Analysis and review of the possibility of using the generative model as a compression technique in DNA data storage: review and future research agenda

Muhammad Rafi Muttaqin [a,b,1,*], Yeni Herdiyeni [a,2], Agus Buono [a,3], Karlisa Priandana [a,4], Iskandar Zulkarnaen Siregar [c,5]

[a] Department of Computer Science, IPB University, Bogor, Indonesia
[b] Informatic Engineering, Sekolah Tinggi Teknologi Wastukancana, Purwakarta, Indonesia
[c] Department of Silviculture, IPB University, Bogor, Indonesia
[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
* corresponding author

ARTICLE INFO

Article history
Received March 29, 2023
Revised May 10, 2023
Accepted May 21, 2023
Available online October 15, 2023

Keywords
DNA data storage
generative model
compression
deep learning
latent space

ABSTRACT

The amount of data in this world is getting higher, and overwriting technology also has severe challenges. Data growth is expected to grow to 175 ZB by 2025. Data storage technology in DNA is an alternative technology with potential in information storage, mainly digital data. One of the stages of storing information on DNA is synthesis. This synthesis process costs very high, so it is necessary to integrate compression techniques for digital data to minimize the costs incurred. One of the models used in compression techniques is the generative model. This paper aims to see if compression using this generative model allows it to be integrated into data storage methods on DNA. To this end, we have conducted a Systematic Literature Review using the PRISMA method in selecting papers. We took the source of the papers from four leading databases and other additional databases. Out of 2440 papers, we finally decided on 34 primary papers for detailed analysis. This systematic literature review (SLR) presents and categorizes based on research questions, namely discussing machine learning methods applied in DNA storage, identifying compression techniques for DNA storage, knowing the role of deep learning in the compression process for DNA storage, knowing how generative models are associated with deep learning, knowing how generative models are applied in the compression process, and knowing latent space can be formed. The study highlights open problems that need to be solved and provides an identified research direction.

1. Introduction

Human information storage and dissemination methods have undergone fundamental changes. In response to surviving in a complex and ever-changing environment, prehistoric man created paper from wood, leaves and clay and used it as a means of transmitting information [1]. With the advances in computer science, the information age has brought about a worldwide revolution. The digitized information stored in magnetic (diskettes), optical (CDs), and electronic (flash disks) media and sent via the internet has helped the development of knowledge, technology, and art for generations to come.

As the world continues to increase the amount of data (Fig. 1), traditional storage methods are facing more complex challenges [2]. Furthermore, the International Data Corporation estimates that

worldwide data storage demand will increase to 175 ZB (Fig. 1), or 1.75 x 1014 GB, by 2025 [3]. The estimated storage capacity will be exceeded by current storage media with a maximum density of 103 GB/mm^3 [4]. Besides, the cost of data maintenance and transmission, limited storage space, and significant data loss requires information storage [4].

The estimated storage capacity will be exceeded by current storage media with a maximum density of 103 GB/mm^3.

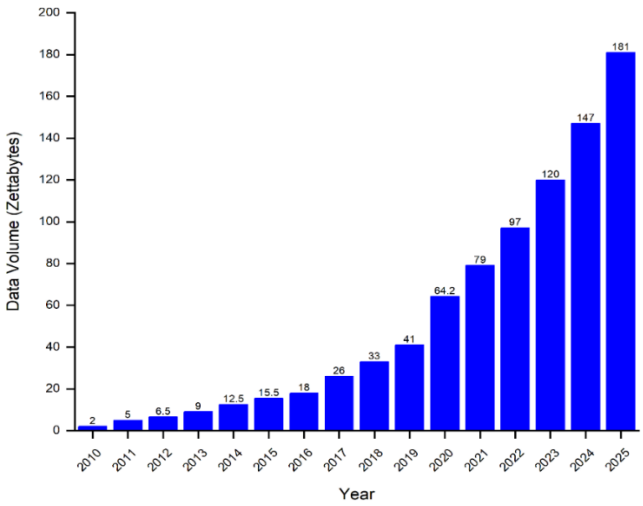

Fig. 1. Annual global data volume (modified from [3])

Almost all digital data is stored using technology that operates for a limited period. The lifespan of memory cards and chips is five years from their first use [5]. Standard hard drives are tolerant of damage caused by high temperatures, magnetic degradation, exposure to ultraviolet light, and mechanical damage. While a solid-state drive (SSD) performed better than a hard drive, it will lose its information if it is not used for more than a few months. [5].

Instead, nature has solved this problem in its own way since the beginning of time on Earth: by storing unique information that characterizes organisms in a unique sequence of bases (A, T, C, G) contained within a small molecule called deoxyribonucleic acid (DNA). For three billion years, this method of information storing has been used. As an information carrier, DNA molecules offer several advantages over conventional storage media. DNA's high storage density, low maintenance costs, and other outstanding properties would make it a durable information storage option in the future [6].

The storage capacity of DNA is phenomenal. Castillo stated that the entirety of the internet's information could be stored on devices smaller than cubic inch units [5]. DNA is considered an ideal medium in this regard since instead of computers that utilize 1's and 0's for storing data, DNA consisting of adenine, guanine, cytosine, and thymine (A, G, C, and T) that have been paired into the two nucleotide base pairs A-T and G-C can be used to store information in the form of binary code [1]. Since a single nucleotide can represent two bits of information, DNA is viewed as an ideal storage medium as the demand for high-capacity storage media increases. Therefore, 1 gram of single-stranded DNA can encode 455 EB of information (ssDNA) [5]. All of the data created by the entire world in a single year can be stored in just 4 grams of DNA [1]. Due to its three-dimensional (3D) structure, DNA provides ample storage space.

DNA has a more significant temperature tolerance (-800 to 800 °C). DNA uses energy millions of times more efficiently than today's personal computers. In addition, DNA has more storage options than most media because it stores data in nonlinear structures, as opposed to linear systems. DNA promises more opportunities for enhancing latency and data extraction because it permits bidirectional data reading. DNA is safe and unlikely to be damaged by living organisms due to the significant fact that it is invisible to the human eye [1].

In light of DNA's potential as a medium for information storage, numerous studies have been conducted to determine how digital information can be stored in DNA. Fig. 2 depicts the process of generally storing information in DNA. The process of converting digital files of various formats, such as images, videos, music, and documents, into binary code is known as the binarization stage. The process of converting files that are already in binary format into the form of a row of four DNA bases is known as encoding. Researchers continue to refine this encoding procedure. After the data is converted into a line of DNA molecules, DNA synthesis is performed, which involves inserting the DNA line into a living organism or creating artificial DNA that will be stored in a location or tube. To retrieve data that has been stored in the DNA medium, the sequencing procedure is performed, which entails reading the baseline of DNA molecules from the DNA medium, resulting in data in the form of rows of DNA molecular bases. The row's data will be decoded, which involves converting it back into binary code. Once the binary code has been recovered, it will be converted back to the original data that was initially stored. In general, this is the stage at which information can be stored and extracted from DNA [7].

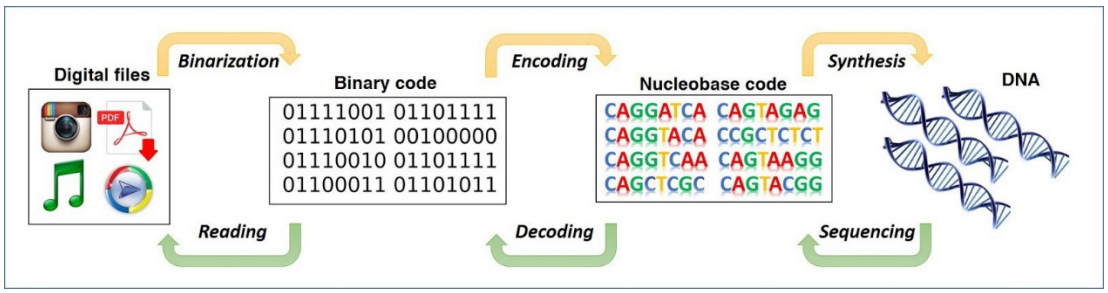

Fig. 2. DNA process data storage

Binarization stage converts digital data such as photos, documents, and videos into binary code. In an image, the pixel consists of 3 channel (RGB) color values with a value range of 0 – 255 in each pixel. For example, the pixel value of 255 will be converted to a binary value with eight digits, namely 11111111, and a value of 25 will become 00011001. In general, computers will read digital data as information that has a binary code [8] [9].

