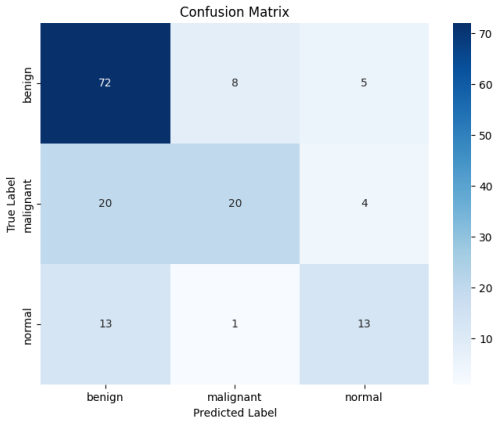
Method	Class	Total	Real	Synthetic	Precision	Recall	F1	Accuracy	Loss
Without DA	Normal	133	133	0	0.59	0.48	0.53	67.31%	2.9162
	Benign	437	437	0	0.69	0.85	0.76		
	Malignant	210	210	0	0.69	0.45	0.55		
DA with VAE	Normal	437	133	304	0.86	0.88	0.87	84.35%	0.7018
	Benign	437	437	0	0.74	0.81	0.78		
	Malignant	437	210	227	0.94	0.83	0.88		
DA with CAT-VAE	Normal	437	133	304	0.95	0.85	0.90	85.11%	0.8320
	Benign	437	437	0	0.79	0.78	0.79		
	Malignant	437	210	227	0.81	0.91	0.86		
DA with CAT-VAE	Normal	1000	133	867	0.86	0.88	0.87	84.35%	0.7018
	Benign	1000	437	563	0.74	0.81	0.78		
	Malignant	1000	210	790	0.94	0.83	0.88		
DA with CAT-VAE	Normal	2000	133	1867	1.00	0.96	0.98	97.50%	0.1167
	Benign	2000	437	1563	0.96	0.98	0.97		
	Malignant	2000	210	1790	0.97	0.99	0.98		

Table 7 summarizes the performance of the model on the Brain Dataset (BRATS20 MRI) for two classes: Tumor and NoTumor. Without data augmentation, the CNN gave an accuracy of 79.64% and a loss of 0.8145, showing moderate performance, but it struggles with the NoTumor class, particularly in terms of precision (0.55) and recall (0.43). The incorporation of VAE-based data augmentation has demonstrated a marked performance improvement across both classes of observations, particularly in the NoTumor class, and there was a notable improvement in precision (0.84) and recall (0.80). CAT-VAE ranked highest, achieved the highest accuracy (93.62%), and the lowest loss (0.2579) when trained on 2000 images. The Tumor class saw particularly significant improvements in recall and F1-scores.

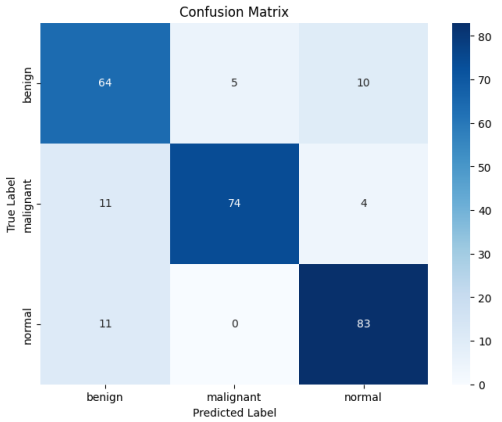
The confusion matrices (Figure5) provide important insights into each model's classification behavior. Without data augmentation, the model did not distinguish well between the two classes, producing misclassifications of 17 Tumor cases as NoTumor and 28 NoTumor cases as Tumor, suggesting significant confusion in classifying NoTumor cases. With VAE-based Data Augmentation (830 imgaes), the model performed better than the baseline, with misclassification of 23 Tumor cases as NoTumor and 30 NoTumor cases as Tumor. When CAT-VAE synthetic images used (830 images), misclassification was further improved to 22 Tumor cases as NoTumor, and 27 NoTumor cases as Tumor. Once the number of synthetic images was increased to 1000, the model misclassified 19 Tumor cases as NoTumor and 29 NoTumor cases as Tumor. When the final size of 2000 synthetic images was included, the model performed even better, misclassifying 29 Tumor cases, and 22 NoTumor cases as Tumor.

