Variational Autoencoder-Based Synthetic Data Generation for Augmenting Mosquito Larvae Image Datasets

Muhammad Anggia Muchtar[1*], Anggi Ester Herna[2], Amer Sharif[2], Pauzi Ibrahim Nainggolan[3], Maya Silvi Lydia[2], Fahrurrozi Lubis[1], Dhani Syahputra Bukit[4], Riza Sulaiman[5]

[1] Department of Information Technology, Universitas Sumatera Utara, Medan, 20155, Indonesia
[2] Department of Computer Science, Universitas Sumatera Utara, Medan, 20155, Indonesia
[3] Computer Vision and Multimedia Laboratory, Universitas Sumatera Utara, Medan, 20155, Indonesia
[4] Department of Public Health, Universitas Sumatera Utara, Medan, 20155, Indonesia
[5] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, 43600 Malaysia
E-mail: anggi.muchtar@usu.ac.id
[*]Corresponding author

Keywords: variational autoencoder, mosquito larvae, synthetic data

Received: June 22, 2025

We propose a convolutional Variational Autoencoder (VAE) to synthesize mosquito-larvae images for dataset augmentation. The dataset comprises 1,300 grayscale microscope images (640×480) resized to 224×224 and split 1,200/100 for train/test. The encoder uses four Conv2D blocks (32, 64, 128, 256) that project to a 100-dimensional latent space; the decoder mirrors this with Conv2DTranspose layers. Training uses Adam with an MSE+KL objective, batch size 64, over 100–400 epochs. Image realism is assessed via Fréchet Inception Distance (Inception-v3 features) across latent sizes {50, 100} and epoch counts. The best configuration (latent = 100, 400 epochs) achieves FID = 0.4668 with total loss in the 29,206–33,806 range, representing a 59.7% reduction relative to a 50-D/100-epoch baseline (FID = 1.1588). Among variants, the standard VAE yields the lowest FID overall, outperforming VQ-VAE and VAE-S, while β-VAE is competitive in lower-capacity settings. These results indicate that the learned generator produces samples statistically close to the real distribution and provide a practical training recipe for generating synthetic larvae imagery to support downstream recognition tasks.

Povzetek: Kako izboljšati učenje na majhnih podatkovnih zbirkah ličink komarjev? Predlagan je konvolucijski VAE, ki iz 1.300 sivinskih mikroskopskih slik generira sintetične nove primere. Model z MSE+KL ciljem in latentnim prostorom 100 dim. doseže najboljšo kakovost.

1 Introduction

As a tropical country, Indonesia has environmental conditions that support the growth and development of various mosquito species with potential risks to human and animal health. The high population density of mosquitoes is one of the main factors contributing to the increasing cases of vector-borne diseases [15]. In this era of rapidly advancing technology, methods involving artificial intelligence have become popular across various topics, including synthetic image generation and image classification [23]. One promising approach in synthetic image generation and image classification is the use of VAE [35], which aims to produce varied image data that can assist in training classification or detection models. VAE is one generative model alternative to Generative Adversarial Network (GAN). The advantage of VAE over GAN is that VAE can work with relatively small datasets, while GAN is more prone to mode collapse due to its lack of an encoder [14].

VAE combines concepts from Autoencoder and probabilistic models to generate hidden representations of data. An autoencoder is an artificial neural network designed to learn compressed representations of input data [28] [29]. Synthetic data refers to artificially created data that mimics real data. According to Ewing [5], synthetic data generation serves as a solution to overcome limitations in existing mosquito data. Thus, synthetic data can be used to train mosquito population models, test control strategies, and predict the spread of transmitted diseases. The image augmentation process using VAE consists of two main stages. First, the VAE model is trained using classified mosquito larvae images. This training produces a model capable of learning the training data distribution and generating latent representations that capture essential features of the larvae images. After the VAE model is trained, the second stage involves using the model to generate new variations of larvae images. In this process, the VAE model takes existing mosquito larvae images as input and produces image variations by manipulating latent representations. Image variations generated by VAE are hypothesized—based on prior studies—to support more effective training of recognition and classification systems; however, our work evaluates generative quality only and does not assess downstream classifiers. Prior work on consistency training shows that

data diversity can improve recognition robustness [31]. In our study, we evaluate generative quality only and do not assess downstream recognition performance. This research produces diverse image data via VAE expressly for dataset enrichment under limited field data; broader impacts (e.g., public-health awareness or improved recognition performance) are outside the scope evaluated here.

1.1 Research questions & hypotheses

RQ1 (Generative capability). Can a convolutional VAE trained on 1,300 grayscale mosquito-larvae images learn the underlying data distribution and generate morphologically plausible and diverse samples?
H1. Under our training protocol, the model will converge stably (decreasing total loss) and achieve low Fréchet Inception Distance (FID), indicating statistical similarity between synthetic and real images.

RQ2 (Configuration sensitivity). Which training configuration (latent dimensionality and number of epochs) yields the best generative quality?
H2. Increasing model capacity and training time (e.g., latent = 100 vs. 50; 400 vs. 100 epochs) will reduce FID relative to a low-capacity baseline.

RQ3 (Variant comparison). How does the standard VAE compare with β-VAE, VQ-VAE, and VAE-S on FID under aligned settings?
H3. The standard VAE will achieve the best overall FID across the tested grid, with β-VAE competitive at lower-capacity settings; VQ-VAE and VAE-S will yield higher FID on this dataset.

RQ4 (Deferred / not evaluated here). Does VAE-based synthetic augmentation improve downstream classifier generalization compared with no augmentation or traditional augmentation?
H4 (Deferred). We expect VAE-generated augmentation to improve macro-F1 or accuracy on larvae classification; this is not evaluated in the present study and is deferred to future work (see Limitations & Future Work).

Scope note. We evaluate generative quality only (loss and FID) in this paper and do not evaluate downstream classification or detection models. The downstream hypothesis in RQ4/H4 is pre-specified for future experiments.