Encoding stage converts the binary code into a DNA sequence (A, C, T, & G). Several encoding techniques have been carried out, as in the research by Goldman *et al.* [10]. Goldman *et al.* [10] encoded 739 kilobytes of hard-disk storage with an estimated Shannon information of 5.2 x 10^6 bits into DNA code, then synthesized this DNA, sequenced it, and rebuilt the original file with 100% accuracy. Erlich and Zielinski [11] used the DNA fountain technique that explored the limit architecture in terms of bytes per molecule and obtained a perfect retrieval from a density of 215 petabytes per gram of DNA, orders of magnitude higher than previous reports. In addition, several researchers also did encoding in the DNA storage process, namely Yi Zhang *et al.* [12], Anavy *et al.* [13], Newman *et al.* [14], Kosuri & Church [15], Lee *et al.* [16] and Takahashi *et al.* [17].

The synthesis stage is the process of forming artificial DNA from a DNA sequence. According to Dong *et al.* [18], there are several synthesis techniques, such as: (1) Based on solid-phase phosphoramidite chemistry, (2) Array-based DNA synthesis, (3) Based on enzymatic synthesis. This synthesis cost is expensive, so the DNA storage technology cannot be adapted into a digital data storage technology. For example, DNA storage costs 800 million USD per terabyte of data, compared to tape storage which only costs 15 USD per terabyte [19]. Thus, a technique is needed to minimize the costs required in the synthesis process.

The sequencing process is reading the DNA sequence from a DNA medium. The result of the sequencing process is the nucleotide base sequence of a DNA, namely Adenine Cytosine, Thymine, and Guanine. According to Dong *et al.* [18], there are several sequencing techniques, such as: (1) Sanger sequencing, (2) Next generation sequencing, (3) Heli Scope single-molecule sequencer (4) Pacific biosciences SMRT technology, (5) Odford nanopore technologies and (6) Single-cell genomic sequencing technologies.

This decoding technique is the opposite of encoding, for example :

00 A (Encoding)
A 00 (Decoding)

Thus, this decoding technique is usually one unit with the encoding algorithm, such as compress and decompress. Reading process returns digital data from binary code to initial data, for example, images, videos, and others [8][9].

The cost of chemical DNA synthesis, which is $3,500 per 1 megabyte of information (Fig. 3) [11], is still quite high. The expensive synthesis process is one of the primary reasons why DNA data storage technology has not been widely adopted [7]. Consequently, both the Binarization stage and the DNA sequence data can be compressed prior to the synthesis process in the DNA data storage stage. The goal is for the amount of DNA sequence data to be synthesized to be minimal so that the cost of the synthesis process can be kept to a minimum.

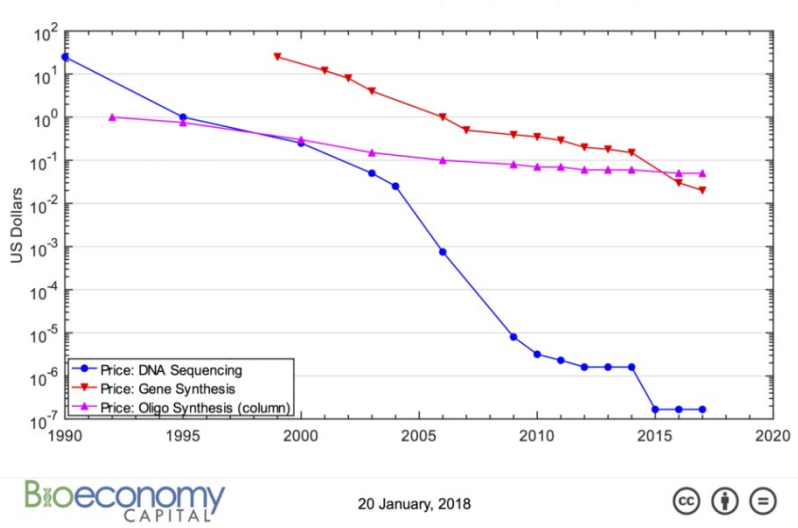

Fig. 3. The cost of DNA synthesis and sequencing [11]

As a result of multimedia and communication technology advancement, multimedia entertainment plays a crucial role in contemporary human life. Imagery and video play a significant role in modern multimedia entertainment. This can provide a method for storing and transmitting image and video data and become essential when internet bandwidth is constrained, especially for large, high-quality digital images. Researchers are concerned about image compression technology due to the internet's bandwidth constraints, which hinder the development of image communication. Image compression aims to represent and transmit large original images using the fewest bytes possible and restore images with acceptable results. Deep Learning-based image compression [20] is one of the image compression techniques currently undergoing development.

The image's features can be automatically rather than manually learned using the deep learning model. Image recognition can be made more efficient with the addition of convenient features. During this time, the image features are always determined manually by the initial knowledge of the model maker, and the number of features is limited. An infinite number of features are learned automatically by the deep learning model. Optimizing image processing requires a method for extracting features that is effective. Using deep learning models, unpredictable image characteristics can be learned and used for image security. Consequently, the deep learning model can also be applied to image compression [20].

Generative models are one of the deep learning techniques utilized in the image compression process. Generative models describe the generation of a data set in terms of the opportunity model. Using the sampling model as an example, we can generate new data [21]. Generative models are also called deep generative models. The word profound is used here because the focus will be on a generative model with neural network representation. Where neural networks exhibit adaptability and strength. With the

development of neural networks and the rise in computing power, the deep generative model has emerged as one of the primary directions for advancing artificial intelligence.

This paper review will focus on image compression using deep generative modeling. We will determine if the compression using generative modeling achieves sufficient compression of the image and if it has been implemented for DNA data storage to reduce the cost of synthesizing the DNA sequence to be stored in the DNA medium.

2. Related Works

2.1. DNA Data Storage

Using DNA, scientists have already begun a major project to develop an alternative to data storage. Watson and Crick published one of the oldest and most influential publications in biology history in the journal Nature in 1953, which suggests that DNA is a transporter of genetic information [22]. Since that time, DNA also known as the genetics of an organism's information has been stored in a four-base linear row. Many researchers proposed storing specialized information in DNA after a decade [23]. However, this was unsuccessful due to the limited knowledge of proper DNA synthesis and sequencing techniques.

In 1988, Joe Davis created the first chance to compile information storage on DNA, or DNA Storage [24]. The information contained in the pixel value of a "Microvenus" image was converted into a line of 0-1 arranged in a 5 x 7 matrix, where 1 indicates a dark pixel and 0 indicates a bright pixel. The information is then encoded into DNA molecules with 28 base pairs (bp) and fed to *Escherichia coli* bacteria. After being successfully recovered using DNA Sequencing, the original image was successfully viewed again. Clelland proposed in 1999 using a method based on "DNA micro-dots" such as steganography to conceal information within DNA molecules [25]. Bancroft, two years later, introduced the idea of using DNA bases to directly modify English writing in the same way that amino acid sequences in DNA are changed.

Church and Goldman led the field of DNA Storage research in 2012 [10][26]. Church could store up to 659 KB of information in the DNA model, whereas the previous maximum size that could be stored successfully was less than 1 KB. Goldman contains more data, amounting to 739 bytes. According to these two studies, the data stored in DNA includes not only text but also images, sounds, PDF files, and so on.

The research of Church and Goldman is the genesis of additional research in the broader field of DNA storage research. Thus, the amount of data that can be stored continues to increase as methods become more complex. By the end of 2018, the maximum amount of data that could be stored amounted to 200 MB, stored in over 13 million oligonucleotides. Alongside the continued advancement of DNA Synthesis and DNA Sequencing technology, this new method of DNA storage continues to evolve, bringing the application of DNA storage closer to fruition (Fig. 4).

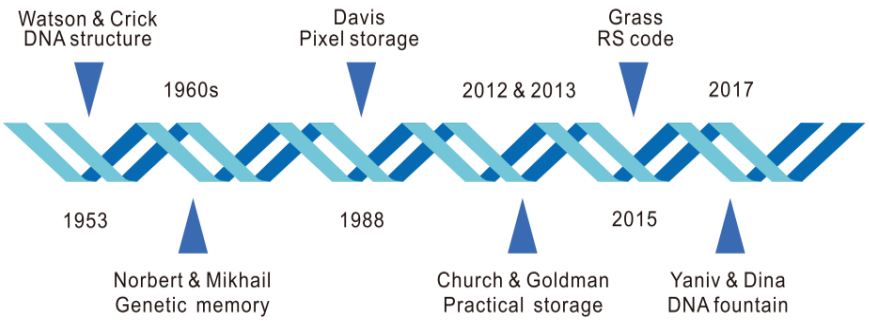

Fig. 4. The origins of DNA storage. Journey in DNA storage research [22], [24]–[26].