The training and validation curves (Figure 7) also demonstrate the learning behavior of the model:

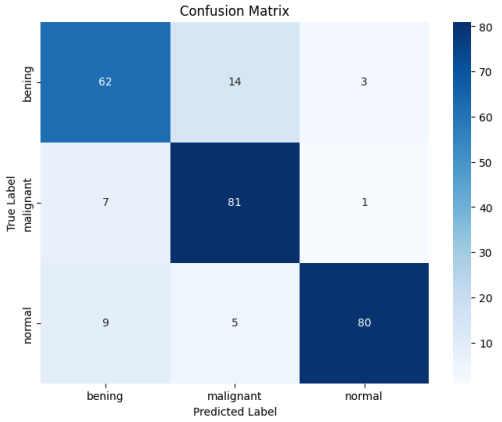
- Training accuracy tends to monotonically increase and approaches perfect values (∼1.0), whereas training loss tends to monotonically decrease, reaching a stable low value. This indicates the model is successfully optimizing.
- Validation accuracy also increases monotonically, but remains slightly below that of the training accuracy, which is expected due to the noise and unpredictability of real-world data. Validation loss tends to also decrease initially, but it displays minor fluctuations, and overall demonstrates stable generalization.
- Despite small differences between training and validation accuracy and loss, which suggests only moderate overfitting, the evidence of stable validation loss indicates the model generalizes very well to unseen data.



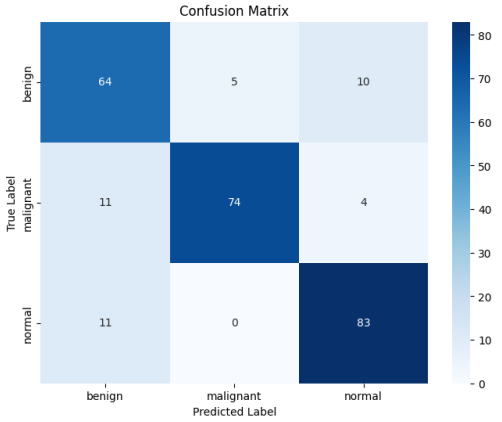
(a) No Augmentation



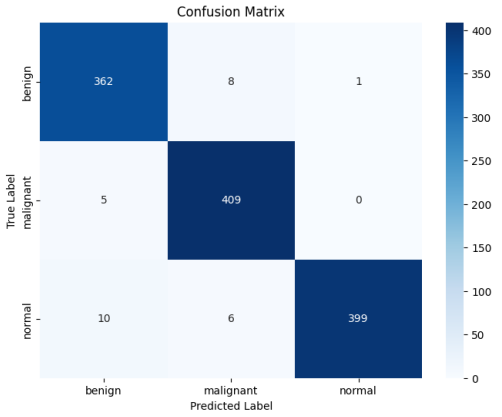
(b) VAE Augmentation



(c) CAT-VAE (Case 1)

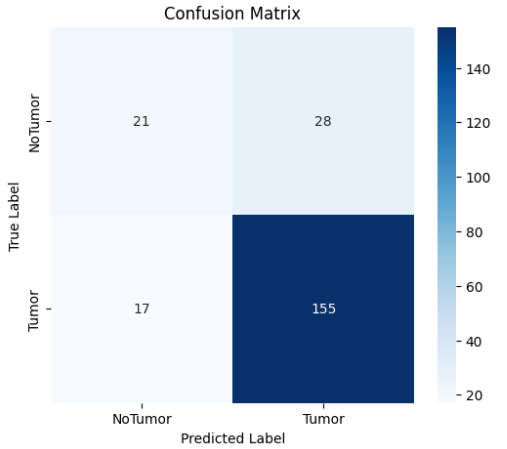


(d) CAT-VAE (Case 2)

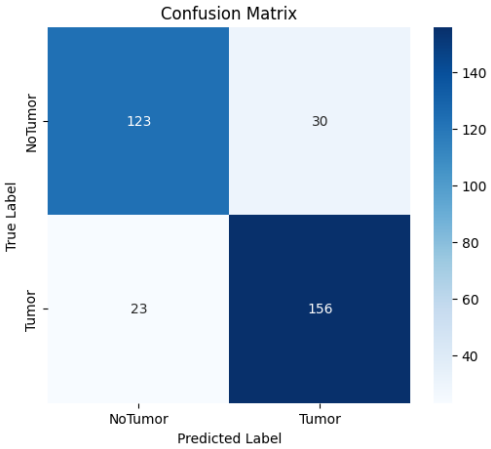


(e) CAT-VAE (Case 3)