2 Related work

Synthetic Data Generation Using VAE has demonstrated significant potential across various fields, including art image generation and data augmentation for biological object classification [24]. Previous studies have validated VAE's capability to learn structured latent space representations and the benefits of data augmentation in enhancing deep learning model performance. Prior Informatica work reports VAE’s expansion to image processing within unsupervised learning, motivating its use here for synthetic mosquito larvae images [38].

Ihsan's (2023) study [25] proved that VAE can learn complex patterns in batik motifs and generate diverse new variations. This research demonstrated VAE's success in mapping batik motif features into latent space, enabling the generation of new motifs while preserving fundamental batik characteristics. These findings are relevant to the current study as they showcase VAE's ability to capture and produce visual variations from datasets, which can be adapted for synthetic mosquito larvae image generation [25].

Further research by Wang [21] in "Comparative Study of VAE and GAN for Mosquito Larvae Image Synthesis" specifically compared VAE and GAN performance in generating synthetic mosquito larvae images. The results showed that while GAN could produce images with sharper texture details, VAE excelled in:

1. Better training stability
2. Capability to generate more diverse yet valid morphological variations
3. More structured latent space interpretability

These findings support using VAE as the primary approach for synthetic mosquito larvae image generation.

Conversely, research by Akter [1] titled "Mosquitoes Classification Using Convolutional Neural Network with Data Augmentation" investigated the impact of data augmentation on mosquito classification accuracy. The results showed accuracy improvement from 70% to 93% after data augmentation implementation, indicating the importance of data variation in training more robust models. However, traditional augmentation techniques (such as rotation, flipping, or color changes) have limitations in creating complex morphological variations. This suggests that VAE-based synthetic data generation can provide better solutions by producing structurally diverse mosquito larvae samples while maintaining their biological characteristics.

Meanwhile, Kim & Myung [8], in their study "Autoencoder-combined Generative Adversarial Networks for Synthetic Image Data Generation and Detection of Jellyfish Swarm," proposed combining Generative Adversarial Networks (GANs) and Autoencoder to generate synthetic jellyfish images. Although model accuracy increased from 80.5% to 83.8%, the GAN-based approach required more complex training and tended to be less stable compared to VAE. These results align with Wang's [21] findings that demonstrate VAE's superiority in training stability for biological data. Based on previous studies, it can be concluded that:

1. VAE has proven effective in learning and generating synthetic image data, including biological images like mosquito larvae [21] and batik motifs [25].
2. VAE offers specific advantages for biological data, including training stability and the ability to generate valid morphological variations.
3. Synthetic data augmentation significantly improves classification accuracy [1], with VAE providing better solutions than traditional augmentation.
4. Although GAN can produce more detailed images, its training complexity makes VAE superior for

applications with limited data [8][21].

This study aims to: (i) leverage a VAE to generate morphologically plausible and diverse synthetic mosquito-larvae images; (ii) quantify generative quality using reconstruction loss and Fréchet Inception Distance (FID); and (iii) compare standard VAE against selected variants (β-VAE, VQ-VAE, VAE-S) under consistent settings. No experiments are conducted on classification or detection models; evaluating downstream performance is beyond this study's scope and is deferred to future work.

The study focuses on the challenge posed by limited mosquito larvae image datasets, which represents a primary obstacle in developing deep learning-based classification systems. These data limitations arise from two key factors: firstly, the difficulties associated with direct data collection, as acquiring mosquito larvae images in the field demands considerable time and resources; and secondly, the complexity of larval variations, as mosquito larvae from different species, such as Aedes aegypti, Anopheles, and Culex, display subtle visual differences, requiring large and diverse datasets for effective model training.

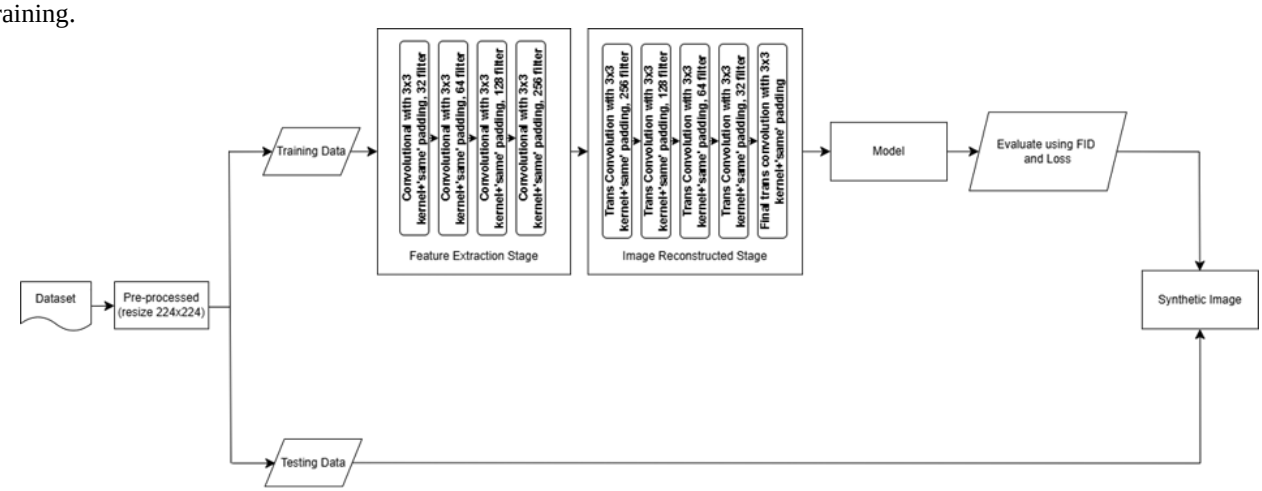

Figure 1: Proposed method

As outlined in Section 1, prior studies indicate that VAEs provide stable training and interpretable latent structures for biological imagery while enabling diversity beyond traditional augmentation [1, 8, 21, 25]. Building on these insights, we employ a VAE to synthesize grayscale larvae images and then detail our architecture and evaluation protocol in Section 3.10.