2.2. Generative Model

Generative models are one of the most indicative areas of artificial intelligence's rapid development [27][28]. Comparable to teams of counterfeiters attempting to produce and use counterfeit currency undetected, generative models can be compared to the police trying to detect counterfeit currency. In this analogy, competition prompted both teams to perfect their techniques until counterfeit goods were identical to the original [29]. The objective of a generative model was to investigate training data sets or examples and the distribution of opportunities that could re-generate those data. The generative model relies on Deep Learning [29].

Deep generative models can be divided into three major categories (Fig. 5): autoregressive generative models (ARM), flow-based models, and latent variable models. Text analysis [30], image analysis [29], audio analysis [31], active learning [32], reinforcement learning [33], graph analysis [34], medical imaging [35], image compression [36], and other applications use deep generative modeling.

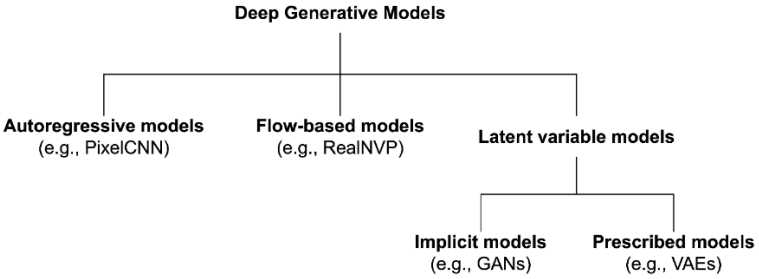

Fig. 5. Diagram representation of deep generative models

Supervised learning is straightforward to ascertain, and all supervised learning algorithms have essentially the same objective: learning to connect new input examples with output correctly. For instance, an object recognition algorithm can associate a cat's photograph with a breed-specific identifier.

Unsupervised learning is a subfield of machine learning that contains numerous algorithms with varying objectives. The primary aim of unsupervised learning is to discover something useful by analyzing datasets containing examples of unlabeled inputs. Typical examples of unsupervised learning include grouping and dimension reduction. Generic modeling is an additional method for unsupervised learning. Examples of x training are taken from the $p_{data}(\boldsymbol{x})$ distribution in generative modeling. The objective of generative modeling algorithms is to examine a $p_{model}(\boldsymbol{x})$ that closely resembles $p_{data}(\boldsymbol{x})$. Using latent variable $\boldsymbol{z}$ with a fixed prior distribution $p(\boldsymbol{z})$, such as a Gaussian distribution and a network of decoders or generators that calculate $\boldsymbol{x} = f(\boldsymbol{z})$, generative models implicitly define the distribution of $p_{model}(\boldsymbol{x})$ [36].

Directly examine pdata approximation by writing the $p_{model}(x;\theta)$ function controlled by the parameter and searching for parameter values that bring pdata and pmodel as close together as possible. In particular, maximum likelihood estimation, which minimizes the Kullback-Leibler divergence between the pdata and the model, is likely the most popular approach to generative modeling. Taking the average of a set of observations to estimate the average parameters of a Gaussian distribution is one of the connotations of maximum likelihood estimation. This method relies on the density function depicted in Fig. 6.

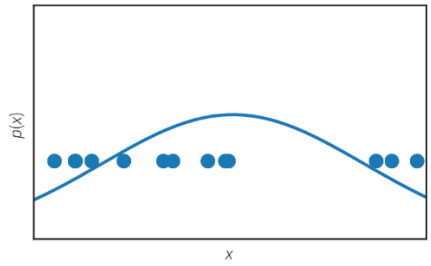

Fig. 6. Illustration of density estimation with multiple data points on the actual number line used to match the Gaussian density function describing the observed example

In recent years, generative models, such as generative adversarial network (GAN) and variational autoencoder (VAE), have dominated unsupervised deep learning techniques [37]. GAN is trained and reused as a fixed feature extractor for supervised tasks [38]. The network is based on the Convolutional Neural Network (CNN) and has demonstrated its superiority in visual data analysis as unsupervised learning. Sparse Autoencoder was trained on large-scale image datasets to study features in another study [39]. This network generates a high-level feature extractor from unlabeled data that can be used for unsupervised face detection. The resulting features are sufficiently discriminatory to identify other high-level objects, such as animals or human bodies.

3. Method

In this section, the author describes research questions, search processes, study selection criteria, and approaches to quality assessment. This review paper utilizes Paganelli *et al.* [40] guidelines and protocols. Once a research question has been identified, a search strategy is formulated, and pertinent articles are extracted from multiple scientific journal databases. The obtained paper is subjected to the study selection criteria, and some of these papers will be selected again for quality evaluation. A collection of successfully identified articles was chosen following rigorous quality testing. The author carefully read this paper and answered the research question satisfactorily.

3.1. Research Question

This study defines the following research question (RQ) for the review. The selection of this RQ is based on the fact that the response to this question will explain the primary purpose of this paper and serve as a model for future research.

- How can Machine Learning be applied to DNA storage?
- How are compression techniques utilized in methods for DNA data storage?
- What role does deep learning play in DNA data storage compression?
- How are generative models associated with deep learning?
- How can generative models be implemented in compression strategies?
- How can a latent space/generative model be constructed?

3.2. Search Process

The search sources are from the online digital databases, such as: IEEE Xplore, Science Direct, ACM Digital Library, Wiley Online Library, and other sources accessible via the Publish and Perish applications. The SLR will search for articles published period from 2012 to 2022, this considered due to the development of DNA data storage began to reappear in 2012 and advanced quite rapidly. Various word combinations are employed to restrict the scope of the search. Each RQ uses a unique query, as presented in Table 1.

Table 1. List of queries for each RQ

ID	Research Question	Query
1	How can Machine Learning be applied to DNA storage?	“machine learning” AND (“dna data storage” OR “dna based storage”)
2	How are compression techniques utilized in methods for DNA data storage?	“compression” AND (“dna data storage” OR “dna based storage”)
3	What role does deep learning play in DNA data storage compression?	"deep learning" AND "compression" AND ("DNA data storage" OR "DNA based storage")
4	How are generative models associated with deep learning?	"generative model" AND "deep learning"
5	How can generative models be implemented in compression strategies?	"generative model" AND "compression"
6	How can a latent space/generative model be constructed?	"method" AND "generative model" AND ("latent space" OR "latent variable")

3.3. Selection Criteria

Many articles are extracted from electronic article search databases to compile a comprehensive review for this article. The numerous articles will be filtered due to the exclusion criteria as in Table 2. Meanwhile, Fig. 7 depicts the results of paper selection using the PRISMA method.

Table 2. Criteria in the selection of articles

No.	Criterion
1	Titles and abstracts are inconsistent with the article's purpose.
2	Articles is written not in english
3	Years were issued outside 2012-2022 (RQ 1-3) and 2018-2022 (RQ 4-6).
4	Articles outside quartile scopus
5	Duplicate articles
6	Furthermore, journals and proceedings

