Efficient Image Encoding for DNA Data Storage Using Latent Space Representation of Variational Autoencoder

Muhammad Rafi Muttaqin[1], Yeni Herdiyeni[2*], Agus Buono[3],
Karlisa Priandana[4], and Iskandar Zulkarnaen Siregar[5]
[1,3,4]Doctoral Program in Computer Science, School of Data Science, Mathematics, and Informatics, IPB University
Jawa Barat, Indonesia 16680
[1]Informatic Engineering, Sekolah Tinggi Teknologi Wastukancana
Jawa Barat, Indonesia 41151
[2]Artificial Intelligence Study Program, School of Data Science, Mathematics, and Informatics, IPB University
Jawa Barat, Indonesia 16680
[5]Department of Silviculture, Faculty of Forestry, IPB University
Jawa Barat, Indonesia 16680
Email: [1]rafiaqinmuttaqin@apps.ipb.ac.id, [2]yeni.herdiyeni@apps.ipb.ac.id,
[3]agusbuono@apps.ipb.ac.id, [4]karlisa@apps.ipb.ac.id, [5]siregar@apps.ipb.ac.id

***Abstract*—Deoxyribonucleic Acid (DNA)-based data storage as a revolutionary approach to long-term, high-density information storage faces the challenge of high costs and biological constraints in DNA synthesis. Unlike conventional binary encoding, DNA storage requires efficient data compression techniques to minimize the length of the resulting DNA sequence. The research explores the use of a Variational Autoencoder (VAE) as a deep learning-based compression method to reduce image complexity. VAE enables significant data reduction while retaining essential structural features. The dataset used is from the Modified National Institute of Standards and Technology (MNIST) database. The proposed method involves encoding images into latent variables, followed by binarization and translation into DNA sequences while applying biological constraints such as maintaining GC balance and avoiding homopolymers, to ensure stability and sequencing accuracy. Reconstruction quality is assessed using the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), yielding an average SSIM of 0.8017 and an average PSNR of 18.91 dB. Analysis of variance shows that digits with greater variability suffer more noticeable degradation, whereas digits with lower variability preserve their structural integrity. These results suggest that integrating VAE into DNA data storage can reduce sequence complexity while still preserving recognizable image structures. The research contributes to the advancement of DNA-based data storage by presenting optimized coding techniques.**

Received: June 10, 2025; received in revised form: Nov. 04, 2025; accepted: Nov. 10, 2025; available online: July 27, 2026.
*Corresponding Author

***Index Terms*—Biological Constraint, Deoxyribonucleic Acid (DNA) Data Storage, Image Compression, Latent Space Representation, Variational Autoencoder (VAE)**

I. Introduction

DEOXYRIBONUCLEIC Acid (DNA)-based digital storage has seen rapid progress in recent years [1–4]. This technology combines massive storage density with exceptional longevity and a tiny physical size relative to traditional storage media. The main barrier to real-world adoption remains economic. Synthesizing the long nucleotide sequences required for data representation is expensive [5]. In a direct comparison, DNA excels over conventional binary media thanks to its extreme density, resistance to chemical degradation, and minimal footprint, while conventional media face challenges in terms of long-term durability and scalability. These realities motivate continued research into DNA-centric encoding methods.

Since image files require large amounts of storage space, minimizing the length of DNA sequences generated from such files requires an efficient compression approach. Variational Autoencoder (VAE), a deep learning model class, generates low-dimensional latent variables that capture the core features of the data [6]. By retaining important information despite the reduced

size, this representation is a promising alternative for usage in DNA data storage systems [5].

Recent work on DNA-based data storage spans a range of strategies crafted to address both efficiency issues and biological constraints. Although schemes that directly translate binary code into DNA have been proposed, they continue to suffer from elevated error rates and sequence instability. This occurs because GC balance and homopolymer limitations are not taken into account [7]. Therefore, to overcome this challenge, fountain codes have been developed that allow for higher storage density and better compliance with biochemical limitations. However, this approach requires complex decoding and repeated reading to ensure data integrity [7, 8]. On the other hand, compression techniques that have been proven to retain critical information while reducing storage size have been discovered, namely vector quantization and standard formats (Joint Photographic Expert Group (JPEG) and DNA) [8, 9]. Deep learning approaches, such as DNA encoding scheme based on Quantized ResNet Variational Autoencoder (QRes-VAE) and Levenshtein Code (LC) (DNA-QLC), have recently been used in research because of their advantages in terms of stability and information density [10–12]. However, approaches using autoencoders and VAE show more potential than all other strategies in producing compact latent representations that can be adjusted to biochemical constraints [7, 12].

Therefore, this research integrates VAE into the digital compression process before the data are encoded into DNA sequences. Then, the image representation is converted into binary and translated into DNA sequences in latent space, taking into account biological constraints (homopolymers and Guanine-Cytosine (GC) content). The output of DNA synthesis is read and decoded to reconstitute the images. In summary, this step improves storage efficiency, preserves reconstruction accuracy, and cuts the cost of synthesis [13].

Beyond reducing sequence length, the research optimizes the full storage retrieval workflow by integrating VAE-based compression step. It also assesses reconstruction fidelity, compliance with biological constraints, and overall compression efficiency to secure savings in both cost and storage. Trade-offs between nucleotide efficiency and image fidelity are measured with Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Overall, the researchers employ VAE to reduce image complexity within a deep-learning compression framework.

II. Research Method

The research uses the VAE model approach to compress digital image data. The compressed data are then converted into DNA sequences for DNA-based data storage. The steps used in the research are shown in Fig. 1.

A. Data Collection

At this stage, the image used is from the Modified National Institute of Standards and Technology (MNIST) database of handwritten digits dataset from Yann Lecun's official website at http://yann.lecun.com/ or the Keras library (https://keras.io/api/datasets/mnist/). The collected data are prepared for the compression process. MNIST is a standard benchmark in pattern recognition and widely used to evaluate image-classification methods in machine learning. It is known for its practical applications, especially in handwriting classification [14]. MNIST comprises 70,000 grayscale images of handwritten digits (0–9), each at 28×28 resolution, split into 60,000 for training and 10,000 for testing samples. Characterized by its small and uniform image dimensions, the MNIST dataset offers low computational overhead, allowing a focus on algorithmic design. Its suitability stems from a normalized and centered structure, which mitigates handwriting irregularities. The resulting consistency among the numerical images creates an ideal and standardized environment for classification. It is effective for testing diverse algorithmic approaches, including Convolutional Neural Networks (CNN) [15–21].