The system is designed with a VAE architecture consisting of two main components [27] [32]:

1. **Encoder**
   a. Input: Mosquito larvae image sized 224x224 pixels (grayscale).
   b. Process:
      i. Feature extraction using convolutional layers (Conv2D) with 32, 64, 128, and 256 filters.
      ii. Feature compression into a 100-dimensional latent space.
      iii. Generation of probabilistic distributions (z_mean, z_log_var) for latent vector sampling
2. **Decoder**
   a. Input: Latent vectors from the latent space.
   b. Process:
      i. Image reconstruction using Conv2DTranspose layers (256, 128, 64, 32 filters).
      ii. The output of synthetic images sized 224×224 pixels.
3. **Evaluation metrics**
   a. Fréchet Inception Distance (FID): Measures statistical similarity between synthetic and original images across different combinations of latent dimensions and epochs. The best FID score obtained was 0.4668.
   b. Loss Function: Combined reconstruction loss (MSE) and KL-divergence. The obtained loss values ranged approximately from 29206 to 33806.
4. **Comparison with VAE variants**
   Overall, the standard VAE achieved the best FID at latent 100 / 400, while β-VAE was competitive at 50 / 400; VQ-VAE and VAE-S yielded higher FID on this dataset.

3 Proposed method

The proposed method begins with data acquisition and preprocessing, followed by feature extraction using the customized VAE architecture. Subsequent stages include latent space optimization, synthetic data generation, and quantitative evaluation. Figure 1 presents the comprehensive workflow diagram detailing the sequential steps of the proposed research methodology.

Figure 2: Sample images of larvae

3.1 Dataset collection

The mosquito larvae image dataset was obtained from the Balai Teknik Kesehatan Lingkungan dan Pengendalian Penyakit (BTKLPP) Kelas I Medan. The data were collected using a digital microscope in the field. The required image formats were .JPG or .PNG. A total of 1300 images were collected with a size of 640×480 pixels, where 1200 images were used as training data and 100 images as testing data. Fig. 2 shows sample images of larvae.

3.2 Dataset preprocessing stage

The collected dataset with original dimensions of 640×480 pixels was resized to 224×224 pixels and converted to grayscale to accelerate computation. The dataset was then divided into 1200 images for training and 100 images for testing. The resulting images had dimensions of (224, 224, 1) for model training purposes. Fig 3 demonstrates the dataset resizing process.

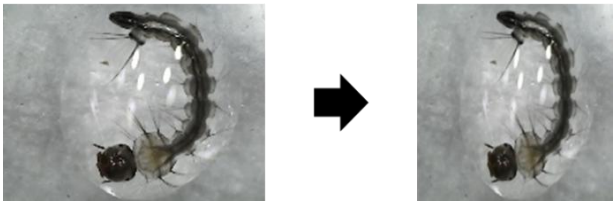

Figure 3: Dataset Pre-processed

3.3 Feature extraction stage

In the feature extraction stage, the encoder will compress high-dimensional input data into a low-dimensional latent space [20]. The VAE encoder is designed with four convolutional layers (Conv2D) that progressively extract visual features from 224×224 pixels mosquito larvae images. The first layer uses 32 filters with a 3×3 kernel and stride 2, followed by layers with 64, 128, and 256 filters using 3×3 kernel size, stride 2, and Leaky ReLU activation function to deepen feature extraction. The output from the final convolutional layer is then flattened and connected to a dense layer of 512 neurons to prepare for compression into the latent space. The extraction results will be mapped to a latent representation with 100-layer dimensionality.

3.4 Image reconstruction stage

The decoder in the VAE architecture is responsible for reconstructing images from latent representations back to their original form through a structured series of processes [20]. This process begins with transforming a 100-dimensional latent vector into a 3D tensor measuring 14×14×256 through a dense layer and reshaping operation. Subsequently, this tensor passes through four transposed convolutional layers (Conv2DTranspose) that perform gradual up sampling. Each transposed convolutional layer employs a 3×3 kernel with stride 2 and 'same' padding, incorporating Leaky ReLU activation functions for intermediate layers and sigmoid activation for the final layer. These layers serve to increase the spatial dimensions of the image while reducing feature depth, starting from 256 filters and ultimately producing a reconstructed image measuring 224×224 pixels with 1 channel (grayscale). The sigmoid activation in the final layer ensures pixel values remain within the (0,1) range, consistent with the input data. This reconstruction process not only generates images resembling the input but also enables the creation of novel variations through latent vector manipulation and forms the basis for calculating reconstruction loss during model training. The reconstruction output can subsequently be used for model quality evaluation or augmentation of mosquito larvae image datasets.

3.5 Variational autoencoder (VAE)

In this study, the VAE serves as the baseline generative model to synthesize additional mosquito-larvae images for data augmentation. The model adopts a convolutional encoder–decoder architecture and leverages the reparameterization trick to enable differentiable sampling from the latent distribution. Training optimizes a composite objective that combines a pixel-level reconstruction term with a Kullback–Leibler (KL) divergence regularizer to promote a well-structured latent space. After training, latent draws are decoded to produce morphology-consistent variants of the original images, which may be integrated into downstream recognition/classification pipelines in future studies; we do not evaluate such integration in this paper. In line with