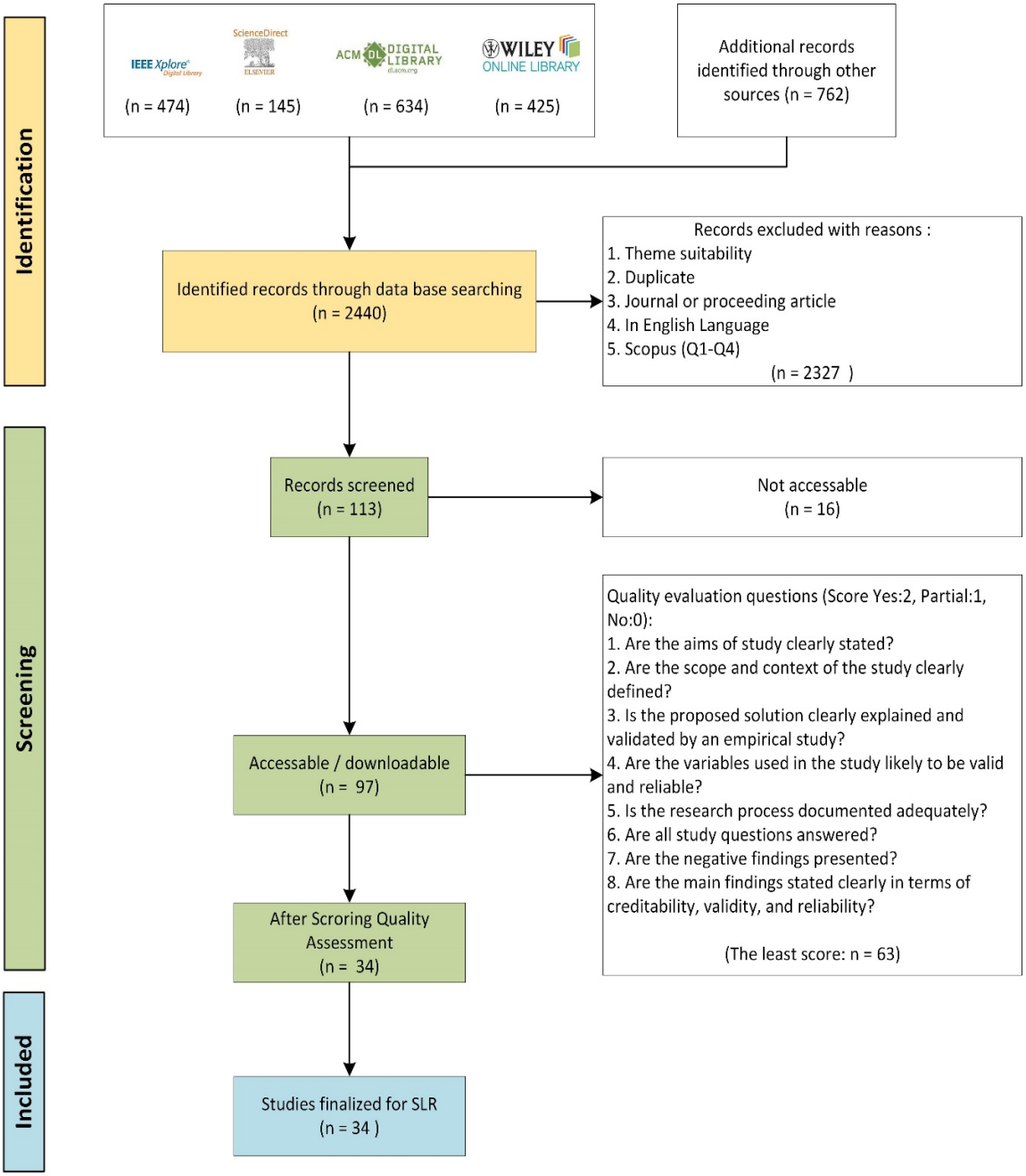

Fig. 7. The selection results use PRISMA

3.4. Quality Assessment

The chosen paper is then evaluated using the quality evaluation method described in the paper [41]. Table 3 contains a list of questions to assess the article's quality. Articles with scores below eight were removed from the list to refine the search further. This quality assessment procedure is included in Fig. 5's eligibility stage. Table 4 also depicts the number of papers selected after the quality assessment. Fig. 8 shows the results of the quality assessment process. After applying the selection criteria and conducting a quality evaluation, 34 articles remained.

Table 3. Quality evaluation questions. Yes score 2; Partial score 1: No score 0

ID	Questions	Yes (2)	Partial (1)	No (0)
Q1	Are the aims of the study clearly stated?			
Q2	Are the scope and context of the study clearly defined?			
Q3	Is the proposed solution clearly explained and validated by an empirical study?			
Q4	Are the variables used in the study likely to be valid and reliable?			
Q5	Is the research process documented adequately?			
Q6	Are all study questions answered?			
Q7	Are the negative findings presented?			
Q8	Are the main findings stated clearly in terms of creditability, validity, and reliability?			

Table 4. Process of paper selection

Source	After query search	After applying the selection criteria	After quality assessment
IEEE Xplorer	474	30	11
Science Direct	145	4	7
ACM Digital Library	634	30	8
Wiley Online Library	425	19	1
Manual Search	762	14	7
Total	2440	97	34

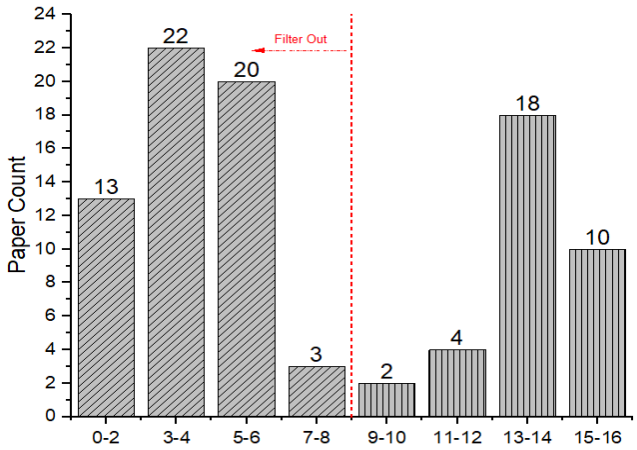

Fig. 8. Quality assessment result

4. Results and Discussion

This section presents the answers to the study's research questions. Each research contest answer is supported by selecting our search results articles.

4.1. RQ1. How can Machine Learning be applied to DNA storage?

The application of machine learning techniques to DNA data storage methods has begun. As Stanley *et al.* [42] demonstrate, machine learning techniques are used to overcome errors caused by repeated oligo or rewriting of DNA oligo during encoding before DNA synthesis. In a separate study, Ben Cao *et al.* [43] developed machine learning through the Damping Multi-Verse Optimizer (DMVO) algorithm to optimize the encoding process based on the constraints of DNA sequence arrangement. In the production of DNA strands, one must consider making DNA a more efficient storage medium and avoiding errors during the synthesis and sequencing processes. According to the study, DNA storage coding is limited by the Hamming distance constraint, the storage edit distance constraint, the GC content constraint, the no-run length constraint, and the uncorrelated address constraint.

Chao Pan *et al.* [44] used signal processing and machine learning techniques to store quantized imagery in DNA without using oligo or excessive rewriting to address error and cost issues. The core of the research method consists of quantizing and compressing color images using a unique encoding method on each of the three color channels. The quantization scheme reduces the image color palette to eight intensity levels per channel and compresses the intensity level by combining Hilbert-space Filing Curves, differential, and Huffman Coding.

The answer to this research question concludes that machine learning methods have been implemented in DNA data storage research at the encoding stage to minimize errors during DNA synthesis and sequencing. The formation of DNA through synthesis must adhere to certain biological constraints. This limitation is due to the DNA base protein's inherent properties.

4.2. RQ2. How are compression techniques utilized in methods for DNA data storage?

DNA synthesis is one of the steps involved in DNA data storage. This method is used to store data on DNA both in Vivo and in Vitro. Currently, the cost of synthesis is typically higher than the cost of sequencing [45]. This is the rationale utilized by Shufang *et al.* to compress digital information data [46]. Before converting the original digital file into a DNA sequence, they proposed Huffman's quaternary coding method to compress the binary data. Based on the statistical properties of the source, Huffman's proposed quaternary process can achieve a very high compression ratio for files by distributing non-uniform opportunities from the source file. Fig. 9 depicts, in general, the process of encoding DNA using the Huffman quaternary coding method. This method can also be proposed to correct standard synthesis and sequencing errors. The study's findings were able to convert a 5,2 KB document into 3,934 bits of DNA base.

Due to the high cost of DNA synthesis, Mishra *et al.* [47] also compress digital data during the DNA data storage process. They proposed a method for efficiently compressing digital data into DNA sequences using Huffman tree drinking variation. This method simultaneously addresses DNA coding's limitations. GC-constraint and run-length constraints are the limitations that have been overcome. Fig. 9 depicts the methods utilized in their research. The illustration of a simple encoding algorithm has a 2-digit binary mapping rule that converts to one of the nucleotides of DNA (A, C, T, G), as shown in Table 5.

Table 5. The Conversion Formula from Binary Code to DNA Nucleotide in Encoding Step

Binary Code	Nucleotide of DNA	Remarks
00	A	Adenine
10	G	Guanine
01	C	Cytosine
11	T	Thymine

For instance, if the binary code is as follows: 1100011010110001, then it will be converted into DNA nucleotide bases: TACGGTAC. For the conversion formula in the decoding stage, it is just the opposite of encoding. However, the DNA sequence must consider "biological constraints" for the DNA produced during the synthesis process to match the character of the DNA compound bonds [48].