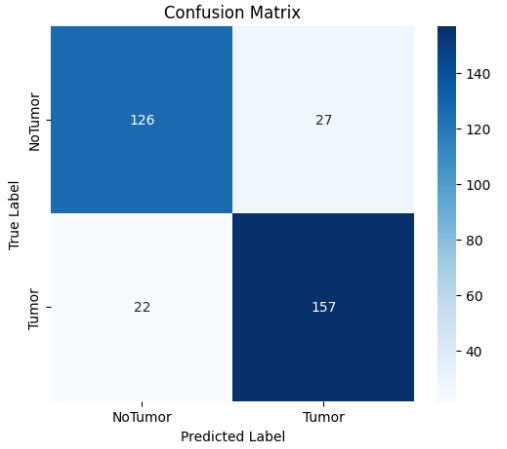
Figure 4. Comparison of confusion matrices for different data augmentation strategies using the Ultrasound breast cancer dataset.



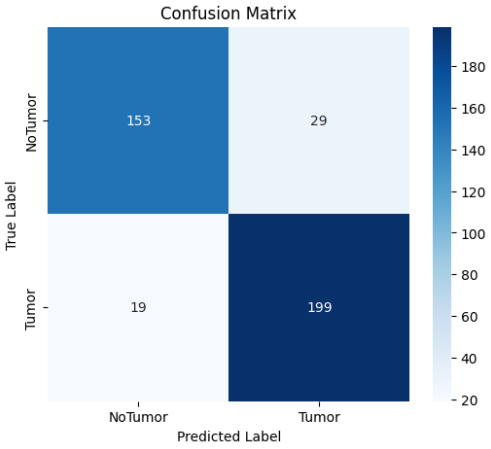
(a) No Augmentation



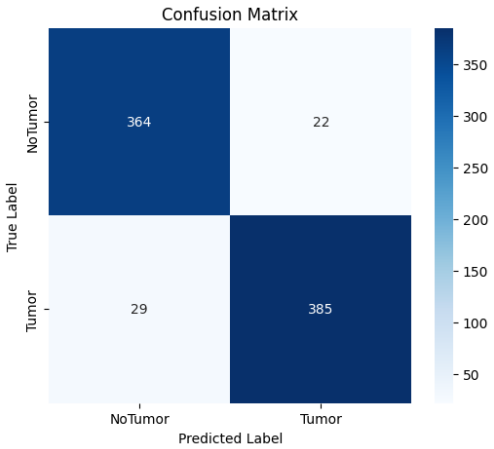
(b) VAE Augmentation



(c) CAT-VAE (Case 1)

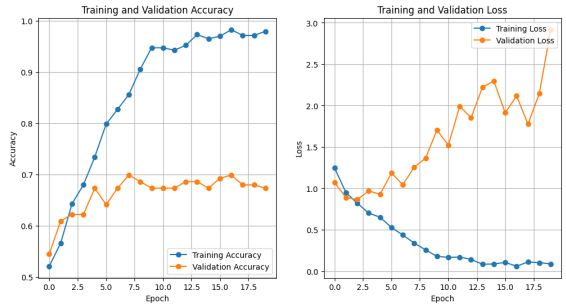


(d) CAT-VAE (Case 2)

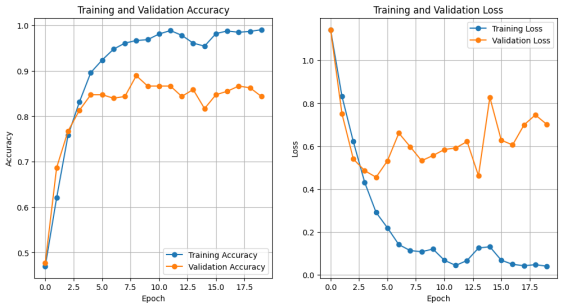


(e) CAT-VAE (Case 3)

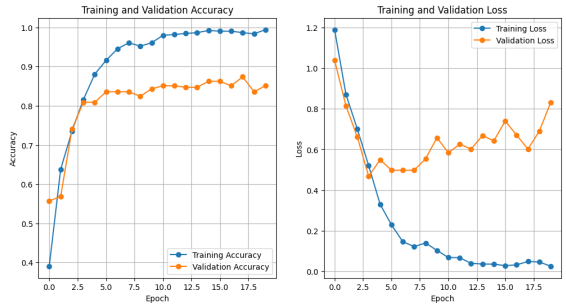
Figure 5. Comparison of confusion matrices for different data augmentation strategies using the BraTS 2020 dataset.