B. Image Compression

The prepared image data are compressed using the VAE model. The VAE is chosen over the autoencoder for several reasons. First, it has probabilistic representation. A VAE produces a representation in the form of a probabilistic distribution (mean and variance) instead of a fixed point, as in a regular autoencoder [22]. It provides more flexibility in the data reconstruction process because the VAE can sample from the resulting distribution. Second, it has control on latent space. With Kullback-Leibler Divergence (KL Divergence)-based regulation, VAE produces a better-organized latent space [23]. It ensures that the latent representation is more meaningful and does not overfit the training dataset. Third, it has a better generalization. The VAE can produce a more robust representation, thus working better for data with high variation [24]. It is suitable for datasets with MNIST handwriting images of various handwriting styles. Fourth, it has optimization for compression. By utilizing the probabilistic distribution property, the VAE can significantly reduce the data dimensions without losing important information [25]. It helps to minimize the length of the generated DNA sequence. Last, it has sampling capability. The VAE

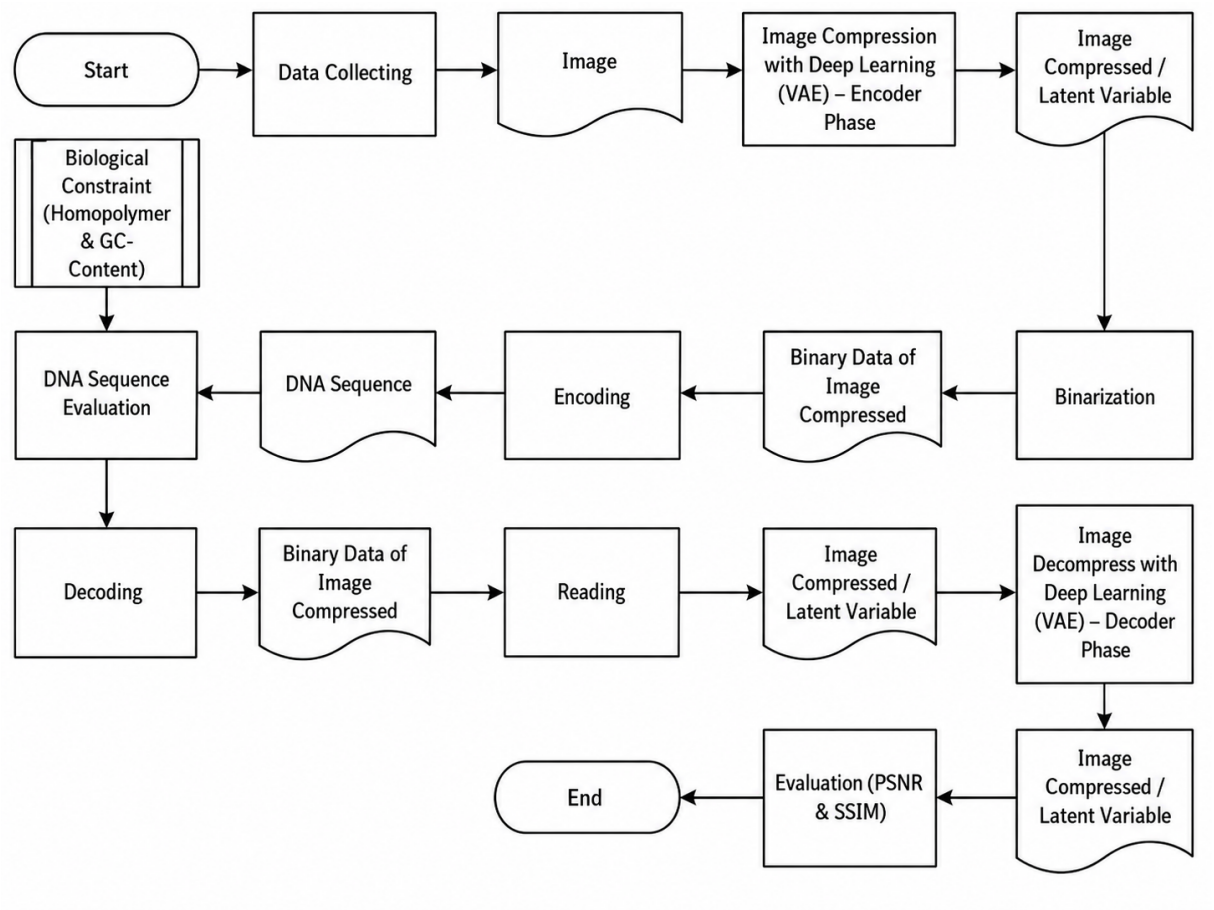


Fig. 1. The research stages. Note: Deoxyribonucleic Acid (DNA), Variational Autoencoder (VAE), Guanine-Cytosine (GC) content, Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR).

TABLE I
DETAILED ARCHITECTURE OF THE FULLY CONNECTED VARIATIONAL AUTOENCODER (VAE)

Encoder				
Layer	Type	Input Dim	Output Dim	Activation
E1	Flatten	1×28, ×28	784	–
E2	Dense	784	400	ReLU
E3	Dense	400	200	ReLU
z_mean	Dense	200	10	Linear
z_logvar	Dense	200	10	Linear
Reparam	-	μ, $\log \sigma^2$	z(10)	-
Decoder				
D1	Dense	10	200	ReLU
D2	Dense	200	400	ReLU
D3	Dense	400	784	Sigmoid
Reshape	-	784	1×28, ×28	-

allows for new sampling of the latent space, which can be used for new data generation or simulation, something that ordinary autoencoders cannot do [26].

The model used is a fully connected VAE for 28×28 single-channel images. Details regarding the encoder and decoder layers, including layer type, dimension size, activation function, z_{mean}, z_{logvar}, and the reparameterization mechanism, are shown in Table I. In the encoder, the input is flattened into a 784-dimensional vector and projected through Dense ($784 \rightarrow 400$, ReLU) followed by Dense ($400 \rightarrow 200$, ReLU). Two linear branches form the parameters $z_{\text{mean}} \in \mathbb{R}^{10}$ and $z_{\text{logvar}} \in \mathbb{R}^{10}$, and the latent sample z is generated using the reparameterization trick. The decoder maps z through Dense ($10 \rightarrow 200$, ReLU) $\rightarrow$ Dense ($200 \rightarrow 400$, ReLU) $\rightarrow$ Dense ($400 \rightarrow 784$, $\mathrm{Sigmoid}$), then reshapes it into 1×28×28. Training is carried out with a reconstruction loss of Binary Cross-Entropy and KL divergence (β = 1.0), using the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 128, and 10 epochs. Then, data are taken from a directory in `ImageFolder` format, processed with `Grayscale`, resized to 28 × 28, and converted to tensors so that pixel values fall within $[0, 1]$. The dataset is split into 80% for training and 20% for testing. The random seed is set to 42 for `torch`, `numpy`, and `random`.