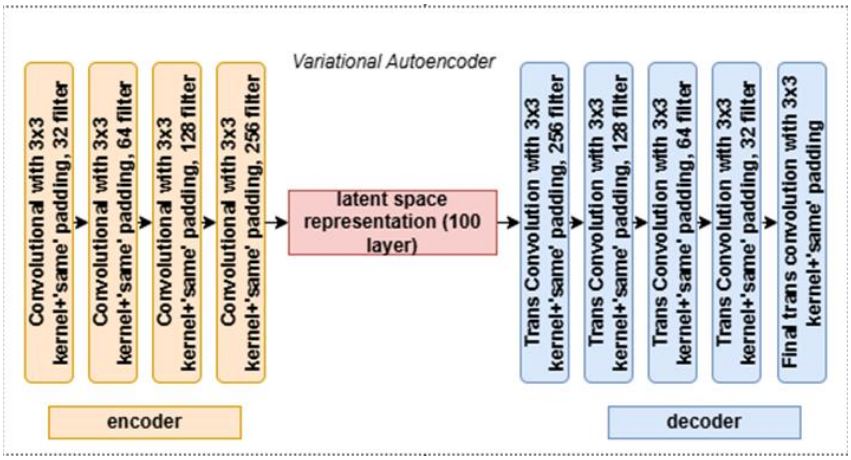

Figure 4: Architecture of VAE

prior uses in image processing, VAEs can generate realistic human faces [30] and remove noise from images [9]; however, a common limitation is that outputs may appear blurry relative to GAN-based models due to the KL-divergence regularization [7]. Unless otherwise specified, training and evaluation settings are aligned with the other model variants to enable fair comparison. A VAE is a generative variant of autoencoders that imposes a prior on the latent code and leverages the reparameterization trick for differentiable sampling, which suits image synthesis tasks [36].

We next detail the specific VAE architecture and training configuration employed in this study. The VAE model was selected for its ability to learn underlying data distributions and produce diverse samples while retaining essential characteristics of the original data. The implemented VAE architecture comprises two main components: an encoder that maps input images to a lower-dimensional latent space and a decoder that reconstructs images from latent representations. The encoder is designed with four convolutional layers (32, 64, 128, and 256 filters) that progressively extract hierarchical features from 224×224-pixel mosquito larvae images. The 100-dimensional latent space effectively captures significant variations across mosquito species. The decoder employs a symmetrical architecture with transposed convolutional layers to reconstruct images from latent vectors (256, 128, 64, and 32 filters). The model is optimized using a combined loss function consisting of reconstruction loss (measuring similarity to input images) and KL divergence loss (regulating latent space distribution). Implementation involved testing various combinations of latent dimensions and epochs to identify optimal parameters using the Adam optimizer. We tune model capacity and training length because VAE performance is sensitive to hyperparameters; Bayesian-optimization studies show systematic tuning can materially improve VAE performance [37]. Fig. 4 illustrates the VAE architecture.

3.6 Beta-variational autoencoder (β-VAE)

Beta-Variational Autoencoder (β-VAE) represents an important variant of the standard VAE that introduces the hyperparameter β as a balancing factor between reconstruction accuracy and latent feature disentanglement capability [6]. The selection of an optimal β value depends heavily on data characteristics and application objectives. β-VAE has proven particularly valuable across various domains requiring high interpretability of latent features, such as in medical image analysis, where the ability to separate independent anatomical factors (including organ shape, tissue texture, and pathological abnormalities) becomes crucial for diagnosis [11].

3.7 Adversarial variational autoencoder (AVAE)

The Adversarial Variational Autoencoder (AVAE) represents an advanced generative architecture that synergistically integrates the VAE framework with GANs [13]. This innovative hybrid configuration strategically employs the VAE decoder as the generator component within the GAN architecture while simultaneously introducing a dedicated discriminator network that learns to differentiate between authentic samples and reconstructed outputs [18]. Within the domain of synthetic data generation, AVAE exhibits significant advantages over conventional VAE implementations, particularly in its capacity to produce high-fidelity synthetic data with superior resolution. This enhanced performance addresses a critical limitation of traditional VAEs, which frequently generate visually degraded outputs characterized by excessive blurring - a direct consequence of over-compression in the latent space representation.

3.8 Vector quantized variational autoencoder (VQ-VAE)

The Vector Quantized Variational Autoencoder (VQ-VAE) represents a generative architecture variation of the VAE that introduces a discrete representation quantization mechanism in the latent space, contrasting with the continuous distribution approach of conventional VAEs. VQ-VAE is a model that demonstrates superior performance in complex data modeling [12]. The primary advantage of VQ-VAE lies in its ability to learn efficient hierarchical representations. This architecture has become a key component in generative models for text-based image generation [17]. VQ-VAE presents several implementation challenges, including:

a. Codebook Collapse, a condition where only 10% to 20% of codes remain active. This may reduce model performance efficiency.
b. Variance Allocation, a situation where the model struggles to distribute variance between the encoder and codebook.
c. Non-Differentiablity, a condition that can cause training instability and may be addressed using the straight-through estimator approach.

3.9 Sparse-variational autoencoder (VAE-S)

The Sparse Variational Autoencoder (VAE-S) represents a specialized variant of the VAE designed to generate sparse latent representations, where only a tiny fraction of units in the latent space remains active [4]. In contrast to conventional VAEs that employ continuous Gaussian distributions and produce dense latent representations, VAE-S enforces sparsity through three primary mechanisms: (1) utilization of non-Gaussian prior distributions [33] (e.g., Laplace or Spike-and-Slab) that naturally promote sparsity, (2) incorporation of L1 (Lasso) or L0 regularization on latent space activations, and (3) implementation of sparse activation functions in the encoder layers [16]. VAE-S offers significant advantages in representation efficiency and interpretability. The model typically activates only 5-20% of latent space dimensions for each input sample [10], making it particularly suitable for high-dimensional data applications such as medical imaging. However, VAE-S also presents several challenges, including complex training procedures due to sparse regularization and heightened sensitivity to λ value selection - where tremendous λ values may induce underfitting. In contrast, overly small values compromise the model's sparsity characteristics [22].