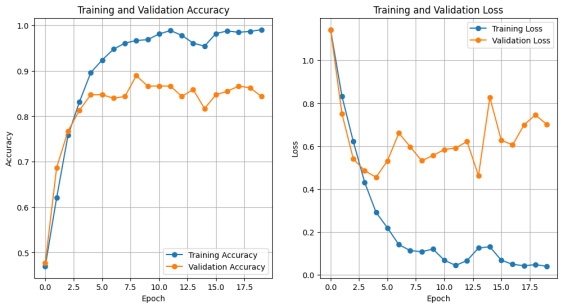
(a) No Augmentation



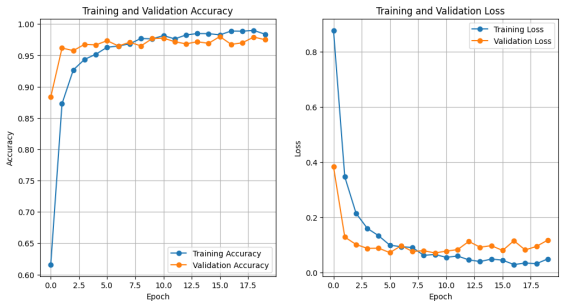
(b) VAE Augmentation



(c) CAT-VAE (Case 1)

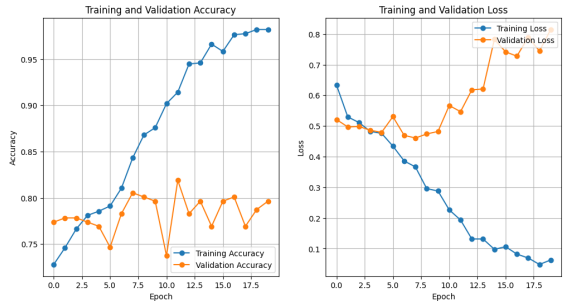


(d) CAT-VAE (Case 2)

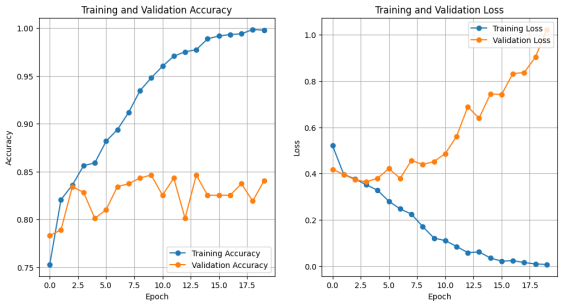


(e) CAT-VAE (Case 3)

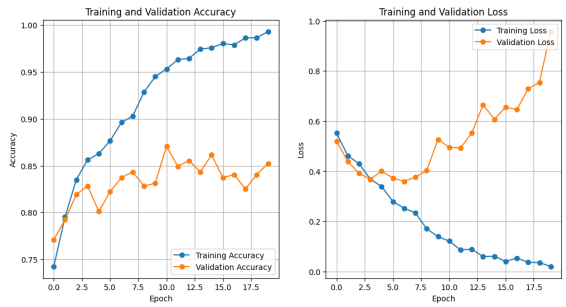
Figure 6. Accuracy and loss curves (training and validation) for different data augmentation strategies using the Urtarsound breast cancer dataset. Each subplot includes both training and validation curves.