Meanwhile, training is executed on Central Processing Unit (CPU) or Graphic Processing Unit (GPU), depending on availability. The resulting latent variable z represents the data in a lower-dimensional space and is subsequently used for the next stage, namely binarization.

C. Binarization

At this stage, the data represented in latent space using the VAE is converted into a binary format. A binarization process is performed to convert the numerical representation into a series of bits (0 and 1) [27]. Before encoding, this research applies binarization to condense the raw data into binary form. The goal is to retain essential characteristics while streamlining the encoding process, thereby preparing clean, efficient input for the later DNA conversion step.

D. Encoding

At this stage, the binary data are encoded into a DNA sequence using an encoding algorithm [28] that considers biological constraints. First, homopolymers are curtailed by capping runs of identical nucleotides at three bases to mitigate synthesis and sequencing issues [29]. Second, GC content is regulated to remain within a balanced range, avoiding extremes that compromise thermal stability [30]. This procedure yields a stable, efficient DNA sequence suitable for synthesis. This stage results in a DNA sequence that is stable, efficient, and ready for synthesis.

E. DNA Sequence Evaluation

The validation of encoded sequences involves an assessment of their adherence to biological constraints and encoding efficiency. This process includes two key checks. First, a homopolymer analysis ensures that no nucleotide is repeated more than three times consecutively. Second, the GC-content is evaluated against a 40–60% threshold, a critical range due to its direct impact on DNA stability, efficiency of synthesis, and sequencing fidelity. After verification, the researchers perform the reverse mapping, converting the DNA sequence back into its binary form to reconstruct the original representation as a series of bits (0 and 1).

F. Reading

The decoded binary data are transformed into numeric latent codes using an inverse map aligned with the prior binarization procedure. Aligning these two steps preserves consistency with the initial latent representation. The recovered codes are subsequently used by the VAE to decompress the image.

G. Image Decompression with Deep Learning

The decoded binary data are first converted back into numerical values representing the latent variables. These recovered values form a 10-dimensional latent vector that is used as the input to the decoder component of the trained VAE. During decompression, the decoder transforms the compact latent representation using a sequence of fully connected layers. The latent vector is projected from 10 to 200 neurons and subsequently to 400 neurons using the Rectified Linear Unit (ReLU) activation function. It is then mapped to 784 output values via a sigmoid-activated layer, ensuring the reconstructed pixel intensities remain between 0 and 1.

The resulting 784-dimensional output vector is reshaped into a grayscale image with a resolution of 28×28 pixels. This process reconstructs the main structural characteristics of the original image from the compressed latent representation. However, because the original 784 pixel values are represented by only 10 latent variables, some fine details, sharp edges, and local intensity variations may not be completely preserved. Consequently, the reconstructed image may exhibit blurring or smoothing while remaining visually recognizable. The similarity between the original and reconstructed images is subsequently evaluated using the SSIM and PSNR.

H. Evaluation

The evaluation stage is conducted to assess the quality of the reconstructed data. The reconstructed images are evaluated with several metrics, such as SSIM and PSNR. They compare the image quality with the original data [31, 32].

III. Results and Discussion

At this stage, compression is carried out in the form of an image size reduction process from 28×28 or has 784 image intensity values into latent variables of 10 values only. The process carried out is to first train the VAE model with the entire MNIST dataset with a predetermined number of latent variables, namely 10 variables. The trained model is then saved in PyTorch format with an extension `.pth`. Files with `.pth` extension can be used to store the trained model parameters.

After the model is stored, it is called to perform the encoder or compression process by reducing the number of input image variables to latent variables. This result is then subjected to the binarization stage. Tests are conducted for each class, each having 10 MNIST images with a size of 28×28 pixels or 784



Fig. 2. Representative Modified National Institute of Standards and Technology (MNIST) samples (digits 0–9) employed for Variational Autoencoder (VAE)-based encoding and evaluation.

variables. An example image from each class is shown in Fig. 2.

The VAE model used compresses the original MNIST images. The images have dimensions of 28×28 pixels, totaling 784 intensity values, into only 10 latent variables. The latent variables serve as compact representations of the original images, with reduced dimensions while still preserving their key structural features.

Encoding begins by inputting MNIST digits into the VAE encoder, which generates latent codes. This operation compresses the high-dimensional input into 10 variables. These variables are sampled from a probability distribution characterized by its statistical parameters, namely the mean (μ) and standard deviation (σ). The objective of this dimensional reduction is to obtain compact storage and facilitate subsequent DNA conversion, while retaining the essential information required for accurate reconstruction of the original image.

The compression efficiency is assessed through Principal Component Analysis (PCA). The analysis demonstrates that the latent variables form distinct clusters, which align with specific digit classes. Consequently, the VAE framework proves capable of discerning the unique characteristics of handwritten digits while enabling a substantial reduction in the storage space required for DNA synthesis.

VAE offers data compression capabilities, which can reduce storage requirements in DNA synthesis. The ability to identify essential features has also been confirmed through PCA. The model's latent variables form separate clusters according to digit classes.

The PCA result confirms that 10 latent variables successfully capture the essential information present in the original imagery. This argument is based on a dataset consisting of 100 images in total, with 10 samples selected from each of the 10 classes. The PCA plot of the latent variables is shown in Fig. 3.

Figure 3 presents the patterns produced by the latent variables derived from the 10 MNIST classes. From the visualization pattern, the images in one class are mostly close to each other owing to the similarity of their features. If those are located far away, it is possible that there are slightly different writing patterns, and there may be patterns that resemble other classes, such that their position is closer to images from other classes.

The MNIST dataset contains image data of handwritten numbers from 0 to 9. There are many handwriting models for just one number. Therefore, the features contained in the image, such as concave, straight, thick, and thin patterns, in each image significantly affect the latent variable values that have been generated and visualized with PCA. For example, images with ID 50.jpg in digit 6 and 84.jpg in digit 8 are located close to each other in the PCA projection, indicating that they share structural similarities. Both digits exhibit closed circular loops and curved strokes, which make their latent representations overlap despite belonging to different classes. It suggests that the VAE captures dominant geometric patterns, such as rounded contours, rather than class-specific details, explaining why digits with similar structural motifs tend to cluster together in the latent space, as shown in Fig. 4.

Based on the analysis results shown in Figs. 3 and 4, it can be concluded that the majority of the latent variable values tend to cluster based on the same class. Therefore, the latent variables successfully represent the original data. In the case of size compression for the development of DNA data storage, this latent space can be used to reduce the number of variable values that will be stored in the DNA medium through synthesis.

The generated latent variables are of the floating type. Therefore, the process of converting latent variables into binary data uses the concept of floating points with the IEEE 754 32-bit single-precision standard [33, 34]. Detailed results of the binarization process using the IEEE 754 32-bit single-precision standard are provided in Table A1 in Appendix.