3.10 Model evaluation (protocol)

We evaluate the proposed VAE and its variants by tracking the total loss (reconstruction MSE + KL divergence) during training and by computing the Fréchet Inception Distance (FID) between real and generated images. The dataset contains 1,300 grayscale microscope images resized to 224 × 224 pixels and split 1,200/100 for train/test. Models are trained with Adam (batch size 64) across latent sizes {50, 100} and epochs {100, 400}. For FID, images are resized to 299 × 299 pixels to match the Inception-v3 backbone. Quantitative and qualitative results following this protocol are presented in Section 4. Scope note: We restrict evaluation to generative metrics (loss and FID) and do not evaluate downstream classification or detection results in this paper to make the study design explicit, we predefine success criteria tied to our metrics and baselines:

- S1 — Absolute FID target (primary). Achieve FID ≤ 0.60 on at least one configuration, indicating close alignment between generated and real feature distributions.
  Outcome: Best configuration reached FID = 0.4668 at latent 100 / 400 epochs.
- S2 — Relative improvement vs. baseline (primary). Achieve ≥ 50% relative reduction in FID compared with a low-capacity baseline (latent 50 / 100 epochs).
  Outcome: Reduction from 1.1588 to 0.4668 (≈59.7%).
- S3 — Training stability (primary). Observe monotonically decreasing total loss trends without divergence across epochs; improved configurations should also reduce FID relative to weaker ones.
  Outcome: Loss curves indicate stable convergence; FID improves with capacity/epochs.
- S4 — Variant comparison (secondary). The standard VAE achieves the best FID in ≥ 75% of tested configurations or the best overall FID, while β-VAE may be competitive at lower capacity.
  Outcome: VAE is best in 3/4 configs and achieves the lowest overall FID (0.4668 at 100/400); β-VAE is best at 50/400.
- S5 — Deferred downstream criterion (not evaluated here). In planned future work, VAE-augmented training should yield a statistically significant improvement (e.g., +≥3 pp macro-F1 under 5-fold CV) versus no-augmentation on a standard classifier (e.g., ResNet-18). (Not assessed in this paper.)

Note: These criteria operationalize RQ1–RQ3 within the scope of loss/FID only and explicitly mark RQ4 as deferred.

We next describe the two metrics that operationalize our criteria: (a) Training Loss and (b) Fréchet Inception Distance (FID).

Table 1: Loss value

Epoch	*Kl Loss*	*Reconstruction Loss*	*Loss*
1	47.4243	33758.8398	33806.2656
2	2.7390	32067.9824	32070.7207
3	5.0820	31858.7715	31863.8535
4	6.9111	21770.0117	31776.9238
5	10.1332	31380.1367	31390.2676
6	13.4444	31318.8809	31332.3262
…	…	…	…
395	25.4491	29453.2715	29478.7207
396	25.2640	29430.8984	29456.1641
397	25.5050	29202.8516	29228.3574
398	25.4908	29346.9512	29372.4395
399	25.4973	29324.8281	29350.3262
400	25.4764	29418.4766	29443.9512

a. Loss value

The loss function in VAE consists of two main components: reconstruction loss and KL-divergence loss. Reconstruction Loss is a function used to evaluate the quality of the model's output compared to its original input. It measures the discrepancy between the original input and the reconstructed image produced by the decoder. The reconstruction loss quantifies how closely the reconstruction approximates the original data, typically using metrics such as mean squared error (MSE) or binary cross-entropy (BCE). The following is the reconstruction loss formula.

$$\boldsymbol{L_{reconstruction}} = \|\boldsymbol{x} - \boldsymbol{x'}\|^2 = \sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{x'_i})^2 \quad (1)$$

Meanwhile, the KL-divergence loss functions as a regularization term that constrains the latent distribution to approximate a prior distribution, typically the standard normal distribution N(0, I) [6]. The Kullback Leibler (KL) Loss represents a fundamental component of the VAE framework, serving to quantify the divergence between the latent distribution predicted by the encoder and the target Gaussian distribution, thereby enabling the generation of realistic synthetic data. The following presents the KL Loss formula.

$$\boldsymbol{L_{KL}}(\mathrm{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\mathbf{1} + \log \boldsymbol{\sigma_j^2}(\mathrm{x}) - \boldsymbol{\mu_j^2}(\mathrm{x}) - \boldsymbol{\sigma_j^2}(\mathrm{x})\right) \quad (2)$$

The loss function represents the combined measure of Reconstruction Loss and KL Loss. These two components work synergistically to train the model, enabling it to simultaneously produce accurate data reconstructions while generating well-structured and generalizable latent representations.

The loss function evaluates how effectively the model predicts the desired output. The loss can be formulated as follows.

$$\boldsymbol{L_{total}} = \boldsymbol{L_{reconstruction}} + \boldsymbol{L_{KL}} \quad (3)$$

b. Fréchet Inception Distance (FID) score

The FID has emerged as a robust quantitative metric for assessing the performance of generative models, particularly GANs and Variational Autoencoders (VAEs), by providing a comprehensive measure of sample quality [2]. This metric compares the statistical properties of generated samples against real data distributions in the feature space of a pre-trained Inception network (Inception-v3) [34].

The fundamental superiority of FID over alternative evaluation metrics (including Inception Score and Precision-Recall measures) stems from its unique capacity to jointly evaluate two critical aspects of generative performance: (1) the perceptual fidelity of individual samples and (2) the diversity across the generated sample set [19]. The interpretation of FID scores follows an inverse relationship with generation quality, where decreasing FID values correspond to improving alignment distributions.