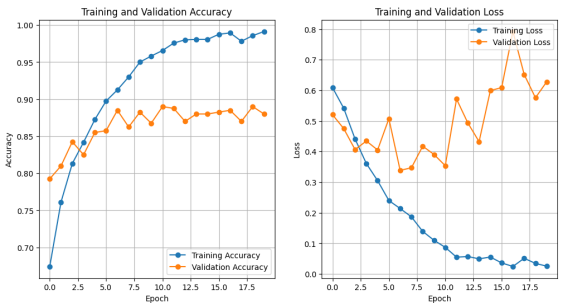
(a) No Augmentation



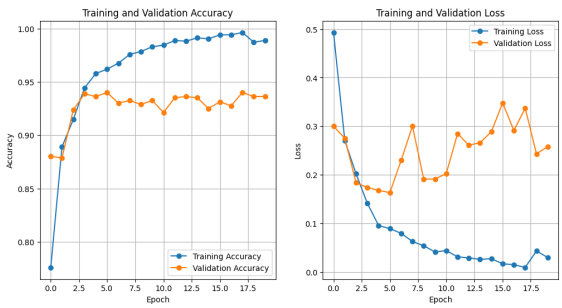
(b) VAE Augmentation



(c) CAT-VAE (Case 1)



(d) CAT-VAE (Case 2)



(e) CAT-VAE (Case 3)

Figure 7. Accuracy and loss curves (training and validation) for different data augmentation strategies using the BraTS 2020 dataset. Each subplot includes both training and validation curves.

Table 7. Performance on Brain Dataset (Epoch = 20)

Method	**Class**	**Total**	**Real**	**Synthetic**	**Precision**	**Recall**	**F1**	**Accuracy**	**Loss**
Without DA	Tumor	830	830	0	0.85	0.90	0.87	79.64%	0.8145
	NoTumor	277	277	0	0.55	0.43	0.48		
DA with VAE	Tumor	830	830	0	0.84	0.87	0.85	84.04%	1.0233
	NoTumor	830	277	553	0.84	0.80	0.82		
DA with CAT-VAE	Tumor	830	830	0	0.85	0.88	0.87	85.24%	0.9574
	NoTumor	830	277	553	0.85	0.82	0.84		
DA with CAT-VAE	Tumor	1000	830	170	0.87	0.91	0.89	88.00%	0.6283
	NoTumor	1000	277	723	0.89	0.84	0.86		
DA with CAT-VAE	Tumor	2000	830	1170	0.95	0.93	0.94	93.62%	0.2579
	NoTumor	2000	277	1723	0.93	0.94	0.93		

5. Discussion

CAT-VAE outperforms conventional VAE using Breast Cancer Ultrasound and Brain Tumor MRI datasets. Results across the evaluation criteria, SSIM, PSNR, FID, and MSE, showed that CAT-VAE consistently performed better than VAE for both structural and aesthetic image quality. Most importantly, CAT-VAE generated images maintained important anatomical details with acceptable fidelity, which may greatly impact a medical diagnosis, in addition to taking part in later classification tasks.

The observed improvement in SSIM and PSNR across both datasets confirms that CAT-VAE is harnessing local and global structural information better than the regular VAE baseline model. The reason for this improvement can be attributed to the attention mechanisms from CAT-VAE, which gather information from contextually relevant areas to reconstruct the image accurately. Furthermore, the increase in SSIM for the tumor and non-tumor brain images conveys that CAT-VAE has been able to learn complicated spatial dependencies, especially in challenging modalities.

Furthermore, FID improvements, including a 25-point improvement in the non-tumor class, suggest that CAT-VAE is capable of producing realistic images that are statistically closer to the real data distribution. Additionally, the lower MSE further establishes this, demonstrating better pixel-wise accuracy and less reconstruction noise.

Quantification of these results suggests a significant increase in classification performance when using synthetic images for data augmentation. The classification experiment on the Breast Ultrasound dataset and BraTs 2020 dataset provides solid evidence of CAT-VAE's value for clinical AI capabilities. With the model trained with 2000 CAT-VAE generated images, the model was able to reach an outstanding 97.5% accuracy, with very little false positive and false negative rates when combined with the aforementioned impact on classification. In an especially high-stakes medical application where diagnostic reliability can make a huge difference, this aspect of performance is especially critical.

Confusion matrices and learning curves also provide additional evidence of the model's efficacy. CAT-VAE has reduced both false negatives and false positives, which are two dire errors found in medical diagnosis. Additionally, the training and validation curves indicate fast convergence and good generalization without overfitting when adequately trained with augmented samples.