After the latent variable data are converted into binary data, the encoding process is carried out on DNA Sequence-based data consisting of four bases, namely adenine, thymine, guanine, and cytosine, which are represented by the letters A, T, G, and C, respectively. In making DNA sequences, it is also necessary to pay attention to the biological limitations of DNA sequences, including homopolymers [35]. A homopolymer refers to a sequence of DNA bases consisting of the same nucleotide repeated in a sequence. In this case, an excessively long homopolymer can increase errors during DNA synthesis and sequencing. Therefore, controlling their length is essential for maintaining the accuracy of the encoded data [36]. Based on several studies, the recommended maximum length of repetition is three repetitions [35, 36]. The results of the encoding from the binarization process are presented in Table A2 in Appendix. The resulting DNA sequence is stored in the DNA medium through the synthesis process, both in Vivo and in Vitro. However, this synthesis process is not performed in the research.

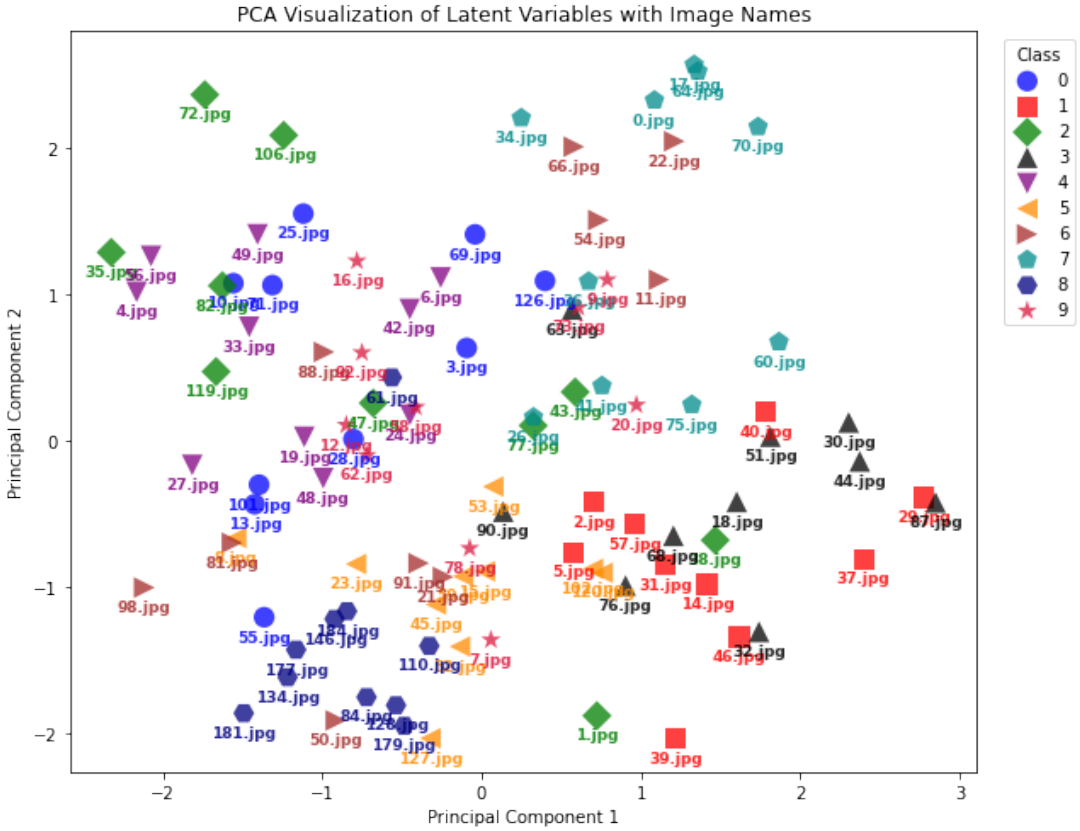


Fig. 3. 2-D Principal Component Analysis (PCA) projection of Variational Autoencoder (VAE) latent representations for Modified National Institute of Standards and Technology (MNIST) digits (0–9), showing class-wise clustering.

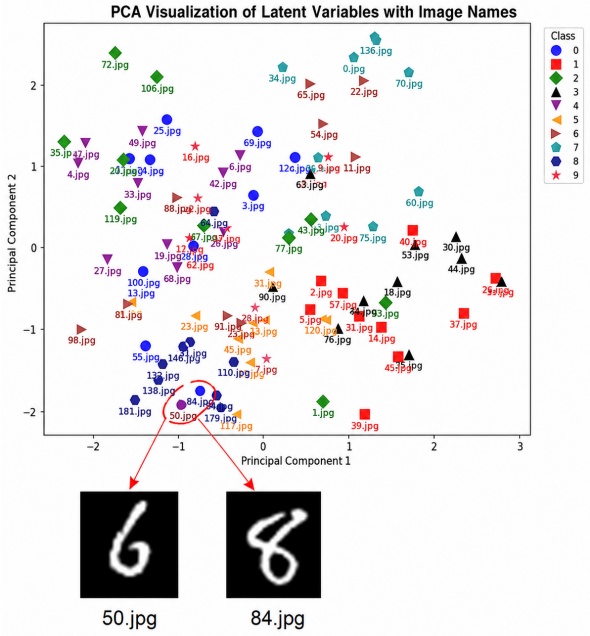


Fig. 4. Two-dimensional Principal Component Analysis (PCA) visualization of the Variational Autoencoder (VAE) latent space showing the neighboring positions of digit ‘6’ (50.jpg) and digit ‘8’ (84.jpg).

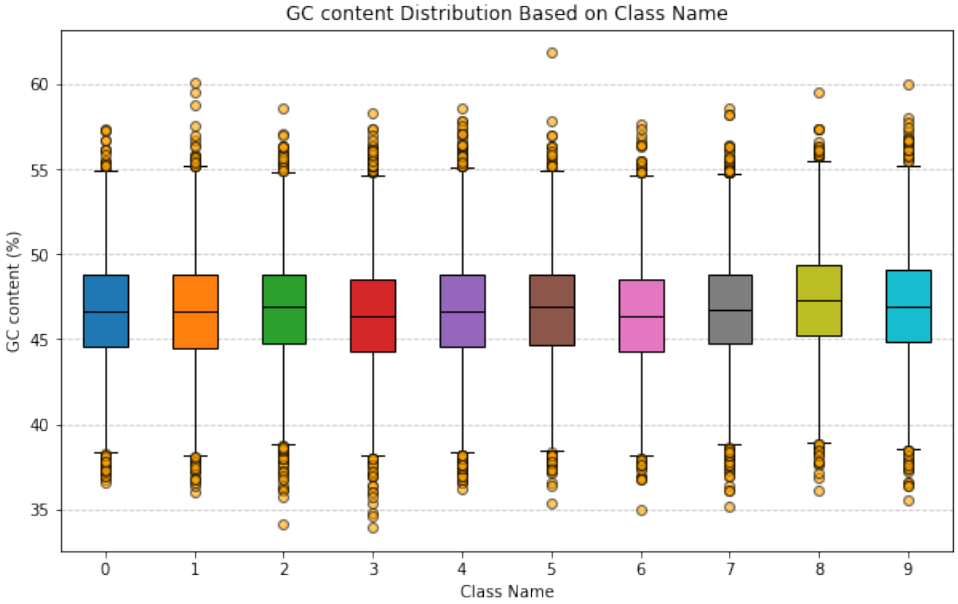


Fig. 5. Guanine-Cytosine (GC)-content boxplots by digit class: Most sequences fall within 40–60%, with a few outliers.