Table 2: Detail Result of FID score

Latent Dimension	*Epoch*	Batch Size	FID
50	100	64	1.1588
50	400	64	0.5360
100	100	64	0.6666
100	400	64	**0.4668**

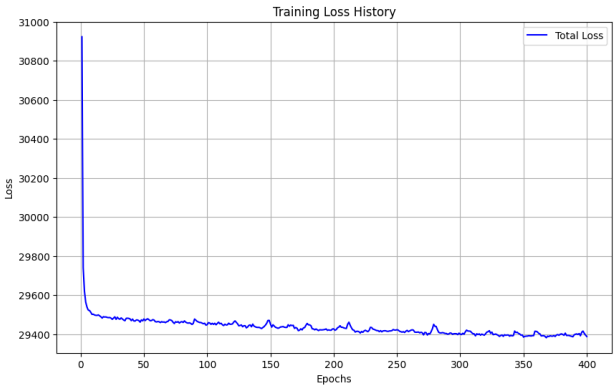

Figure 5(a): Loss chart latent dimension 50 epoch 100

Figure 5(b): Loss chart latent dimension 50 epoch 400

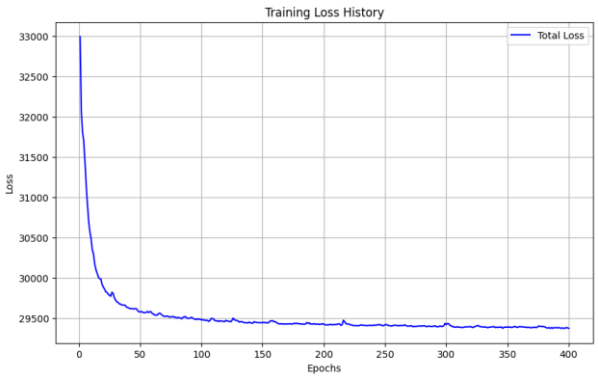

Figure 5(c): Loss chart latent dimension 100 epoch 100

Figure 5(d): Loss chart latent dimension 100 epoch 400

The metric ranges theoretically from 0 to ∞, with scores approaching 0 indicating nearly identical feature statistics between generated and authentic samples, while higher values reflect more significant divergence [3]. This characteristic makes FID particularly valuable for comparative analysis of model iterations or different architectures. The mathematical formulation of FID incorporates the Fréchet distance between multivariate Gaussian distributions fitted to the activation features of real and generated samples, as expressed by the following equation: between generated and real data

$$\boldsymbol{FID} = \|\boldsymbol{\mu_r} - \boldsymbol{\mu_g}\|^2 + \boldsymbol{Tr}\left(\sum_r + \sum_g - 2\left(\sum_r \sum_g\right)^{\frac{1}{2}}\right) \quad (4)$$

4 Performance analysis

4.1 Quantitative results (loss & FID)

Using the protocol in Section 3.10, we evaluate four configurations: (i) latent 50 / 100 epochs, (ii) 50 / 400, (iii) 100 / 100, and (iv) 100 / 400. Loss curves across epochs indicate stable convergence (Fig. 5a–d). The FID scores are 1.1588 (50/100), 0.5360 (50/400), 0.6666 (100/100), and 0.4668 (100/400), with the best performance achieved by latent 100 / 400 epochs (FID = 0.4668). Table 1 summarizes representative loss values, and Table 2 reports FID per configuration. The evaluation results are presented as follows.

A. Loss values

Testing revealed optimal performance with latent dimensions 100 and 400 epochs, demonstrating loss values ranging from 29,206 to 33,806. The loss charts for all dimension-epoch combinations are displayed in Fig. 5a, 5b, 5c, and 5d. Table 1 summarizes the loss values for the optimal configuration.

According to Fig. 5a, the loss values displayed in the loss chart for the combination of latent dimension 50 and 100 epochs range between 29508 to 33900. In Fig. 5b, the loss values shown in the loss chart for latent dimension 50 with 400 epochs range from 29380 to 30950.

According to Fig. 5c, the loss values displayed in the loss chart for the combination of latent dimension 100 and 100 epochs range between 29450 to 30800. In Fig. 5d, the loss values shown in the loss chart for latent dimension 100 with 400 epochs range from 29206 to 33806.

B. FID score

In the evaluation process using FID, the model resizes images to 299×299 pixels in accordance with the Inception-v3 library requirements. The FID results are presented in Table 2. Table 2 shows varied FID scores for each tested combination of latent dimension values and epochs.

Table 3: Comparison of FID Scores for VAE, β-VAE, VQ-VAE, and VAE-S.

Latent Dimension	Epoch	FID			
		VAE	β-VAE	VQ-VAE	VAE-S
50	100	1.1588	1.2956	3.5829	4.7286
50	400	0.5360	0.4826	1.9788	4.7518
100	100	0.6666	0.6679	2.1580	4.7452
100	400	**0.4668**	0.7364	2.1088	4.7683

4.2 Result and discussion

This study successfully implemented the VAE to generate synthetic image data of mosquito larvae with the aim of enhancing a limited dataset. Based on evaluations using loss metrics and the FID, the VAE model demonstrated satisfactory performance. The loss values ranged from 29206 to 33806, with the optimal combination observed at a latent dimension of 100 and 400 epochs. The losses show the model reconstructs images well, though improvement is still possible.

In addition, the evaluation results using FID revealed that the model successfully generated synthetic images that closely resemble the original data. The best FID score achieved was 0.4668, obtained with a latent dimension of 100 and 400 epochs. An FID value close to zero indicates high statistical similarity between the synthetic and original images, suggesting that the VAE effectively captured the visual characteristics of mosquito larvae.

These findings align with previous studies indicating that VAEs excel in training stability and the ability to generate valid larval variations, particularly for biological data such as mosquito larvae. The model's success is further evidenced by comparisons with other VAE variants, including β-VAE, VQ-VAE, and VAE-S, where the standard VAE produced better FID results. This strengthens the argument that VAE is well-suited for limited datasets, as demonstrated in the studies by Wang and Akter. However, an identified drawback is the tendency of the synthetic images generated by the VAE to appear blurry, which is attributed to the strong KL-divergence regularization.

Overall, this study demonstrates that synthetic data augmentation using VAE can serve as an effective solution to address the limitations of mosquito larvae datasets. The generated synthetic data enriches dataset diversity, and low FID values indicate statistical similarity to real images. Nevertheless, this study does not measure the effect of such data on the generalization of classification models. These findings provide a practical recipe for synthesizing larvae imagery that may support future development of larval identification systems; validating epidemiological or surveillance impacts requires downstream studies not performed here. For future research, further exploration of VAE architecture optimization or integration with other techniques, such as GANs, is necessary to reduce the blurriness effect in synthetic images. The researchers also conducted a comparison of FID values between the VAE and four VAE variant models, β-VAE, VQ-VAE, VAE-S, and VAE-S. The comparison results of FID values between the

Figure 6: Synthetic image from model

VAE and the four VAE variant models will be presented in Table 3. The optimal FID score was achieved using the VAE configuration with latent dimension 100 and 400 epochs. Figure 6 presents the synthetic mosquito larvae images generated by the model.