From a radiologist's perspective (Figures 8 and 9), the generated breast and brain images presented across VAE and CAT-VAE models show varying degrees of structural fidelity compared to their real counterparts. In the breast cancer images, fine-grained tissue patterns and lesion borders are notably blurred in the VAE outputs, while the CAT-VAE model restores some anatomical coherence, particularly in the central regions. Likewise, in the brain tumor dataset, samples from all four VAEs promote the smoothing of important features, which could potentially hide the existence of disease and the absence of disease. CAT-VAE images show some improvement over deep VAE visuals and would be better at preserving the asymmetry of the tumor or the layout of the brain tissue, but still exhibit most of the lost detail. Importantly, we note that it is possible for us to visually lose the original class of the image during the generative process (e.g., as in a tumor case appearing as no tumor), which is indicative of a risk of semantic loss. The important point is that we have to recognize that these images are only outputs of deep generative systems that perform processing in latent and abstract representations. Therefore, visual inspection alone cannot be reliable for evaluating the fidelity or trustworthiness of deep generative sample images for diagnostic decision making. The AI model may retain class-relevant features in ways that we cannot visually see, meaning we cannot deem these outputs as acceptable or rejected simply through human visual peculiarities.

Nevertheless, there are still some limitations to take into consideration. First, the model was only evaluated on two datasets: breast ultrasound and brain tumor MRI. This sample covers two different modalities and structural features, but not the full scale of medical imaging (e.g., CT scans, retinal images, histopathology, etc.). Validation with generalizability across other domains still remains to be determined. Next, we rely heavily on accurate labelling and segmentation, and obtaining quality annotations can be costly and time-consuming in a real clinical setting. Finally, the attention mechanism we used is fixed during training, which may also not be flexible enough to adapt across different types of and resolutions of abnormality when used in more heterogeneous datasets.