The next step is to evaluate the DNA sequence results with predetermined constraints, namely homopolymer and GC Content. Homopolymer constraints occur during the encoding process to generate DNA sequences. GC content is the percentage of the number of guanine (G) and cytosine (C) bases in a DNA sequence compared to the total length of the sequence [37]. The recommended threshold is 40–60%. The formula for obtaining the GC content is shown as follows:

$$\text{GC-Content} = \frac{G + C}{A + T + G + C}$$

The variables are defined as follows. The G represents the number of guanine bases. Then, C represents the number of cytosine bases, A is the number of adenine bases, and T shows the number of thymine bases.

The GC content evaluation result of DNA sequences generated from the MNIST images is presented in Fig. 5. Most sequences fall within the recommended threshold of 40–60%, although outliers appear across nearly all digit classes (0–9), indicating that the deviations are widespread rather than confined to specific classes. Several factors may contribute to this variation. First, differences in image structure influence GC content distribution, as thinner or irregularly shaped digits often produce more extreme values. Second, several constraints may affect the GC content and stability of the DNA sequence produced, especially when the thickness of the scratch varies. Third, this model struggles to handle unstandardized handwriting. For some complex numbers, such as '9' and '4', it produces a wider variety of shapes compared to simple numbers, such as '1', which in turn causes output fluctuations. Noise in the image causes the risk of reconstruction inaccuracy and errors that damage GC content. The key metrics for assessing this impact are detailed in Table II, including sequence length, GC content, and homopolymer, which is encoded by VAE. The final sequence result is determined by these factors.

TABLE II
SUMMARY STATISTICS OF DNA SEQUENCES GENERATED FROM VARIATIONAL AUTOENCODER (VAE)-BASED ENCODING METRIC.

Metric	Value
Average sequence length	155 bases
Minimum sequence length	152 bases
Maximum sequence length	158 bases
Average GC-content	48.7%
GC-content within recommended 40–60%	96.5% of sequences
Maximum homopolymer length (min–max)	3 bases (all sequences)
Sequences complying with homopolymer ≤ 3	100%

After evaluating the DNA sequences, the next stage is decoding, where the DNA sequences are converted into binary latent variable data again. The binary values are transformed into decimal latent variables and inserted into the decoder phase of the VAE. In this stage, every set of 10 latent variable values is reconstructed into an image. Figure 6 illustrates an example of a reconstructed image derived from its

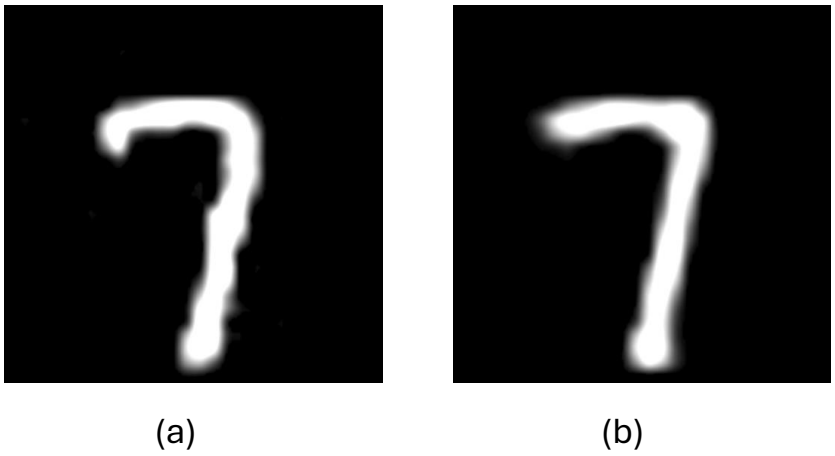


Fig. 6. Digit '7': (a) original image, (b) reconstruction from Variational Autoencoder (VAE) latent code after Deoxyribonucleic Acid (DNA) encoding–decoding pipeline.

TABLE III
CLASS-WISE STRUCTURAL SIMILARITY INDEX MEASURE (SSIM) AND PEAK SIGNAL-TO-NOISE RATIO (PSNR) ACROSS DIGIT CLASSES.

Class Name	PSNR (dB)	SSIM
0	18.70	0.8606
1	24.38	0.8916
2	16.85	0.7383
3	17.63	0.7639
4	18.51	0.7768
5	17.34	0.7537
6	18.64	0.8184
7	20.05	0.8160
8	16.79	0.7573
9	19.32	0.8247

original counterpart.

Visually, the original image (Fig. 6a) has a sharper contour of the '7' lift, while the reconstructed image (Fig. 6b) looks blurrier and less sharp. The shape of the number '7' is still recognizable, but there is a change in the upper left corner, where the tang looks more diffuse compared to the original image. The center of the number '7' is faded, which indicates that the contrast information in the original image is not fully retained in the reconstruction. The main difference in the reconstructed image is the decrease in sharpness due to excessive smoothing, which is probably caused by the loss of information in the encoding-decoding process and the decoder phase of the VAE architecture used.

While earlier studies have examined JPEG, DNA, and fountain code compression techniques, VAE-based methods achieve advantages in terms of the balance between nucleotide efficiency and reconstructed image quality. Nevertheless, there are some remaining challenges, one of which is accurately reconstructing fine details in complex figures. VAE models are able to outperform classical dimension reduction tools, such as PCA and standard autoencoders. Its main advantage is the adaptive latent representation, which allows it to consider biochemical constraints directly, such as GC balance and homopolymer constraints, without compromising reconstruction accuracy. These benefits are consistent with previous studies, demonstrating that learning-based image compression and deep representation models can improve the efficiency and robustness of DNA-based image storage [8, 11, 12]. In particular, Qres-VAE-based image compression is combined with Levenshtein coding and DNA mapping rules to improve information density while satisfying GC-content and homopolymer constraints [11]. Nevertheless, due to the limited scope of the present study, further comparative evaluations involving other compression methods are required.