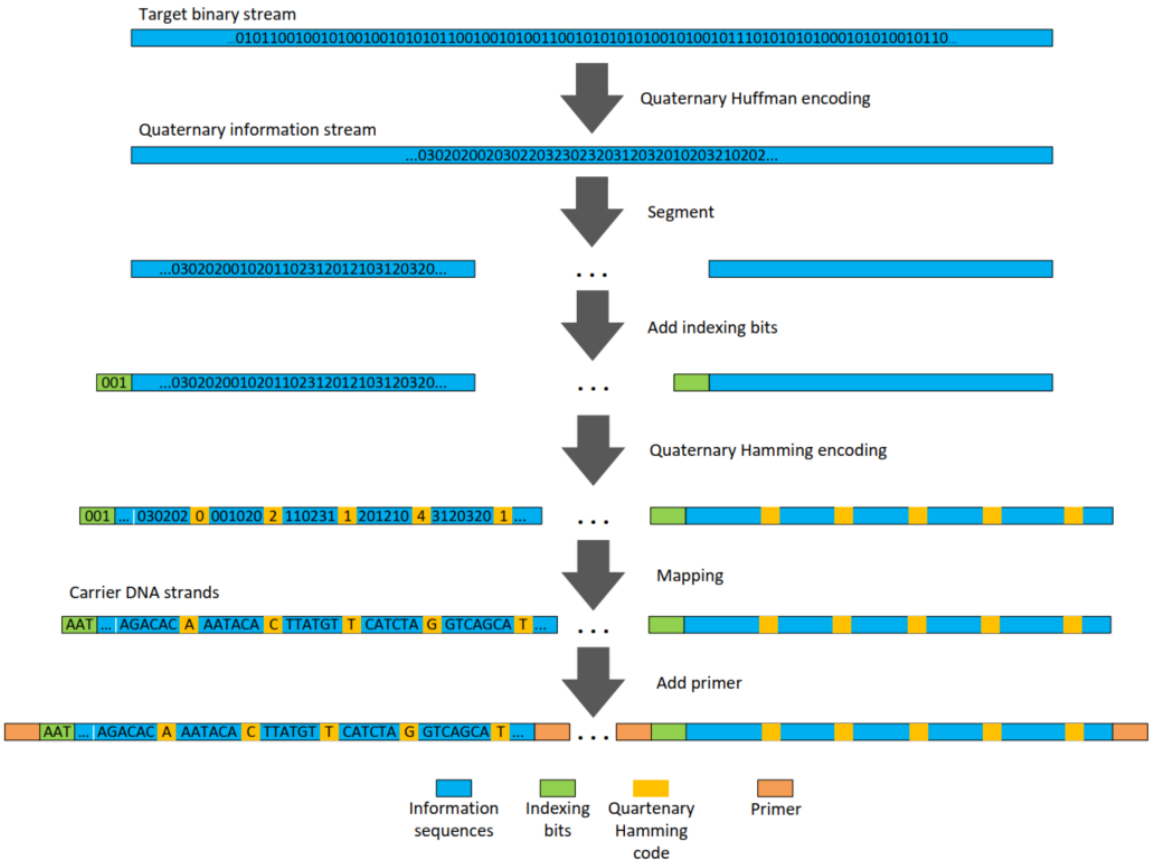

Fig. 9. Encoding DNA process with Quaternary Huffman Coding (modified from [46])

Fig. 10 illustrates the study's methodology. In the initial step (1), input data is converted to a binary row. Using Huffman's Minimum Variance coding, step two (2) of the DNA Tree is constructed. Encoding algorithms obtain DNA data utilizing the DNA Tree and binary rows. The DNA for the fourth line is stored. The fifth (5) stored DNA is extracted. The sixth (6) uses recovered DNA data and the DNA Tree algorithm, which restores the same binary sequence using decoding algorithms. The seventh (7) original input data is retrieved from the recovered binary sequence.

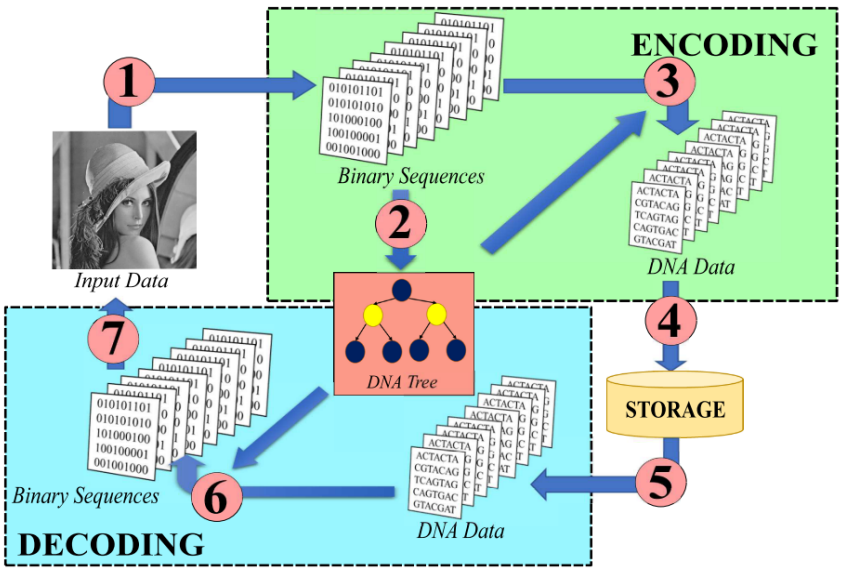

Fig. 10. Binary code of DNA data storage method used by Mishra *et al.* [47]

The answer to this research question concludes that scientists utilize compression techniques to reduce the price of DNA synthesis. Most scientists compress the input/digital data of the file/document stored on the DNA medium. With all of its development, the Huffman Code method is the most widely used technique for digital data compression.

4.3. RQ3. What role does deep learning play in DNA data storage compression?

As described in the third research question section on compression techniques used in the DNA data storage method, several studies have employed compression techniques to address the issue of the expensive DNA synthesis procedure. This section will present some research that uses compression techniques for DNA data storage, but compression techniques are based on deep learning.

In their research, Franzese *et al.* [49] utilized neural network-based compression techniques. They convert an image into a latent space representation that is then stored in DNA. Compression based on neural networks produces excellent results. The generative model technique achieves compression outcomes ten times superior to the conventional scheme. In addition to reducing the cost of synthesis, this technique can also cover data back with lower coverage and tolerate numerous errors, thereby reducing the cost of sequencing.

Franzese *et al.* [49] then utilized Huffman coding algorithms to convert binary data from Latent Space to ternary, then converted to quaternary DNA using the rotational code method to ensure biological constraints were met. The questioned biological constraint is the G-C content (ratio of base G to base C), and there is no homopolymer repetition. These constraints may cause difficulties during synthesis and sequencing.

This third research question concludes that few researchers have employed deep learning techniques to compress digital data for DNA data storage. Due to their high potential, it is worthwhile to attempt to integrate network-based compression or deep learning into the concept of DNA data storage. This must be done to provide a diverse option for the future development or implementation of DNA data storage.

4.4. RQ4. How are generative models associated with deep learning?

Commonly, generative models are utilized as potent instruments for feature extraction, regression, clustering, and classification. Generative models are also used for pattern recognition via data generation, recommendation generation, topic modeling, text generation, etc. [50]. Several deep learning and machine learning algorithms derive their concept or operation from the generative model method. The algorithms Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Latent Dirichlet Allocation (LDA), Boltzmann Machine (BM), Variational Auto Encoder (VAE), and Generative Adversarial Network (GAN) use generative model concepts [50]. Fig. 11 illustrates the distinction between algorithm machine learning and deep learning classification within the generative model concept. Generative models are widely used in deep learning algorithms, particularly for generating new outputs that resemble the input.

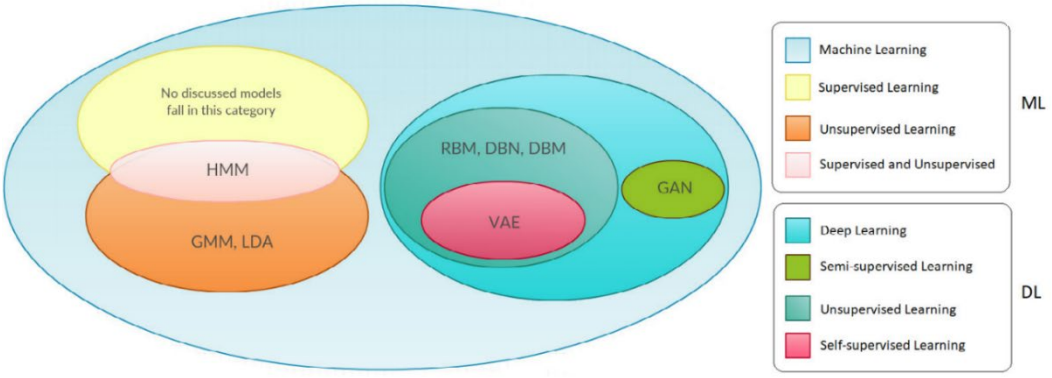

Fig. 11. The classification of machine learning (ML) and deep learning (DL) algorithms with generative model concepts [50]

Discriminative is the antonym of generative. This phrase is similar to the concepts of supervised and unsupervised in machine learning or deep learning [51]. The primary objective of generative models is to capture the joint distribution of all system variables, including input variables. Due to the lack of labels associated with each input pattern, such learning abilities are typically referred to as unsupervised learning. The objective is to create an internal model of the input environment by identifying a set of latent features that precisely characterize the correlation between statistically observed variables.