4.3 Findings vs. research questions

- RQ1/H1 (Generative capability): Achieved FID 0.4668 (latent 100 / 400) with stable convergence, supporting the hypothesis that the VAE learns the larvae image distribution and yields diverse samples. S1 and S3 satisfied.
- RQ2/H2 (Configuration sensitivity): Increasing capacity/epochs lowered FID relative to the 50/100 baseline (1.1588 → 0.4668, ≈59.7% reduction). S2 satisfied.
- RQ3/H3 (Variant comparison): The standard VAE attained the best overall FID and was top in 3/4 configs; β-VAE was competitive at 50/400. S4 satisfied with nuance.
- RQ4/H4 (Deferred): Not evaluated in this study; downstream augmentation benefits to be tested in future work. (See Limitations & Future Work.)

5 Limitations & future work

First, our evaluation focuses exclusively on generative quality (loss, FID) and does not include downstream classification/detection experiments. Second, we use a single grayscale dataset with a fixed train/test split, which may limit generalizability. Third, image sharpness—often a limitation of VAEs—was not addressed with perceptual or adversarial refinements.

Future work will (i) measure augmentation gains using standard classifiers (e.g., ResNet-18) under k-fold cross-validation and external held-out sets, (ii) test domain generalization to different acquisition devices/labs, and (iii) explore hybrid VAE–GAN to reduce blur while maintaining stability.

6 Conclusion

This study provides theoretical contributions by demonstrating the effectiveness of VAE in learning the data distribution of mosquito larvae images and generating structured latent representations [9], while reinforcing the theory that VAE can overcome data limitations through synthetic data that preserves essential characteristics of the original data and confirming the importance of KL-divergence regularization for model stability despite causing blurring effects in the output [7][21]. Practically, this research offers an efficient dataset augmentation solution by producing high-quality synthetic images (FID 0.4668) through VAE implementation (latent dimension 100, epoch 400 , offering a reproducible recipe for dataset enrichment under limited field data. Our evaluation is limited to generative metrics (loss and FID); we do not test classification or detection models. Accordingly, any improvement in downstream accuracy or generalization should be regarded as a hypothesis informed by prior literature rather than a finding of this study. Future work will explicitly assess augmentation benefits by training and testing classifiers with and without VAE-generated images.

Comparative analysis of various VAE architecture variants reveals that the standard VAE outperforms β-VAE, VQ-VAE, and VAE-S for this specific biological data context. Furthermore, the stringent KL-divergence regularization creates a trade-off between training stability and visual output sharpness, while the 100-dimensional latent space proves optimal for capturing mosquito larvae morphological variations. Potential implications include supporting the future development of larvae-classification systems via synthetic augmentation and offering a framework that may extend to other vector species in tropical regions; however, these application-level gains require dedicated downstream (and where relevant, clinical/operational) validation. For future research, the exploration of hybrid VAE-GAN architectures is recommended to enhance visual output sharpness, along

with clinical validation of diagnostic system accuracy using synthetic data and the development of integrative pipelines with vector-borne disease surveillance systems. These findings make significant contributions to advancing computer vision methods in tropical public health while opening new research opportunities for applying generative models to address global health challenges.

References

[1] Akter, M., Hossain, M.S., Ahmed, T.U., Andersson, K. (2021). Mosquito Classification Using Convolutional Neural Network with Data Augmentation. In: Vasant, P., Zelinka, I., Weber, GW. (eds) Intelligent Computing and Optimization. ICO 2020. Advances in Intelligent Systems and Computing, vol 1324. Springer, Cham. https://doi.org/10.1007/978-3-030-68154-8_74 15

[2] Asperti et al. 2022. Enhancing Variational Generation Through Self-Decomposition. Digital Object Identifier, 10.1109/ACCESS.2022.3185654. IEEE. https://doi.org/10.1109/ACCESS.2022.3185654

[3] Borji, A. (2022). "Pros and Cons of GAN Evaluation Measures: New Developments". Computer Vision and Image Understanding, 215, 103329. https://doi.org/10.1016/j.cviu.2022.103329

[4] Dai, B., & Wipf, D. (2019). Diagnosing and Enhancing VAE Models: A Sparse Coding Perspective. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1906.02691

[5] Ewing, D. A., Cobbold, C. A., Purse, B. V., Nunn, M. A., & White, S. M. (2016). Modelling the effect of temperature on the seasonal population dynamics of temperate mosquitoes. Journal of theoretical biology, 400, 65-79. https://doi.org/10.1016/j.jtbi.2016.04.008

[6] Hinggis et al. 2017. β -VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK. Under review as a Conference Paper at ICLR. https://doi.org/10.48550/arXiv.1804.03599

[7] Huang, T., Ding, Z., Zhang, J., Tai, Y., Zhang, Z., Chen, M., Wang, C., & Liu, Y. (2023). Learning to Measure the Point Cloud Reconstruction Loss in a Representation Space. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12208–12217. https://doi.org/10.1109/CVPR52729.2023.01175

[8] Kim, Kyukwang, Myung, Hyun. (2018). Autoencoder-Combined Generative Adversarial Networks for Synthetic Image Data Generation and Detection of Jellyfish Swarm. Journal Article, 6, 54207-54214. IEEE. https://doi.org/10.1109/ACCESS.2018.2869250

[9] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). arXiv:1312.6114. https://doi.org/10.1609/aaai.v33i01.3301492

[10] Li, X., et al. (2019). Sparse Variational Autoencoder for Unsupervised Feature Learning. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1109/CVPR52729.2023.01175