Image quality for the original and reconstructed versions is quantified using the standard SSIM and PSNR metrics. A comparative analysis of the findings from this assessment is detailed in Table III. As presented in Table III, the model's ability to reconstruct images varies. This model performs optimally on digit '1' (PSNR of 24.38 dB, SSIM of 0.8916). This finding indicates that a simple and direct structure enables the capture of features with the highest accuracy. On the other hand, digit '2' produces the worst results (PSNR of 16.85 dB, SSIM of 0.7383), indicating a degradation in quality. This degradation also occurs for digits '3', '5', and '8'. For instance, digit '8' has a complex double-circle shape and complex curves, which seem to make it difficult for the model to maintain clear boundaries in the latent space. Similarly, digits '3' and '5', which combine straight lines with

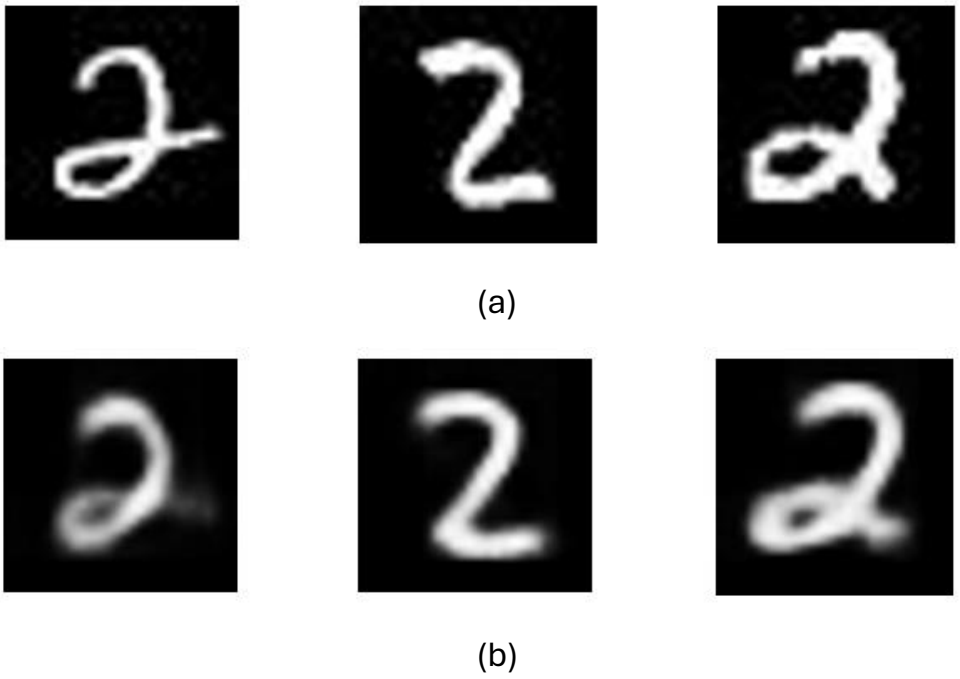


Fig. 7. Digit ‘2’: (a) original inputs and (b) reconstructed

curved sections, experience a loss of important fine details. These findings strongly support the suggestion that the structural complexity of an object is a key factor in determining the success of reconstruction, with an increase in variation leading to a decrease in reconstruction performance.

The inherent structural variation of the digits greatly influences the ability of VAE to reconstruct MNIST digits. Digits with higher complexity, such as ‘2’ ($\approx$ 7,501) and ‘8’ ($\approx$ 6,617), receive the worst reconstruction scores. For instance, digit ‘2’ produces less detailed output as measured by PSNR of 16.85 dB and SSIM of 0.7383, although its variance has decreased. Meanwhile, digit ‘1’, which has the lowest variance ($\approx$ 3,768), is successfully reconstructed with high accuracy, achieving a PSNR of 24.38 dB and an SSIM of 0.8916. These results indicate that the complexity of the digit determines the accuracy of the model reconstruction.

Statistical analysis finds a significant negative connection between original variance and PSNR (r = -0.80, $p < 0.01$; ρ = -0.64, $p < 0.05$). This result implies that quantitatively, reconstruction accuracy is affected by data complexity. Although in correlation analysis is observed with SSIM, the lower values indicate sensitivity compared to pixel-level metrics. This result is illustrated in Fig. 7, where the reconstructed number ‘2’ is degraded in its curve and joints, which is a direct result of the high variance in its class.

Although VAE-based methods are capable of shortening sequences without sacrificing digit structure, their performance has not yet been directly benchmarked against conventional 2-bit mapping techniques. Future studies should critically evaluate the improvements offered by VAE in terms of sequence efficiency and accuracy. Moreover, the research is unable to conduct this evaluation due to time and scope constraints. However, it is recommended that this kind of comparative analysis for future studies in this field.

IV. Conclusion

The research discovers that prior to DNA encoding, VAE application for image compression successfully reduces sequence complexity without damaging visual structure. The results show that the reconstructed images are visually similar to the originals, this finding is supported by SSIM values close to 1 (in the range of 0.80–0.89) and PSNR below 25 dB. Despite the numbers retaining their shape, the reconstruction experiences blurring and white noise. The accuracy of this process also correlates with the digits’ complexity. For instance, digit ‘1’ is reconstructed more accurately than the more complex digits ‘2’ or ‘8’. This relationship is also measured through correlation analysis, which reveals a significant negative relationship between image variance and PSNR, reflecting that reconstruction results are highly dependent on image complexity.

Overall, the research successfully demonstrates that integrating the VAE model into a DNA storage system can reduce the amount of encoded information, maintain storage efficiency, and preserve the main features of the image after reconstruction. However, there are challenges in maintaining high PSNR values and preserving fine details, especially in digits with high variance and complex structures. There is an absence of a direct comparison with 2-bit mapping, practical validation through actual DNA synthesis and sequencing, and evaluation limited to MNIST. Future research should focus on optimizing the VAE architecture and improving encoding techniques, exploring alternative VAE variants (e.g., β-VAE or Vector-Quantized Variational Autoencoder (VQ-VAE)), conducting head-to-head benchmarking (JPEG DNA, PCA, and standard autoencoders), and extending the evaluation to more complex datasets (e.g., Fashion-MNIST or CIFAR-10 grayscale).

Furthermore, to address the PSNR of < 25 dB issue, future studies will need to systematically increase the VAE latent dimensionality ($z > 10$) until it reaches similar biochemical constraints, and determine the optimal distortion-ratio point at a fixed nucleotide budget. As for direct comparison, future studies can adopt a unified assessment protocol with the same binarization, GC/homopolymer threshold, and matched nucleotide budget, reporting PSNR, SSIM, nucleotides per image, and constraint compliance, including an incomplete autoencoder baseline (latent ≈ 10) alongside JPEG-DNA and PCA. Subsequently, further research will be conducted using a stratified MNIST subset and extending the same protocol to broader datasets described above.