This fourth research question concludes that generative models and deep learning have a strong relationship. In machine learning, generative and discriminatory models share similarities with unsupervised and supervised learning. We are also aware that deep learning is a subset of machine learning. Consequently, some deep learning algorithms employ concepts from generative models.

4.5. RQ5. How can generative models be implemented in compression strategies?

Generative models are widely implemented in deep learning algorithms for various beneficial purposes, including data compression [52]. Compression is one of the existing steganography methods [53]. This procedure necessitates the explicit distribution of generative objects. In this compression-based steganography method, generative models provide an advantage. This is because generative models can offer the ideal "sampler" or explicit distribution of generative media. Variational Autoencoder (VAE), Generative Adversarial Network (GAN), and similar algorithms are generative model-based algorithms. The algorithm can produce objects derived from latent variables that adhere to past distribution patterns, such as the Gaussian Distribution.

Variational Autoencoder (VAE) was used to perform encryption and compression in the study [54]. This is because VAE can compress and encrypt images more efficiently and accelerate image encryption. VAE is a generative model that can generate similar images through neural network training and unsupervised learning. Changing the weight and bias of the generative model resulted in the generation of an unidentified noise image, an encrypted image, for the first time in this study. Using two images to train weights and biases from VAE to generate different images is the method. The system then divides the weight and bias of two different training images and divides the data into generated models to create noise images.

One of the generative model implementations in the deep learning method is the autoencoder architecture. There are various kinds of development of Autoencoder, including the Variational Autoencoder (VAE). Several researchers use variational autoencoders to compress digital data, for example, image data. In the autoencoder architecture, there are two main parts: Encoder and Decoder. The output from the Encoder is a representation of Latent Space with much smaller dimensions. This is meant by "compression," which is performed by the Generative Model. This compression intends to minimize the image dimensions. This compression will be used to overcome the problem of the high cost of synthesis in the DNA storage process at this time. With the image dimension compression or reduction process, the resulting DNA sequence will have a much smaller number compared to images that are not compressed. Image compression using a generative model has been carried out by Liu *et al.* [55] with predetermined input images.

4.6. RQ6. How can a latent space/generative model be constructed?

A latent variable is a low-dimensional subspace generated by projecting a monitored multivariable sample room [56]. Autoencoder (AE) is one of the most effective and versatile unsupervised learning techniques for reducing the dimensions of big data models. Variational Autoencoder (VAE), an extensification of AE, can discover an efficient latent variable space as a magnitude of multivariate normal distribution by adding constraints on the coding network; VAE is a nonlinear form of probabilistic principal component analysis (PCA). VAE uses Bayesian variational inference for parameter estimation and integrates AE into a generative framework. This technique can be used for dimensionality reduction, reconstruction, and generation. Recent studies have shown great interest in this regard.

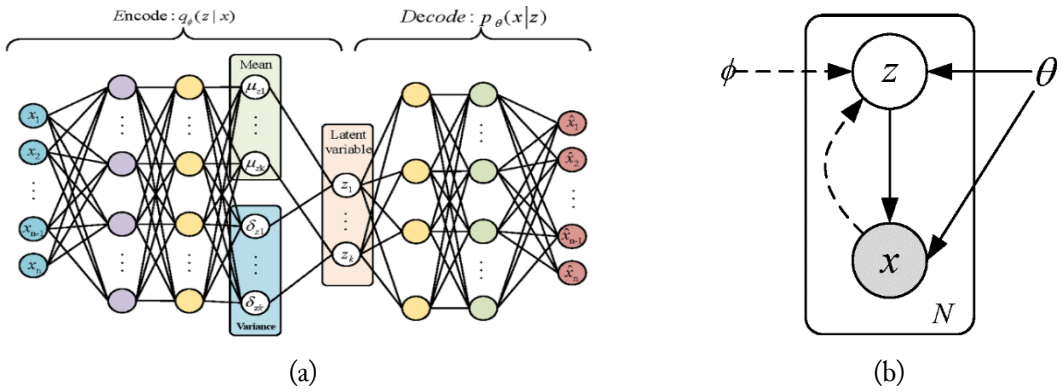

Fig. 12. VAE structure illustration [56]

Fig. 12(a) depicts the structure of the VAE, which consists of two components: encoder and decoder. Observation data $x = [x_1, x_2, \dots, x_n]^T$ is generated by a random process involving the latent variable $z = [z_1, z_2, \dots, z_n]^T$, and $\boldsymbol{x}$ can be reconstructed from $\boldsymbol{z}$ using generation models. The model $q_\theta(z|x)$ is considered a probabilistic encoder and $p_\theta(x|z)$ a probabilistic encoder. It is referred to as a probabilistic decoder. The VAE is graphically represented in Fig. 12(b). The filled and outlined circles represent the observed and latent variables, respectively. Arrows represent dependencies, whereas plates represent the number of instances.

VAE's capability of encapsulating complex data distributions into continuous, low-dimensional latent spaces makes it ideal for design applications due in large part to the characteristics of its latent space [57]. It is more difficult to replicate the latent space in generative adversarial network (GAN) algorithms, but GAN has proven to be an excellent candidate for generative design applications [58].

In deep learning, latent space / latent variables are produced from the Autoencoder architecture, which has two primary components, the Encoder and Decoder a shown in Fig. 13.

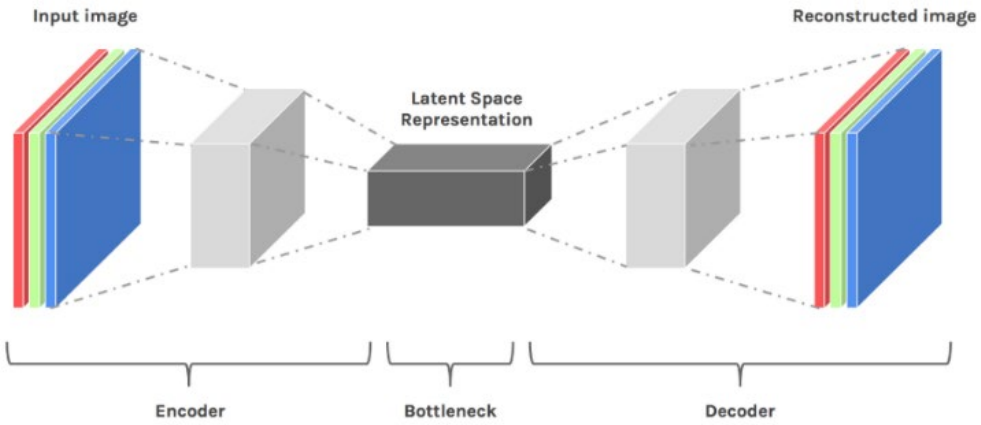

Fig. 13. Autoencoder architecture component [59]

From Fig. 13, it can be seen that the latent space representation can be generated from the Autoencoder architecture and its derivatives. The contents of the latent space will be used in the binarization stage in the initial process of DNA data storage. The contents of this latent space are compressed or reduced data.

This sixth research question concludes that generative algorithm models, including VAE, can generate latent space. A latent variable is a variable that fills the latent space, where the amount depends on the requirements and will serve as the basis for generating new data based on input. In general, a less dimensional latent space extracts more meaningful change directions from the data but suffers more significant compression losses, making it more difficult to reconstruct the input data from its

representation of latent. This means that latent spaces of lower dimensions tend to generate fewer and more diverse design candidates in the exploration context of the design space. However, the smaller the latent space dimension will become, the simpler it will be to explore.

4.7. Discussion

DNA data storage research journey from the 1960s to the present has been extensive. From discovering DNA compounds to how digital data can be stored on DNA. DNA data storage research begins with a problem, as with most scientific discoveries and research. This research is motivated by the current limitations of data storage media. The boundaries that arise consist of the materials required to create the storage technology and the increasing volume of data produced. This issue prompted the development of research into alternative data storage media.

DNA base compounds have been demonstrated to be a storage medium for living organisms' biological data. Numerous studies have utilized information from biological data collected on living things thousands of years ago. Based on this study, the potential of DNA for data storage is enormous. Therefore, research into the viability of DNA as a digital data storage medium was initiated.

Storing digital data on DNA involves six steps: binarization, encoding, synthesis, sequencing, decoding, and reading. Existing research focuses on each stage individually or as a whole. There is research on creating or developing algorithms during the encoding process or steps. This stage converts binary data extracted from a document to be stored into a row of four DNA bases, specifically A, T, G, and C. The research results on this DNA sequence indicate that constraints must be considered during its preparation. GC constraints and the numerous repetitions of rows in the form of homopolymers are two limitations. Therefore, researchers continue to develop algorithms aided by machine learning techniques, to facilitate adaptation to DNA's biological constraints.