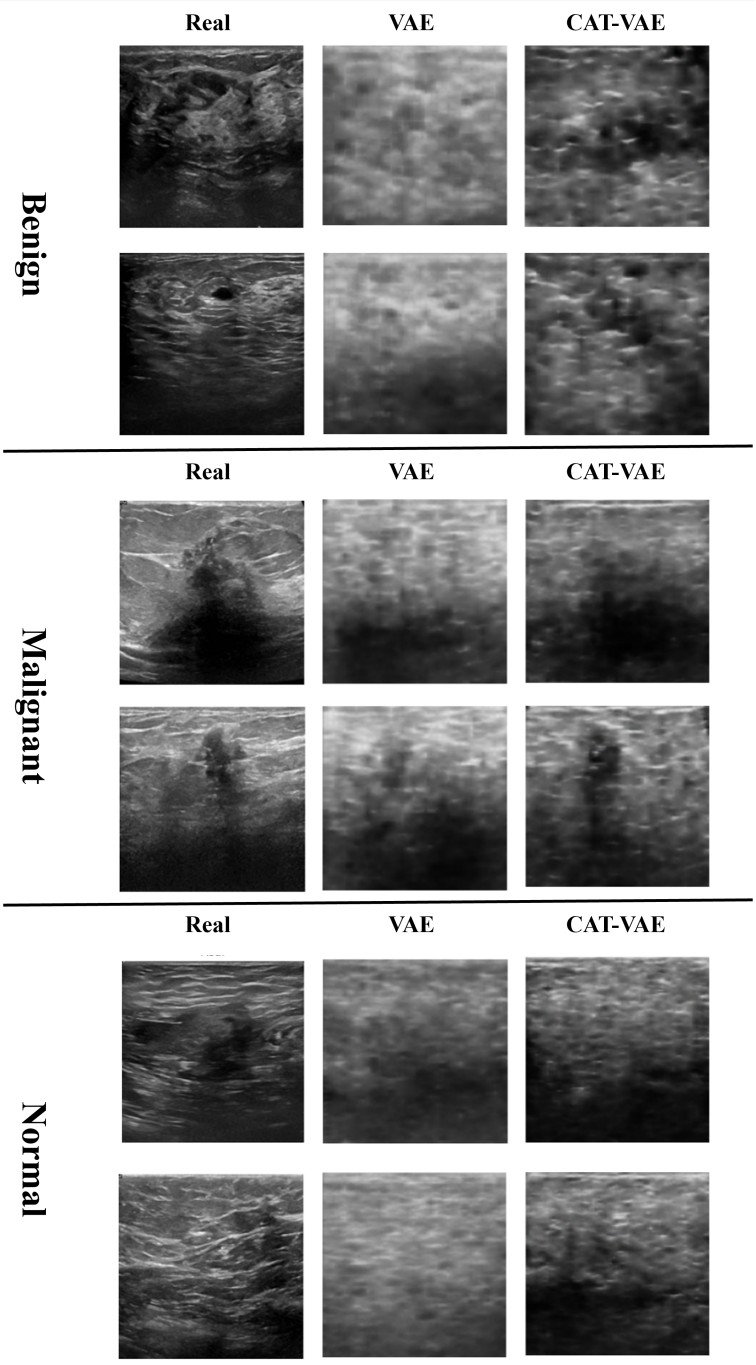


Figure 8. Comparative visual analysis of Ultrasound breast samples

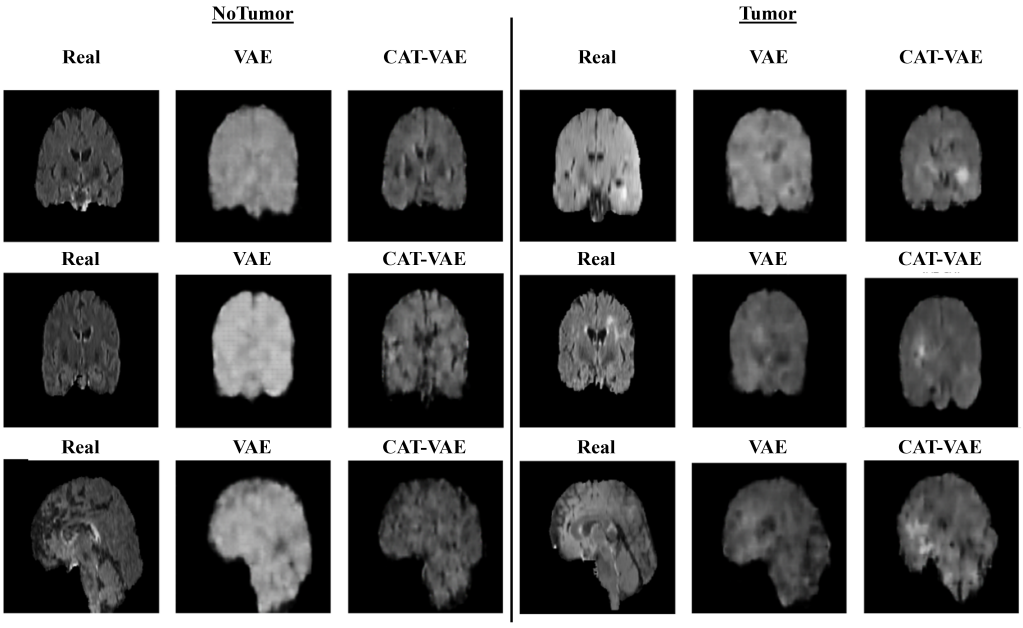


Figure 9. Comparative visual analysis of brain MRI samples

6. Conclusion

In this study, we propose a Category-Aware VAE (CAT-VAE), which improves the generation and reconstruction of medical images by incorporating class information into the latent space. We evaluated CAT-VAE in two vastly different medical imaging datasets, breast ultrasound data and MRI data of a brain tumor. The CAT-VAE consistently improved image quality and related downstream performance compared to the baseline VAE. Quantitative measures of similarity (e.g., SSIM, PSNR) and data quality (e.g., FID, MSE) indicated that CAT-VAE generated high-fidelity images while maintaining class consistencies.

Moreover, CAT-VAE generated samples were enhanced data augmentation that improved classification accuracy, particularly with small training datasets. The implications are meaningful to improve data diversity and generalizability for actual diagnostic systems in healthcare. Despite CAT-VAE being computationally cumbersome and reliant on correct categorical labels, its power in image generation and augmentation means it has potential as a novel methodology in medical imaging. Future research will focus on model efficiency, applying it to other modalities, looking at multiple datasets, and then implementing it in real-life use. In summary, CAT-VAE aimed for a partnership between generative modelling and medical use, and this has been a strong framework for image generation and intelligent diagnostic capability across low-data and imbalanced situations.

Code Availability:

The code used for training, simulation, and evaluation is publicly available at: https://zenodo.org/records/15570016

Acknowledgement

The authors have no acknowledgments to declare.