ACKNOWLEDGEMENT

The research was supported by an internal grant from Sekolah Tinggi Teknologi Wastukancana (No: 011/KL/STT-WKN/PWK/I/2022). The study was conducted at Sekolah Sains Data, Matematika, dan Informatika, IPB University, which served as the host institution and provided academic and facility support that enabled this work.

AUTHOR CONTRIBUTION

Developed the research idea, experimental design, and integration of VAE-based compression with DNA encoding, M. R. M.; Prepared the MNIST image dataset and organized the input data for training, testing, and evaluation, M. R. M.; Implemented the VAE architecture, binarization process, DNA encoding–decoding pipeline, and evaluation scripts, M. R. M.; Conducted model training, latent variable analysis, DNA sequence evaluation, PSNR or SSIM assessment, and interpretation of the results, M. R. M.; Prepared the original manuscript draft, including the methodology, results, discussion, conclusion, tables, and figures, M. R. M.; Coordinated the overall research process and incorporated feedback from all supervisors, M. R. M.; Provided conceptual direction and methodological guidance for the VAE-based DNA storage framework, Y. H.; Provided critical input on the interpretation of the experimental results and evaluation metrics, Y. H.; Supervised the research process and provided academic feedback for manuscript improvement, Y. H.; Provided guidance on the research design, validation strategy, and alignment between the proposed method and research objectives, A. B.; Provided feedback on the analysis of reconstruction quality, latent representation, and DNA sequence evaluation, A. B.; Provided supervisory support and constructive input during manuscript preparation, A. B.; Provided methodological advice related to model evaluation, result interpretation, and research contribution, K. P.; Provided input on the interpretation of PSNR and SSIM, K. P.; Provided academic supervision and recommendations for strengthening the research findings, K. P.; Provided conceptual guidance on biological constraints, DNA sequence feasibility, GC-content balance, and homopolymer limitations in the proposed DNA storage framework, I. Z. S.; Contributed domain knowledge related to DNA encoding rules, nucleotide representation, GC-content evaluation, and homopolymer constraint assessment, I. Z. S.; Provided input on the interpretation of DNA sequence evaluation results, including GC-content distribution, homopolymer length, and biological constraint compliance, I. Z. S.; and Provided supervisory guidance and recommendations to strengthen the biological validity of the DNA-based storage approach, I. Z. S.

DATA AVAILABILITY

The data that support the findings of the research are available from the MNIST database of handwritten digits and were accessed using the Keras MNIST dataset loader at https://keras.io/api/datasets/mnist/. These data are derived from the following resources available in the public domain: MNIST Handwritten Digit Database by Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The derived data generated in this study, including latent variables, binary representations, DNA sequences, and evaluation results, are available within the article and Appendix.

Cite this article as: M. R. Muttaqin, Y. Herdiyeni, A. Buono, K. Priandana, and I. Z. Siregar, "Efficient image encoding for DNA data storage using latent space representation of variational autoencoder", CommIT Journal 20(2), 219–231, 2026.

REFERENCES

[1] A. Doricchi *et al.*, "Emerging approaches to DNA data storage: Challenges and prospects," *ACS Nano*, vol. 16, no. 11, pp. 17 552–17 571, 2022.

[2] A. Akash, E. Bencurova, and T. Dandekar, "How to make DNA data storage more applicable," *Trends in Biotechnology*, vol. 42, no. 1, pp. 17–30, 2024.

[3] Y. Hao, Q. Li, C. Fan, and F. Wang, "Data storage based on DNA," *Small Structures*, vol. 2, no. 2, 2021.

[4] M. R. Muttaqin, Y. Herdiyeni, A. Buono, K. Priandana, and I. Z. Siregar, "Analysis and review of the possibility of using the generative model as a compression technique in DNA data storage: Review and future research agenda," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 3, pp. 472–489, 2023.

[5] L. Organick *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[6] Y. Ji *et al.*, "Towards automatic feature extraction and sample generation of grain structure by variational autoencoder," *Computational Materials Science*, vol. 232, 2024.

[7] C. Wang *et al.*, "Mainstream encoding–decoding methods of DNA data storage," *CCF Transactions on High Performance Computing*, vol. 4, no. 1, pp. 23–33, 2022.

[8] E. Upenik, D. Lazzarotto, M. Testolina, and T. Ebrahimi, "On the performance of learning-based image compression as source coding for JPEG DNA," in *Applications of Digital Image Processing XLVII*, vol. 13137. California, United States: SPIE, 2024, pp. 257–268.

[9] A. A. Jovith *et al.*, "DNA computing with water strider based vector quantization for data storage systems," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 6429–6444, 2023.

[10] Y. Feng, Q. Guo, W. Chen, and C. Han, "A low-complexity deep learning model for predicting targeted sequencing depth from probe sequence," *Applied Sciences*, vol. 13, no. 12, pp. 1–16, 2023.

[11] Y. Zheng, B. Cao, X. Zhang, S. Cui, B. Wang, and Q. Zhang, "DNA-QLC: An efficient and reliable image encoding scheme for dna storage," *BMC Genomics*, vol. 25, pp. 1–10, 2024.

[12] Y. Su, L. Chu, W. Lin, X. Yao, P. Xu, and W. Liu, "A robust and efficient representation-based DNA storage architecture by deep learning," *Small Methods*, vol. 9, no. 3, 2025.

[13] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, 2013.

[14] H. Shao, E. Ma, M. Zhu, X. Deng, and S. Zhai, "MNIST handwritten digit classification based on convolutional neural network with hyperparameter optimization," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 3595–3606, 2023.

[15] S. Bera, V. K. Shrivastava, and S. C. Satapathy, "Advances in hyperspectral image classification based on convolutional neural networks: A review," *Computer Modeling in Engineering & Sciences*, vol. 133, no. 2, pp. 219–250, 2022.

[16] Z. Gao, H. Z. Xuan, H. Zhang, S. Wan, and K. K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9280–9293, 2019.

[17] Z. Gao, Y. Li, and S. Wan, "Exploring deep learning for view-based 3D model retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1, pp. 1–21, 2020.

[18] R. Herrera-Pereda, A. T. Crispi, D. Babin, W. Philips, and M. H. Costa, "A review on digital image processing techniques for in-Vivo confocal images of the cornea," *Medical Image Analysis*, vol. 73, 2021.

[19] R. Janeliukstis and X. Chen, "Review of digital image correlation application to large-scale composite structure testing," *Composite Structures*, vol. 271, 2021.