The DNA data storage process is also hindered by the expense of the Synthesis stage, which entails creating artificial DNA or synthesizing DNA based on the Encoding stage's sequence. Compared to the sequencing process, synthesis is still costly. This is possible due to the restricted ownership of DNA synthesis-capable instrumentation. Due to this cost barrier, the researchers are attempting to reduce the cost of synthesis by compressing the data stored on DNA. Whether digital data compression or DNA row data occurs first, the synthesis process will involve compression.

This process of DNA data storage incorporates a variety of compression techniques. In research, Huffman coding is the most frequently used compression method. In addition to Huffman coding, compression techniques based on deep learning are being tested for incorporation into the DNA data storage process. However, this compression method based on deep learning has not been applied extensively to DNA data storage techniques. This compression method based on deep learning employs the concept of generative models. This method generates a latent variable with a smaller dimension than the input. This latent variable represents the data input distribution despite its smaller dimension size. Therefore, the technique can generate or generate new data resembling input data.

Limitations of the research reviewed in this paper related to its implementation in the DNA data storage process is that existing research has not specifically used the generative model method or utilized latent space in digital data compression to minimize the cost of the synthesis process. The compression carried out by previous research is at the digital data stage, which has been through the binarization step. There is also compression carried out at the DNA Sequence data stage. From the results of this paper, it is possible to compress digital data from the start (beginning), especially digital image data. In this compression, generative models can be used to reduce the size of the image dimensions to represent latent space with one of the deep learning architectures, such as the Autoencoder.

5. Conclusion

The possibility of using generative models as compression techniques in DNA data storage remains extremely open, even though this article has addressed some research questions. Several deep learning

algorithms employ generative model concepts and are designed for various applications, including data compression. Latent variables contain the compression technique's underlying principle using this generative model. In this generative model, latent variables have a smaller dimension than input data. Consequently, latent variables will be used to carry out the encoding process during the DNA data storage phase.

Variational Autoencoder is a deep learning algorithm that uses generative model concepts for data compression (VAE). In addition to VAE, implementing the Generative Adversarial Network (GAN) algorithm includes a generative model concept. Research on digital data storage methods on DNA enables deep learning algorithms in the data compression procedure to reduce synthesis costs. The results of this SLR show that the use of the generative model method on DNA data storage has not been used, so it is still widely potentials to use the generative model method to be integrated at the stage of digital data storage in the DNA medium.

This review paper provides information regarding the use of generative models in data compression in general, particularly about DNA data storage. This study hows research opportunities in utilizing the Generative model to compress data and integrate it with DNA data storage stages. This can develop existing methods of using DNA as a digital data storage medium. With the increase in DNA data storage methods, there will be further opportunities for developing the implementation of digital data storage technology in DNA. Furthermore, the cost of DNA synthesis technology will be cheaper. Future research is recommended to investigate how to integrate the benefits of genetic models in deep learning algorithms into the stages of storing digital data on DNA. Thus, this application's primary objective is to maximize data compression to reduce synthesis costs. In addition to compressing data to be stored, it must also be capable of decompressing data to be restored as its original state quality.

Acknowledgment

The authors would like to express the sincere gratitude to the Bunga Bangsa Foundation, Sekolah Tinggi Teknologi (STT) Wastukancana, and the Department of Computer Science of IPB University for the generous support and assistance in the research project that led to this publication. Their commitment to advancing research and education in the field of computer science is greatly appreciated.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. No supporting funding for this work.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] U. J. Lee, S. Hwang, K. E. Kim, and M. Kim, "DNA Data Storage in Perl," *Biotechnol. Bioprocess Eng.*, vol. 25, no. 4, pp. 607–615, Aug. 2020, doi: 10.1007/s12257-020-0022-9.

[2] A. Doricchi *et al.*, "Emerging Approaches to DNA Data Storage: Challenges and Prospects," *ACS Nano*, vol. 16, no. 11, pp. 17552–17571, Nov. 2022, doi: 10.1021/acsnano.2c06748.

[3] IDC, Seagate, and Statista estimates, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025," *IDC; Statista*, 2021. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/#:~:text=The total amount of data,replicated reached a new high.

[4] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, and W. L. Hughes, "Nucleic acid memory," *Nat. Mater.*, vol. 15, no. 4, pp. 366–370, 2016, doi: 10.1038/nmat4594.

[5] M. Castillo, "From hard drives to flash drives to DNA drives," *Am. J. Neuroradiol.*, vol. 35, no. 1, pp. 1–2, 2014, doi: 10.3174/ajnr.A3482.

[6] Z. Ping *et al.*, "Towards practical and robust DNA-based data archiving using the yin–yang codec system," *Nat. Comput. Sci.*, vol. 2, no. 4, pp. 234–242, Apr. 2022, doi: 10.1038/s43588-022-00231-2.

[7] A. C. Patel and C. G. Joshi, "Deoxyribonucleic Acid as a Tool for Digital Information Storage: An Overview," *Indian J. Vet. Sci. Biotechnol.*, vol. 15, no. 01, pp. 1–8, 2019, doi: 10.21887/ijvsbt.15.1.1.

[8] C. K. Lim, S. Nirantar, W. S. Yew, and C. L. Poh, "Novel Modalities in DNA Data Storage," *Trends Biotechnol.*, vol. 39, no. 10, pp. 990–1003, 2021, doi: 10.1016/j.tibtech.2020.12.008.

[9] Y. Hao *et al.*, "Data Storage Based on DNA," *Small Struct.*, vol. 2, no. 2, p. 2000046, 2021, doi: 10.1002/sstr.202000046.

[10] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013, doi: 10.1038/nature11875.

[11] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science (80-.).*, vol. 355, no. 6328, pp. 950–954, 2017, doi: 10.1126/science.aaj2038.

[12] Y. Zhang *et al.*, "Information stored in nanoscale: Encoding data in a single DNA strand with Base64," *Nano Today*, vol. 33, pp. 6–11, 2020, doi: 10.1016/j.nantod.2020.100871.

[13] L. Anavy, I. Vaknin, O. Atar, R. Amit, and Z. Yakhini, "Data storage in DNA with fewer synthesis cycles using composite DNA letters," *Nat. Biotechnol.*, vol. 37, no. 10, pp. 1229–1236, 2019, doi: 10.1038/s41587-019-0240-x.

[14] S. Newman *et al.*, "High density DNA data storage library via dehydration with digital microfluidic retrieval," *Nat. Commun.*, vol. 10, no. 1, pp. 1–6, 2019, doi: 10.1038/s41467-019-09517-y.

[15] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: Technologies and applications," *Nat. Methods*, vol. 11, no. 5, pp. 499–507, 2014, doi: 10.1038/nmeth.2918.

[16] H. H. Lee, R. Kalhor, N. Goela, J. Bolot, and G. M. Church, "Terminator-free template-independent enzymatic DNA synthesis for digital information storage," *Nat. Commun.*, vol. 10, no. 1, p. 2383, Jun. 2019, doi: 10.1038/s41467-019-10258-1.

[17] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, "Demonstration of End-to-End Automation of DNA Data Storage," *Sci. Rep.*, vol. 9, no. 1, pp. 1–5, 2019, doi: 10.1038/s41598-019-41228-8.

[18] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, and L. Qian, "DNA storage: Research landscape and future prospects," *Natl. Sci. Rev.*, vol. 7, no. 6, pp. 1092–1107, 2020, doi: 10.1093/nsr/nwaa007.

[19] L. Ceze, J. Nivala, and K. Strauss, "Molecular digital data storage using DNA," *Nat. Rev. Genet.*, vol. 20, no. 8, pp. 456–466, Aug. 2019, doi: 10.1038/s41576-019-0125-3.

[20] H. M. Yasin and A. M. Abdulazeez, "Image Compression Based on Deep Learning: A Review," *Asian J. Res. Comput. Sci.*, no. May, pp. 62–76, 2021, doi: 10.9734/ajrcos/2021/v8i130193.

[21] D. Foster, *Generative Deep Learning*, vol. 6, no. November, p. 308 2019. [Online]. Available: https://books.google.co.id/books?hl=en&lr=&id=BEq8EAAAQBAJ&oi=fnd&pg=PT13&dq=D.+Foster,+Generative+Deep+Learning,+vol.+6,+no.+November.+2019. .

[22] X. Wu, K. Wang, X. Wang, H. Kan, and J. Kurths, "Color image DNA encryption using NCA map-based CML and one-time keys," *Signal Process. 22*, vol. 148, pp. 272–287, 2018, doi: 10.1016/j.sigpro.2018.02.028.

[23] X. Li, S. Zhou, and L. Zou, "Design of DNA Storage Coding with Enhanced Constraints," *Entropy*, vol. 24, no. 8, p. 1151, Aug. 2022, doi: 10.3390/e24081151.