REFERENCES

1. Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
2. Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.
3. Sumash Chandra Bandaru, G Bharathi Mohan, R Prasanna Kumar, and Ali Altalbe. Swingale: fusion of swin transformer and attention mechanism for gan-augmented liver tumor classification with enhanced deep learning. *International Journal of Information Technology*, 16(8):5351–5369, 2024.
4. Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.
5. Dimitrios E Diamantis, Panagiota Gatoula, and Dimitris K Iakovidis. Endovae: Generating endoscopic images with a variational autoencoder. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.
6. Xuejie Hao, Lu Liu, Rongjin Yang, Lizeyan Yin, Le Zhang, and Xiuhong Li. A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, 15(3):827, 2023.
7. Md Momenul Haque, Subrata Kumer Paul, Rakhi Rani Paul, Nurnama Islam, Mirza AFM Rashidul Hasan, and Md Ekramul Hamid. Improving performance of a brain tumor detection on mri images using dcgan-based data augmentation and vision transformer (vit) approach. In *GANs for Data Augmentation in Healthcare*, pages 157–186. Springer, 2023.
8. Jiayu Huo, Vejay Vakharia, Chengyuan Wu, Ashwini Sharan, Andrew Ko, Sébastien Ourselin, and Rachel Sparks. Brain lesion synthesis via progressive adversarial variational auto-encoder. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 101–111. Springer, 2022.
9. R John, J Penning, H Chandler, P Fielding, C Marshall, and R Smith. Quantitative evaluation of synthesized brain pet using a variational autoencoder. In *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–4. IEEE, 2021.
10. Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: a review. *Journal of imaging*, 9(4):81, 2023.
11. Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. End-to-end autoencoding architecture for the simultaneous generation of medical images and corresponding segmentation masks. In *International Conference on Medical Imaging and Computer-Aided Diagnosis*, pages 32–40. Springer, 2023.
12. Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, 55(3):2351–2377, 2022.
13. Xiang Li, Minglei Li, Pengfei Yan, Guanyi Li, Yuchen Jiang, Hao Luo, and Shen Yin. Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, pages 93–116, 2023.
14. Jianing Liu. Research on the application of variational autoencoder in image generation. In *ITM Web of Conferences*, volume 70, page 02001. EDP Sciences, 2025.
15. Xiaofeng Liu, Fangxu Xing, Jerry L Prince, Aaron Carass, Maureen Stone, Georges El Fakhri, and Jonghye Woo. Dual-cycle constrained bijective vae-gan for tagged-to-cine magnetic resonance image synthesis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1448–1452. IEEE, 2021.
16. Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
17. Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al. Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548, 2024.
18. Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 68(10):105004, 2023.
19. Khadija Rais, Mohamed Amroune, and Mohamed Yassine Haouam. Medical image generation techniques for data augmentation: Disc-vae versus gan. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–8. IEEE, 2024.
20. Khadija Rais, Mohamed Amroune, Mohamed Yassine Haouam, and Issam Bendib. Comparative study of data augmentation approaches for improving medical image classification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1226–1234. IEEE, 2023.
21. Sara Rajaram and Cassie S Mitchell. Data augmentation with cross-modal variational autoencoders (dacmva) for cancer survival prediction. *Information*, 15(1):7, 2023.
22. Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. A self-attention guided multi-scale gradient gan for diversified x-ray image synthesis. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 18–31. Springer, 2022.
23. Manisha Saini and Seba Susan. Tackling class imbalance in computer vision: a contemporary review. *Artificial Intelligence Review*, 56(Suppl 1):1279–1335, 2023.
24. Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
25. An Zhao, Moucheng Xu, Ahmed H Shahin, Wim Wuyts, Mark G Jones, Joseph Jacob, and Daniel C Alexander. 4d vq-gan: Synthesising medical scans at any time point for personalised disease progression modelling of idiopathic pulmonary fibrosis. *arXiv preprint arXiv:2502.05713*, 2025.

26. Xinmiao Zhu and Yuan Wang. Calf-gan: Multi-scale convolutional attention for latent feature-guided cross-modality mr image synthesis. *Molecular & Cellular Biomechanics*, 22(3):1431–1431, 2025.
27. Qiankun Zuo, Hao Tian, Ruiheng Li, Jia Guo, Jianmin Hu, Long Tang, Yi Di, and Heng Kong. Hemisphere-separated cross-connectome aggregating learning via vae-gan for brain structural connectivity synthesis. *IEEE Access*, 11:48493–48505, 2023.