[11] Locatello, F., et al. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. ICML. https://doi.org/10.48550/arXiv.1811.12359

[12] Luo et al. 2024. Quaternion Vector Quantized Variational Autoencoder. IEEE SIGNAL PROCESSING LETTERS, Vol. 32, 1070-9908. IEEE. https://doi.org/0.1109/LSP.2024.3385109

[13] Mescheder, L., Nowozin, S., & Geiger, A. (2017). "Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks". Proceedings of the 34th International Conference on Machine Learning (ICML), 70, 2391-2400. https://doi.org/10.5555/3305890.3306030

[14] Naderi, H., B. H. Soleimani, dan S. Matwin. 2020. Generating High-Fidelity Images with Disentangled Adversarial VAEs and Structure-Aware Loss. 2020 International Joint Conference on Neural Networks (IJCNN), 978-1-7281-6926-2, 20. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206846

[15] Ndione, R. D., O. Faye, M. Ndiaye, A. Dieye, dan J. M. Afoutou. 2007. Toxic effects of neem products (Azadirachta indica A. Juss) on Aedes aegypti Linnaeus 1762 larvae. In African Journal of Biotechnology. 6(24): 2846- 2854. https://doi.org/10.5897/AJB2007.000-2348

[16] Nguyen, T., et al. (2023). Adaptive Sparse Variational Autoencoder with L0 Regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2022.3220765

[17] Ramesh et al. 2021. Zero-Shot Text-to-Image Generation. International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2102.12092

[18] Rosca, M., Lakshminarayanan, B., & Mohamed, S. (2019). "Improving Generalization in Generative Adversarial Networks via Hierarchical Variational Inference". International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1807.03653

[19] Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. Advances in Neural Information Processing Systems, 31. https://doi.org/10.48550/arXiv.1806.00035

[20] Tai, X.-C., Liu, H., & Chan, R. (2023). PottsMGNet: A Mathematical Explanation of Encoder-Decoder Based Neural Networks (arXiv:2307.09039). https://doi.org/10.48550/arXiv.2307.09039

[21] Wang, R., et al. (2022). Comparative Study of VAE and GAN for Mosquito Larvae Image Synthesis. Computers in Biology and Medicine. https://doi.org/10.1016/j.compbiomed.2022.105440

[22] Wang, Y., Zhang, L., Chen, X., & Liu, Q. (2023). Adaptive sparse variational autoencoders: Balancing sparsity and reconstruction accuracy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 1895-1909. https://doi.org/10.1109/TPAMI.2022.3189841

[23] P. I. Nainggolan et al., (2023). Classification Of Aedes Mosquito Larva Using Convolutional Neural Networks and Extreme Learning Mechine. 2023 7th International Conference on Electrical, Telecommunication and Computer Engineer (ELTICOM). doi: 10.1109/ELTICOM61905.2023.10443125.

[24] P. I. Nainggolan et al., (2025). Detection and Classification of Mosquito Larvae Based on Deep Learning Approach. Eng. Lett. 2025, 33, 198–206.

[25] Ihsan. (2023). Initial Study of Batik Generation using Variational Autoencoder. 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).

[26] Azis. (2023). Comparative Analysis of Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) Algorithms for image classification. Journal Eletronik Sistem InformasI (JESII).

[27] M. Sami and I. Mobin, "A Comparative Study on Variational Autoencoders and Generative Adversarial Networks," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), IEEE, Mar. 2019, pp. 1–5. https://doi.org/10.1109/ICAIIT.2019.8834544

[28] L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto, "Variational Autoencoder," in Variational Methods for Machine Learning with Applications to Deep Networks, Cham: Springer International Publishing, 2021, pp. 111–149. https://doi.org/10.1007/978-3-030-70679-1_5

[29] Q. Xu, Z. Wu, Y. Yang, and L. Zhang, "The difference learning of hidden layer between autoencoder and variational autoencoder," in 2017 29th Chinese Control and Decision Conference (CCDC), IEEE, May 2017, pp. 4801–4804. https://doi.org/10.1109/CCDC.2017.7979344

[30] Mansour, R. F., Escorcia-Gutierrez, J., Gamarra, M., Gupta, D., Castillo, O., & Kumar, S. (2021). Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. Pattern Recognition Letters, 151, 267–274. https://doi.org/10.1016/j.patrec.2021.08.020

[31] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33, 6256–6268.

[32] M. Sami and I. Mobin, "A Comparative Study on Variational Autoencoders and Generative Adversarial Networks," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), IEEE, Mar. 2019, pp. 1–5. doi: 10.1109/ICAIIT.2019.8834544.

[33] Yu. (2024). CS-Intro VAE: Cauchy-Schwarz Divergence-Based Introspective Variational Autoencoder. IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 26. https://doi.org/10.1109/TMM.2024.3365439

[34] Nash, C., Menick, J., Dieleman, S., Battaglia, P.W., 2021. Generating images with sparse representations. arXiv preprint arXiv:2103.03841.

[35] Razavi, A., Oord, A.v.d., Vinyals, O., 2019. Generating diverse high-fidelity images with vq-vae-2. arXiv preprint arXiv:1906.0044.

[36] A. Chefrour and L. Souici-Meslati, "Unsupervised Deep Learning: Taxonomy and Algorithms," Informatica (Slovenia), vol. 46, no. 2, pp. 151–168, 2022, doi: 10.31449/inf.v46i2.3820.

[37] B. Fu, "Variational Autoencoder-based High-dimensional Feature Extraction for Economic Analysis of Power Cost Data," Informatica (Slovenia), vol. 49, no. 25, pp. 75–91, 2025, doi: 10.31449/inf.v49i25.8012.

[38] L. Zou and M. Zhang, "Variational Autoencoder Model Combining Deep Learning and Probability Statistics and Its Application in Large-scale Data Analysis," Informatica (Slovenia), vol. 48, no. 22, pp. 31–46, 2024, doi: 10.31449/inf.v48i22.6921.