[24] M. Dimopoulou and M. Antonini, "Data and image storage on synthetic DNA: existing solutions and challenges," *EURASIP J. Image Video Process.*, vol. 2022, no. 1, p. 23, Oct. 2022, doi: 10.1186/s13640-022-00600-x.

[25] D. Na, "DNA steganography: Hiding undetectable secret messages within the single nucleotide polymorphisms of a genome and detecting mutation-induced errors," *Microb. Cell Fact.*, vol. 19, no. 1, pp. 1–9, 2020, doi: 10.1186/s12934-020-01387-0.

[26] L. Piantanida and william l. Hughes, "A PCR-free approach to random access in Dna," *Nat. Mater.*, vol. 20, no. 9, p. 1172, 2021, doi: 10.1038/s41563-021-01090-4.

[27] A. Brock, J. Donahue, and K. Simonyan, "Large scale GaN training for high fidelity natural image synthesis," in *7th International Conference on Learning Representations, ICLR 2019*, 2019, pp. 1–35, [Online]. Available: https://arxiv.org/abs/1809.11096.

[28] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, 2021, doi: 10.1109/TPAMI.2020.2970919.

[29] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[30] Z. Wang and Q. Wu, "An Integrated Deep Generative Model for Text Classification and Generation," *Math. Probl. Eng.*, vol. 2018, pp. 1–8, Aug. 2018, doi: 10.1155/2018/7529286.

[31] Y. Zhao, X. Xia, and R. Togneri, "Applications of Deep Learning to Audio Generation," *IEEE Circuits Syst. Mag.*, vol. 19, no. 4, pp. 19–38, 2019, doi: 10.1109/MCAS.2019.2945210.

[32] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 5971–5980, 2019, doi: 10.1109/ICCV.2019.00607.

[33] Y. Liu, N. Qiao, and Y. Altinel, "Reinforcement Learning in Neurocritical and Neurosurgical Care: Principles and Possible Applications," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–6, 2021, doi: 10.1155/2021/6657119.

[34] M. Simonovsky and N. Komodakis, "GraphVAE: Towards generation of small graphs using variational autoencoders," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11139 LNCS, pp. 412–422, 2018, doi: 10.1007/978-3-030-01418-6_41.

[35] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoder," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, 2019, no. 2014, pp. 1–18, [Online]. Available: http://proceedings.mlr.press/v121/ilse20a.html.

[36] S. Huang, A. Makhzani, Y. Cao, and R. Grosse, "Evaluating lossy compression rates of deep generative models," in *37th International Conference on Machine Learning, ICML 2020*, 2020, vol. 119, pp. 4394–4404, [Online]. Available: https://proceedings.mlr.press/v119/huang20c.html.

[37] M. Shyu, S. Chen, and S. S. Iyengar, "A Survey on Deep Learning Techniques," *Strad Res.*, vol. 7, no. 8, Aug, pp. 1- 6 2020, doi: 10.37896/sr7.8/037.

[38] K. Raza and N. K. Singh, "A Tour of Unsupervised Deep Learning for Medical Image Analysis," *Curr. Med. Imaging Rev.*, vol. 17, no. 9, pp. 1059–1077, 2021, doi: 10.2174/18756603mtezonzmk0.

[39] A. Rezvani, M. Bigverdi, and M. H. Rohban, "Image-based cell profiling enhancement via data cleaning methods," *PLoS One*, vol. 17, no. 5 May, pp. 1–19, 2022, doi: 10.1371/journal.pone.0267280.

[40] A. I. Paganelli *et al.*, "Real-time data analysis in health monitoring systems: A comprehensive systematic literature review," *J. Biomed. Inform.*, vol. 127, no. September 2021, p. 104009, Mar. 2022, doi: 10.1016/j.jbi.2022.104009.

[41] Z. Kang, C. Catal, and B. Tekinerdogan, "Machine learning applications in production lines: A systematic literature review," *Comput. Ind. Eng.*, vol. 149, no. April, p. 106773, 2020, doi: 10.1016/j.cie.2020.106773.

[42] P. M. Stanley, L. M. Strittmatter, A. M. Vickers, and K. C. K. Lee, "Decoding DNA data storage for investment," *Biotechnol. Adv.*, vol. 45, no. September, p. 107639, 2020, doi: 10.1016/j.biotechadv.2020.107639.

[43] B. Cao, X. Ii, X. Zhang, B. Wang, Q. Zhang, and X. Wei, "Designing Uncorrelated Address Constrain for DNA Storage by DMVO Algorithm," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 19, no. 2, pp. 866–877, Mar. 2022, doi: 10.1109/TCBB.2020.3011582.

[44] C. Pan, S. M. Hossein Tabatabaei Yazdi, S. Kasra Tabatabaei, A. G. Hernandez, C. Schroeder, and O. Milenkovic, "Image Processing in DNA," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 8831–8835, 2020, doi: 10.1109/ICASSP40776.2020.9054262.

[45] L.-F. Song, Z.-H. Deng, Z.-Y. Gong, L.-L. Li, and B.-Z. Li, "Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities," *Front. Bioeng. Biotechnol.*, vol. 9, no. June, p. 13, Jun. 2021, doi: 10.3389/fbioe.2021.689797.

[46] S. Zhang, B. Huang, X. Song, T. Zhang, H. Wang, and Y. Liu, "A high storage density strategy for digital information based on synthetic DNA," *3 Biotech*, vol. 9, no. 9, p. 342, Sep. 2019, doi: 10.1007/s13205-019-1868-4.

[47] P. Mishra, C. Bhaya, A. K. Pal, and A. K. Singh, "Compressed DNA Coding Using Minimum Variance Huffman Tree," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1602–1606, 2020, doi: 10.1109/LCOMM.2020.2991461.

[48] A. Rasool, Q. Qu, Y. Wang, and Q. Jiang, "Bio-Constrained Codes with Neural Network for Density-Based DNA Data Storage," *Mathematics*, vol. 10, no. 5, p. 845, Mar. 2022, doi: 10.3390/math10050845.

[49] J. Zrimec *et al.*, "Controlling gene expression with deep generative design of regulatory DNA," *Nat. Commun.*, vol. 13, no. 1, p. 5099, Aug. 2022, doi: 10.1038/s41467-022-32818-8.

[50] G. M. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Comput. Sci. Rev.*, vol. 38, p. 100285, 2020, doi: 10.1016/j.cosrev.2020.100285.

[51] A. Testolin, M. Piccolini, and S. Suweis, "Deep learning systems as complex networks," *J. Complex Networks*, vol. 8, no. 1, pp. 1–21, 2020, doi: 10.1093/comnet/cnz018.

[52] C.-Y. Zhang, Q. Zhao, C. L. P. Chen, and W. Liu, "Deep compression of probabilistic graphical networks," *Pattern Recognit.*, vol. 96, p. 106979, Dec. 2019, doi: 10.1016/j.patcog.2019.106979.

[53] K. Chen, H. Zhou, H. Zhao, D. Chen, W. Zhang, and N. Yu, "Distribution-Preserving Steganography Based on Text-to-Speech Generative Models," *IEEE Trans. Dependable Secur. Comput.*, vol. 19, no. 5, pp. 3343–3356, Sep. 2022, doi: 10.1109/TDSC.2021.3095072.

[54] X. Duan, J. Liu, and E. Zhang, "Efficient image encryption and compression based on a VAE generative model," *Journal of Real-Time Image Processing*, vol. 16, no. 3. pp. 765–773, 2019, doi: 10.1007/s11554-018-0826-4.

[55] X. Liu *et al.*, "Medical Image Compression Based on Variational Autoencoder," *Math. Probl. Eng.*, vol. 2022, pp. 1–12, Dec. 2022, doi: 10.1155/2022/7088137.

[56] C. Huang, Y. Chai, Z. Zhu, B. Liu, and Q. Tang, "A Novel Distributed Fault Detection Approach Based on the Variational Autoencoder Model," *ACS Omega*, vol. 7, no. 3, pp. 2996–3006, 2022, doi: 10.1021/acsomega.1c06033.

[57] R. Danhaive and C. T. Mueller, "Design subspace learning: Structural design space exploration using performance-conditioned generative modeling," *Autom. Constr.*, vol. 127, p. 103664, Jul. 2021, doi: 10.1016/j.autcon.2021.103664.

[58] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for Medical Image Synthesis: An Empirical Study," *J. Imaging*, vol. 9, no. 3, p. 69, 2023, doi: 10.3390/jimaging9030069.

[59] F. Blom, "Unsupervised Feature Extraction of Clothing Using Deep Convolutional Variational Autoencoders," p. 83, 2018. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1230233&dswid=-6899.