[20] S. R. Lowe, "Pattern recognition in handwriting," in *Advances in pattern recognition and artificial intelligence*. World Scientific, 2022, pp. 77–95.

[21] Y. Zhao *et al.*, "Knowledge-aided convolutional neural network for small organ segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1363–1373, 2019.

[22] H. Akrami, A. A. Joshi, J. Li, S. Aydöre, and R. M. Leahy, "A robust variational autoencoder using beta divergence," *Knowledge-Based Systems*, vol. 238, 2022.

[23] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, "On the use of Kullback–Leibler divergence for kernel selection and interpretation in variational autoencoders for feature creation," *Information*, vol. 14, no. 10, pp. 1–15, 2023.

[24] S. Cao, J. Li, K. P. Nelson, and M. A. Kon, "Coupled VAE: Improved accuracy and robustness of a variational autoencoder," *Entropy*, vol. 24, no. 3, pp. 1–25, 2022.

[25] T. Ahmed and L. Longo, "Examining the size

of the latent space of convolutional variational autoencoders trained with spectral topographic maps of EEG frequency bands," *IEEE Access*, vol. 10, pp. 107 575–107 586, 2022.

[26] H. Tian, X. Jiang, S. Xiao, H. La Force, E. C. Larson, and P. Tao, "LAST: Latent Space-Assisted Adaptive Sampling for Protein Trajectories," *Journal of Chemical Information and Modeling*, vol. 63, no. 1, pp. 67–75, 2022.

[27] A. Raj and R. D'Souza, "DNA computing-enabled big data anonymization framework with a comparative study through existing anonymization tools," in *2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-RVITM)*. Bangalore, India: IEEE, Nov. 28–29, 2023, pp. 1–6.

[28] J. Zhang, D. Fang, and H. Ren, "Image encryption algorithm based on DNA encoding and chaotic maps," *Mathematical Problems in Engineering*, vol. 2014, no. 1, pp. 1–10, 2014.

[29] L. Organick *et al.*, "An empirical comparison of preservation methods for synthetic DNA data storage," *Small Methods*, vol. 5, no. 5, pp. 1–10, 2021.

[30] Y. Liu, X. He, and X. Tang, "Capacity-achieving constrained codes with GC-content and runlength limits for DNA storage," in *2022 IEEE International Symposium on Information Theory (ISIT)*. Espoo, Finland: IEEE, June 26–July 1, 2022, pp. 198–203.

[31] M. R. Raigonda, "Signature verification system using SSIM in image processing," *Journal of Scientific Research and Technology*, vol. 2, no. 1, pp. 5–11, 2024.

[32] Z. Wang *et al.*, "Vectorial-optics-enabled multi-view non-line-of-sight imaging with high signal-to-noise ratio," *Laser & Photonics Reviews*, vol. 18, no. 6, 2024.

[33] J. Yang, "acceleration system for single-precision floating-point arithmetic based on IEEE754 standard," *Highlights in Science, Engineering and Technology*, vol. 87, p. 82–89, 2024.

[34] M. Roy, S. Chakraborty, and K. Mali, "Audio encryption framework based on chaotic map and DNA encoding," *Applied Acoustics*, vol. 224, 2024.

[35] G. G. Dagher, A. P. Machado, E. C. Davis, T. Green, J. Martin, and M. Ferguson, "Data storage in cellular DNA: Contextualizing diverse encoding schemes," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 331–343, 2021.

[36] M. Welzel *et al.*, "DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage," *Nature Communications*, vol. 14, pp. 1–10, 2023.

[37] W. Song, K. Cai, M. Zhang, and C. Yuen, "Codes with run-length and GC-content constraints for DNA-based data storage," *IEEE Communications Letters*, vol. 22, no. 10, pp. 2004–2007, 2018.

APPENDIX

The Appendix can be seen in the next page.

TABLE A1
RESULTS OF BINARIZING VARIATIONAL AUTOENCODER (VAE) LATENT VARIABLES INTO IEEE 754 32-BIT FLOATING-POINT BINARY STRINGS. THESE BINARY REPRESENTATIONS SERVE AS THE BASIS FOR SUBSEQUENT DEOXYRIBONUCLEIC ACID (DNA) ENCODING.

Class	ID Image	Binary Data
0	1.jpg	00111111001000010110000000010000
		10111111100000100001100001110111
		10111111011001011001000011011011
		10111101010100010011101100010111
		00111110110101110000101010111101
		00111111100111111011011110110101
		10111111101000101111111110010110
		00111111010101011100101001111000
		00111111101010100001100111000101
		00111101110011100110011001001000
1	1002.jpg	00111111011101001000010011001000
		00111111101110000011111110110010
		10111110001011110111010000110100
		00111101001001000011001001000001
		00111110000001100001101000010000
		00111111110101001001010111000000
		00111111000000001001001101000100
		10111110111110110001011010111011
		10111111101011101111000100101001
		00111111000111000011010111111101
2	1011.jpg	10111111100110011011001011011101
		11000000000000100111000001010101
		10111111000000111011100110001101
		10111011111100111111100101011110
		10111110110111001110111101011111
		00111111001001001001001011111110
		10111111001100010011001010001010
		10111101101100101110111001000110
		00111110110101110001001101001110
		00111111001010100111000001111000

TABLE A2
DEOXYRIBONUCLEIC ACID (DNA) SEQUENCE GENERATED FROM BINARY DATA.

Class	ID Image	DNA Sequence
0	1.jpg	ATTTCCAAGTAAACCAAGTTTG
		AACAATAATGTTCTTGTAGTAG
		ACCTTCTTGGGTACATGTACCT
		ATTGTCCTAAGGGTTGGCTTTC
		CTTGTCTGTCCGTTTGCACCTTT
		GCCCATTTTCGGTTAGGCTGA
		ATTTGCCCAACGCTACCATTCT
		ATGCCGCAGA
1	1002.jpg	ATTCTGGCAAGCGCAACTTTC
		TGAATTTGCGCCCTTGAGTTCT
		CAATCAATCAGCAATAGCAAC
		ATTGAACGACGGACAAAGTTTG
		CCAGCCCTGAAACTTTCAAAGA
		GCGGAGAGTTGTTGTACCGGTG
		TGTTTGCCTCTGAGCCAGCTTG
		ATGACGGTTGC
2	1011.jpg	GTTTGATATCTAGCGTGAAAC
		ACATGACCCAGTCTTGAACTCT
		ATACGTCTCTTGCTTCCCATTC
		CTTCGTTATGTTCCGGTTGCAG
		CAGCCTTCCCTTCCGAGCGCC
		ACCCATTCGTAGTGTCACGAT
		TGTCTACATCATGATTCCCA
		GCTAACTGA